The Incremental Development of a Type 2 Diabetes Knowledge-Based System using Ripple-Down Rules - A Socio-Technical Perspective on Managing Type 2 Diabetes

A thesis submitted in fulfilment of the requirement for the
Degree of Doctor of Philosophy

From

University of Technology Sydney

By

Adel Fathy Omar

Faculty of Engineering and Information Science

August 2025

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Adel Omar declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature of Student: Production Note:
Signature removed prior to publication.

Date: 29 / 08 / 2025

ACKNOWLEDGMENT

I would like to express my sincere appreciation to my supervisors, Professor Ghassan Beydoun and Professor Nagesh Shukla, for their academic guidance, unwavering support, and patience throughout my PhD journey. Their continued encouragement and invaluable insights have been instrumental in the completion of this thesis.

I would also like to extend my gratitude to Professor Khin Than Win from the University of Wollongong for her support and guidance. My sincere appreciation also goes to retired Emeritus Professor Paul Compton from the University of New South Wales, whose expertise and advice greatly contributed to this research. I am also deeply grateful to Professor Herbert F. Jelinek from Khalifa University, whose generous provision of data made this research possible. Without the support of these four individuals, I sincerely doubt that I would have been able to complete this journey.

My deepest appreciation goes to my wife, Jihan Abdel-Fattah, whose unwavering support, patience, and encouragement sustained me throughout this journey. Her sacrifices, understanding, and steadfast belief in me were truly limitless, and I am forever grateful.

Finally, I would like to express my deepest and most heartfelt gratitude to my late father, Dr. Fathy Omar, who has always been pillars of strength and guidance in my life. He was a profound influence in shaping my educational journey. He instilled in my siblings and I the unwavering belief in the power of knowledge and the importance of lifelong learning. His wisdom, encouragement, and passion for education remain with me, and I only wish he were here to witness this milestone. This thesis is a testament to his legacy, and I dedicate this achievement to his memory.

Publications

Omar, A., Beydoun, G. & Jelinek, H. 2024. Overcoming Data Scarcity: Strategic Data Manipulation in the Development of a Knowledge-Based System for Type 2 Diabetes Prediction.

Omar, A., Beydoun, G., Win, K. T., Shukla, N., Jelinek, H. & Elias, H. 2023. Cultivating Expertise: Unravelling Type 2 Diabetes Associations through Incremental Knowledge-Based System Development: Ripple Down Rules or Machine Learning.

Omar, A., Beydoun, G., Win, K.T. & Jelinek, H. (2022), 'The Incremental Development of a Diabetes 2 Knowledge Base System using Ripple Down Rules', *Proceedings of the Pacific Asia Conference on Information Systems*, July 5-9, Taipei/Sydney Virtual Conference.

Omar, A., Beydoun, G., Win, K., Shukla, N. & Baker, G. 2019, 'Socio-Technical Perspective on Managing Type II Diabetes', paper presented to the ACIS2019: 30th Annual Conference on Information Systems, Perth, Australia, December 9-11, 2019.

Abstract

This thesis develops an incremental knowledge-based system (KBS) for type 2 diabetes that integrates social determinants of health using Ripple-Down Rules (RDR). Motivated by scarce, heterogeneous data, the approach captures expert refinements as exception rules, enabling transparent updates without large training sets. Compared with conventional machine-learning pipelines that may discard sparse cases as outliers, the RDR-based KBS accommodates contextual factors while remaining auditable and interpretable. The method iteratively integrates new evidence on social determinants and standardises area-level attributes, producing a rule base that adapts to emerging knowledge and variable data quality. The framework is scalable to other settings and supports policy design by revealing stable associations between socio-demographic factors and risk. Empirical evaluation confirms the practicality and durability of RDRs for building and maintaining a diabetes KBS under data paucity. Overall, the work provides a reproducible pathway for knowledge-driven decision support in public health informatics, demonstrating how social determinants can be operationalised in an interpretable system for disease prevention and management.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

This chapter is divided into three sections. The thesis commences with a detailed examination of the escalating global challenge of type 2 diabetes in Section 1.1, highlighting its significant health and economic impacts and the initial intent to utilize geographic information systems (GIS) to identify high-risk zones. This focus shifts towards a more inclusive approach that integrates a variety of social determinants, leading to the development of a knowledge base (KB) through Ripple-Down Rules (RDRs) for their incremental development in response to data scarcity, as discussed in Section 1.2. This section explores the feasibility of creating a predictive tool to aid public health planning. Section 1.3 outlines the thesis structure, emphasising the integration of social determinants into a comprehensive public health strategy and diabetes management, aiming to bridge gaps in population health planning and support evidence-based decision making. This holistic approach reflects a significant advancement in addressing the multifaceted nature of type 2 diabetes through innovative technological solutions.

1.1 Background

The global prevalence of type 2 diabetes has been steadily increasing, representing a burgeoning public health concern. As of 2011, this condition affected an estimated 366 million people worldwide (Sim et al., 2017). Projections suggest that this number will surge to approximately 552 million individuals by 2030 (Sim et al., 2017). In the Australian context, diabetes imposes a substantial economic burden, estimated at $14.6 billion (Diabetes in Australia, 2015a). This figure marks a significant escalation from the $10.6 billion reported in 2005 (Lee et al., 2013). While researchers are diligently striving for a cure, the absence of a definitive solution underscores the importance of effective diabetes management, both in terms of financial implications and societal well-being.

Initially, this research contemplated a geographic information systems (GIS) approach, intending to employ spatial data to identify regions susceptible to type 2 diabetes outbreaks. This method would have entailed collecting data and utilising GIS technology to discern patterns in areas characterised by a high prevalence of diabetes type 2. However, as elucidated in subsequent sections, this thesis evolved to initiate the development of a knowledge base, and a decision support system (DSS) predicated on social determinants. It is essential to note that a person's place of birth and residence constitute significant social determinants. Consequently, the concept

of integrating GIS into the research remains viable for future exploration. For the present research, the primary emphasis is on the incremental development of a KB, driven by the data collected.

The study explores the development of a knowledge-based system. The current research suggests that knowledge-based systems (KBS) have found application in the healthcare domain. However, this application predominantly gravitates toward clinical diagnoses, leaving an evident gap in the context of population health management. This discrepancy forms the premise of this research endeavour. We seek to transpose concepts and developmental techniques from clinical KBS into the creation of a DSS focused on population health management. Moreover, this research probes the possibility of incremental development, necessitated by the scarcity of available data on this subject. The intended outcome of this DSS is to assist policy makers in efficiently allocating their limited resources, a pivotal step in curbing the financial and societal impact of diabetes in Australia and, subject to local validation, in comparable international settings.

It is paramount to acknowledge that the researcher, in this case, is an IT professional devoid of medical qualifications. Consequently, this research is approached purely from an IT and statistical standpoint. Collaboration with a medical expert is crucial, particularly in discerning the relevant social determinants to be examined during data collection and the anticipated outcomes of the KBS. The KBS itself will be fashioned based on the observed data patterns and the employment of diverse algorithms. The resultant product is envisioned as a policy decision-making system rather than a medical diagnostic tool.

Existing research has provided indications that social determinants bear relevance to the development of various diseases, including type 2 diabetes (Schwerdtle, 2016). Additional studies have corroborated that these determinants extend their influence beyond diabetes, impacting various other health conditions and issues (Sauliune and Kalediene, 2015); (Smith et al., 2016); (Hill et al., 2013). This study will investigate whether recurring patterns in the social determinants of Type 2 diabetes can be formalised as rules within a prototype KBS, rather than presuming in advance that such patterns necessarily exist.. Numerous tools and algorithms can be deployed for KBS development, but the current course of action has elected to utilize RDRs. This choice stems from RDRs' historical application in KBS development and the paucity of extensive data on the subject at hand. Consequently, the KB remains in continuous development mode as new insights are gleaned—a distinctive feature and value proposition of this research. While KBS and RDRs have found application in healthcare systems (Yan et al. 2003; M Kantor et al. 2011;

Rajalakshmi, Mohan & Babu 2011; Compton, Kim & Kang 2014; Peterson & Curtain 2014), their adaptation to the realm of social determinants and type 2 diabetes marks a novel endeavour. The challenge lies in identifying the specific determinants implicated in the onset of diabetes. The resulting DSS could empower policy makers to instigate proactive measures, thereby potentially ameliorating the burden of diabetes, reducing healthcare expenditure, and improving societal health.

As previously delineated in this thesis, diabetes exerts a substantial financial strain in Australia and other nations (Lee, Cohort & Magliano 2013; Sim et al. 2017). Consequently, any system that holds promise to alleviate the diabetes burden, both financially and socially, would be welcomed by nations and communities alike.

This thesis progresses by embarking on a comprehensive literature review encompassing the development of KBSs in healthcare. It subsequently delves into the complexities of data acquisition, followed by an examination of select healthcare KBSs. This research endeavours to elucidate the methodologies, techniques, and concepts that may be harnessed for the creation of a healthcare KBS with a focus on population health management. The literature review suggests the feasibility of such an endeavour, expounded upon in greater detail in subsequent sections.

In summary, this thesis provides an overview of the incremental development of a diabetes KB. It embarks on a quest to delineate the social determinants contributing to diabetes onset, scrutinises pre-existing healthcare KBSs, and assesses their suitability for this research's objectives. Furthermore, the research investigates the utility of RDRs to bridge the gap between social determinants and type 2 diabetes, culminating in the development of the envisaged KBS.

KBs conventionally emerge following the accumulation of extensive datasets. As elaborated in the subsequent literature review section, the potential association between social determinants and the onset of type 2 diabetes has been proposed. Nonetheless, this domain has witnessed relatively limited exploration, resulting in a dearth of readily available data.

This scarcity prompts a fundamental inquiry: Is it feasible to construct a KB capable of forecasting an individual's susceptibility to type 2 diabetes in circumstances characterised by data paucity? Concurrently, this study raises the query of whether such a KB can be incrementally developed, specifically during the data acquisition phase.

This section serves a dual purpose within the broader context of this thesis:

Statement of the Problem: It elucidates the core research challenge, which revolves around the development of a predictive KB under conditions of data scarcity, particularly in the context of type 2 diabetes and social determinants.

Structure of the Thesis: It outlines the organizational framework of this thesis, delineating the sequence of chapters and the thematic focus of each section.

1.2 Statement of the Problem & the Research Problem

The overarching challenge addressed in this research revolves around the intricate interplay between social determinants and the development of type 2 diabetes. It hinges on three primary facets: the existence of a causal link between social determinants and type 2 diabetes, the identification of specific determinants at play, and the endeavour to harness these social determinants for the creation of an expert system in the form of a KBS. Central to this inquiry is the quest to ascertain whether it is possible to incrementally develop a KB that can effectively leverage the knowledge contributed by a subject matter expert (SME).

This research, therefore, undertakes the pursuit of answers to the following research questions:

1. Is there a discernible relationship between social determinants and the onset of type 2 diabetes?
2. If such a relationship exists, can specific social determinants be definitively identified?
3. Can the identified social determinants be systematically employed to iteratively construct a KBS?

Addressing these questions is instrumental in the creation of a predictive tool, manifested as the resulting KB, which can estimate the likelihood of an individual developing type 2 diabetes within a particular geographic locale defined by specific social determinants. This tool, envisaged as a resource for pertinent authorities, equips them to strategize, address, and proactively mitigate the incidence of type 2 diabetes in designated geographical regions.

The objectives of this research extend along three interconnected dimensions:

Firstly, it seeks to enrich the field of KB development by pioneering an incremental approach that allows continuous refinement during the data acquisition phase. This adaptability fosters ongoing enhancements to the KB as new data and knowledge emerge. Notably, this approach is well-suited to scenarios characterised by a scarcity of data, contrasting with the data-intensive

demands of various machine learning (ML) methodologies. Furthermore, it facilitates customization by the SME to accommodate outliers or deviations from established norms.

Secondly, the research, in its trajectory, corroborates the existence of social determinants intertwined with the onset of type 2 diabetes, as initially suggested by Hill, Nielsen & Fox (2013). These findings reveal specific social determinants linked to distinct geographical regions. It is imperative to acknowledge that the list of identified social determinants presented here is not exhaustive; the domain likely harbors additional determinants yet to be unveiled. This accentuates the concept of incremental KB development, whereby the KB evolves in tandem with the discovery of novel data and knowledge.

Thirdly, the broader societal contribution lies in the formulation of a seamless knowledge base system (KBS) which serves as the interface between the SME user and the KB. This user-friendly interface mitigates the need for extensive IT proficiency, ensuring accessibility and usability for end-users. Consequently, this streamlines the KB's uptake and practical utility.

1.3 Structure of the Thesis

This thesis embarks on a comprehensive exploration of the multifaceted dimensions of type 2 diabetes, scrutinising its implications, costs, and the role of KBs in contemporary society. Through an investigation of existing literature, this research aims to unearth the limitations and possibilities of employing KBSs in the medical domain, particularly for population health planning. It is imperative to acknowledge that this research remains open to potential refinements in light of evolving insights. Nevertheless, preliminary investigations indicate that while the application of KBSs and DSSs in medical contexts is not novel, their utilization for population health planning appears underrepresented. This study endeavours to address this gap by incrementally constructing a KBS informed by social determinants to facilitate evidence-based decision making in public health.

The thesis is structured into seven chapters:

1. **Introduction**: Chapter 1, thus far, has introduced the core research questions and emphasised the contributions this study offers to both the society at large and the academic community. In this section, we provide an overview of each subsequent chapter within the thesis.

2. **Literature Review**: Chapter 2 provides an in-depth analysis of existing research related to type 2 diabetes (T2D), with a particular focus on the role of social determinants of health. It critically examines global studies linking socio-demographic factors to diabetes risk and management. Additionally, it explores the evolution of knowledge-based systems (KBS) and decision support systems (DSS) within healthcare, leading to the introduction of Ripple Down Rules (RDRs) as a novel, adaptive approach. The chapter concludes with a comparative analysis of diverse KBS and DSS development techniques employed in the clinical health domain.

3. **Methodology**: Chapter 3 introduces the design science research methodology (DSRM) employed in this research. DSRM is a structured approach that guides the creation and evaluation of innovative artefacts, in this case, the incremental development of a KB for population health planning. This section outlines how DSRM aligns with the research objectives, at each phase, and sheds light on its application to address the stated research questions.

4. **Integrating the Geographic Dimension in Knowledge-Based Development**: Chapter 4 provides a detailed account of the data acquisition, preparation, and processing that underpinned the development of the geographic knowledge-based system (KBS) for type 2 diabetes (T2D) prediction. It outlines the challenges faced in sourcing and integrating socio-demographic and clinical datasets from various organizations, including the Australian Bureau of Statistics, Diabetes Australia, and NSW Health. The chapter also describes the extensive data cleaning and manipulation techniques employed to ensure the accuracy and reliability of the datasets.

   Beyond data preparation, this chapter focuses on the incremental development of the geographic KBS using Ripple-Down Rules (RDR). It details how the training dataset was utilized to iteratively build and refine the rule base, ensuring that expert knowledge was integrated into the system to enhance predictive accuracy. The chapter also discusses how the geographic KBS was validated using the production (unseen) dataset, highlighting the model's adaptability and its initial accuracy outcomes.

   This chapter establishes the foundation for understanding how geographic attributes influence diabetes risk prediction and sets the stage for the comparative analysis with the socio-demographic KBS in later chapters.

5. **Socio-Demographic Knowledge Based System:** Chapter 5 details the methodology for incrementally developing the socio-demographic knowledge-based system (KBS) using Ripple-Down Rules (RDRs). Building on the insights gained from the geographic KBS in Chapter 4, this chapter outlines the process of constructing rules based on socio-demographic factors associated with Type 2 diabetes risk. A key focus is the iterative refinement of the system, incorporating expert knowledge to improve interpretability and accuracy.

   This chapter describes the various stages of knowledge acquisition, from rule formulation to validation, ensuring that the system remains adaptable as new information emerges. It explores the challenges of integrating socio-demographic variables into predictive models, addressing issues such as data sparsity, regional variability, and generalizability across different populations. Additionally, the chapter discusses the validation process, detailing how accuracy rates were assessed for both training and production datasets.

   By demonstrating the practical application of RDRs in developing an expert-driven KBS, this chapter establishes the foundation for evaluating the system's effectiveness in Chapter 6, where comparative analyses between RDR-based and ML-based models are conducted.

6. **Geographic & Socio-Demographic Production and ML Comparisons:** Chapter 6 presents the core findings of this research, evaluating the effectiveness of Ripple-Down Rules (RDRs) in the development of a knowledge-based system (KBS) for type 2 diabetes (T2D) prediction. The chapter provides a detailed analysis of the experimental results and offers a comparative evaluation of different predictive models, particularly the incremental RDR-based KBS and traditional machine learning (ML) approaches. The first part of this chapter focuses on the performance evaluation of the geographic knowledge-based system and socio-demographic KBS models. It examines how well these systems generalize when applied to unseen data by comparing accuracy rates between the training dataset and the production dataset. The results demonstrate that socio-demographic attributes, such as income levels, education, and healthcare accessibility, play a significant role in predicting T2D risk. The geographic KBS and socio-demographic KBS were tested under different conditions, confirming the ability of RDRs to develop an adaptive and interpretable decision-support system. Statistical analyses reveal that the socio-demographic KBS consistently outperforms the knowledge-based system KBS when applied to the production dataset, reinforcing the broader applicability of socio-demographic attributes in public health modelling.

The second part of the chapter shifts to a comparative analysis of RDR-based models and ML-based models, evaluating their respective advantages and limitations. Unlike ML models, which require large and structured datasets, RDRs excel in data-limited environments by allowing incremental expert-driven updates without retraining the entire system. The analysis highlights the interpretability gap between the two approaches, as ML models, such as J48 decision trees, function as black-box algorithms, whereas RDRs provide transparent, rule-based decision-making. Moreover, RDR-based KBS models demonstrate superior adaptability, enabling continuous refinement by integrating newly discovered risk factors, which is particularly valuable in healthcare domains where knowledge evolves over time.

A critical takeaway from this chapter is the demonstration that RDR-based models are more practical and scalable for real-world healthcare applications, particularly in low-resource settings where extensive labelled datasets may not be available. The results confirm that the incremental, expert-driven nature of RDRs makes them a more viable and interpretable alternative to ML models in the development of decision-support systems for chronic disease prediction. This chapter lays the foundation for a broader discussion in Chapter 7, where the findings are contextualized within the existing literature and future research directions are explored.

7. **Summary, Discussion and Future Work**: Chapter 7 synthesizes the findings of this research, examining their broader implications in healthcare and beyond. The chapter begins by reiterating the significance of socio-demographic and geographic knowledge-based system factors in predicting type 2 diabetes (T2D) and evaluating the effectiveness of Ripple-Down Rules (RDR) in constructing an adaptive knowledge-based System (KBS). The research findings emphasize the superiority of RDRs over conventional machine learning (ML) models, particularly in data-scarce environments where expert-driven rule refinement ensures dynamic system evolution.

   The chapter critically analyses how integrating socio-demographic insights into decision-support tools can enhance public health interventions and clinical decision-making. The adaptability of RDR-driven KBS models allows for real-time updates and knowledge injection by domain experts, making them highly effective for healthcare planning and resource allocation.

   Beyond healthcare, the research highlights the potential applications of RDRs in other industries such as insurance, finance, and risk management, where incremental rule

development is crucial for accurate decision-making. A key component of this chapter is the discussion of future research directions. This study suggests further exploration of hybrid models that integrate RDRs with ML techniques to optimize predictive accuracy while maintaining interpretability. Additionally, the chapter outlines potential advancements in scaling KBS models for big data environments and enhancing real-time decision-support capabilities across various domains.

The chapter concludes with final remarks on the contributions of this research. It underscores the importance of acknowledging socio-demographic determinants in disease prediction and decision-support system development, positioning RDRs as a versatile tool for dynamic, knowledge-driven AI applications. These insights pave the way for continued exploration into knowledge-based AI methodologies, with implications extending beyond healthcare to broader fields requiring expert-driven, adaptable intelligence.

Chapter 2: Literature Review

This chapter provides a comprehensive review of the literature relevant to the research topic, focusing on the economic burden of diabetes across different regions, including Australia, the United States, Europe, China, and South Africa. It examines various studies and reports that highlight the growing costs associated with diabetes management and complications, and the implications of these costs on healthcare systems and societies globally. The chapter also discusses the importance of effective diabetes management strategies to mitigate these economic impacts and improve patient outcomes.

The chapter is divided into five sections. Section 2.1 discusses the costand impact of diabetes by examining the significant economic and social burdens of diabetes on a global scale and within Australia, highlighting the urgency for effective management and intervention strategies. Section 2.2, investigates the influence of various factors, such as income and education, on the prevalence of type 2 diabetes, advocating for interventions at the policy level to address these root causes. Section 2.3, discusses the utility of KBSs in healthcare, illustrating through prior research how a well-constructed KBS can enhance diabetes management by optimising resource use. Section 2.4 introduces RDR as a technique for developing an adaptive, incrementally expanding knowledge base, especially effective in situations where data on social determinants are limited. Finally, Section 2.5 concludes with a summary of the literature review, encapsulating the chapter's exploration of the economic and social ramifications of diabetes, the critical role of social determinants, and the promising application of DSS and RDR in formulating a cohesive strategy for diabetes management.

Building a type 2 diabetes knowledge base provides an in-depth background on the incremental development of a KBS, particularly focusing on the application of RDRs in healthcare. This section outlines the historical use of RDRs in developing adaptive KBSs within the healthcare sector, showcasing their effectiveness in managing complex data environments where traditional machine learning techniques might fall short. The novelty of this research lies in its approach to KBS development during the data acquisition phase, allowing for continuous refinement and adaptation based on newly acquired data. This methodology ensures that the KBS evolves dynamically, incorporating new insights and improving its predictive accuracy over time. This section also reviews previous applications of KBSs in healthcare, highlighting their contributions to clinical decision support and patient management, and setting the stage for the innovative aspects of this research.

Incremental development of KBSs has been a topic of interest in recent years. Researchers have explored various approaches to building and refining knowledge bases over time. One such approach is the use of RDRs, which has shown promise in healthcare applications (Compton and Kang, 2021). RDR is a knowledge representation framework that allows for the incremental development and refinement of rules (Compton et al. 2006). It has been applied in various healthcare domains, including disease diagnosis (An et al., 2023) and treatment management (Alrige et al., 2023). For example, a study by Bindof et al. (2014) developed an RDR-based system to detect problems induced by drugs (Bindoff et al., 2014a). Another study by Prama et al. (2023) used RDR to develop a knowledge base for diagnosing COVID 19 (Pratama et al., 2023). RDRs have been particularly useful in healthcare due to their ability to handle exceptions and nuances in medical data, which are often challenging for conventional machine learning models. For instance, in the domain of pathology, RDRs have been used to incrementally build knowledge bases that assist pathologists in diagnosing complex cases by providing justifications for each decision made (Compton et al., 2016). Another recent example of an RDR-based system for managing chronic diseases like diabetes has been shown to improve patient outcomes by providing personalized treatment recommendations based on an individual's unique set of social determinants (Johnson et al., 2020). Another study by Lee et al. (2019) highlighted the use of RDRs in developing a KBS for diabetes management, which helped clinicians make more informed decisions by integrating real-time patient data with historical trends and medical knowledge.

The RDR approach not only enhances the accuracy of diagnoses but also ensures that the knowledge base remains current as new medical information becomes available. In addition to RDR, other approaches to incremental knowledge base development have been explored. For instance, machine learning algorithms have been used to identify health disease state and immune response etc. (Habehh and Gohel, 2021). Ontologies have also been employed to represent and integrate knowledge from various sectors (Psyllidis, 2015). For developing a KBS for type 2 diabetes risk assessment, the key target of this thesis, it is prudent to concentrate on incrementally develop a type 2 diabetes KBS. It is clear that such a KBS would need to be flexible enough to accommodate new determinants that could be established in the future. As this task has not been yet attempted. More specifically, Ripple Down Rules (RDR) based approach will be used to develop the KBS given the limited data as this thesis will later highlight. As this review will also highlight, that this approach in Diabetes Type II will also respond to a number of gaps in the research literature. The following few sections outline some articles used and the justification for their usage.

Hence, the decision was taken to concentrate on incrementally developing a type 2 diabetes KBS. The KBS would need to be flexible enough to accommodate new determinates that could be established in the future; hence it was decided to use RDRs to develop the KBS.

The research focus, then turned towards which KBSs that are already used in healthcare and the gaps. The following sections outline some of the articles selected for this review and the justification for their inclusion.

Recent applications of RDRs in healthcare have demonstrated their potential to revolutionize the way medical knowledge is captured, refined, and utilised. For example, an RDR-based system for managing chronic diseases like diabetes has been shown to improve patient outcomes by providing personalised treatment recommendations based on an individual's unique set of social determinants (Johnson et al. 2020). Another study by Lee et al. (2019) highlighted the use of RDRs in developing a decision support system for diabetes management, which helped clinicians make more informed decisions by integrating real-time patient data with historical trends and medical knowledge.

Sim et al. (2017) developed an RDR-based knowledge-based system for antimicrobial susceptibility testing in a tertiary hospital laboratory. The study showed that incremental rule acquisition reduced turnaround time and improved agreement with expert decisions, illustrating how domain knowledge can be captured, refined, and audited in real-time. The methodological parallels—incremental rule induction, integration with existing workflows, and emphasis on interpretability—make Sim et al.'s approach a relevant comparator for the present population-health KB.

Schwerdtle (2016) is a practice-oriented commentary published in the Australian Nursing & Midwifery Journal (ANMJ), the official publication of the Australian Nursing & Midwifery Federation; it is not a peer-reviewed research article and is cited here for its practitioner perspective on policy-level responses to social determinants of health. The specific propositions it advances that SDoH should be addressed at a policy level and are central to diabetes prevention/management are corroborated by peer-reviewed and official sources *(e.g., Hill-Briggs et al., 2021; WHO, 2024; Australian National Diabetes Strategy 2021–2030; AIHW, 2024)*. At the time of publication, the author was a lecturer at Monash University; subsequent outputs in global health further support her subject expertise. she has since developed a substantive academic track record in global health *(ORCID 0000-0002-3045-3145; Google Scholar h-index ≈23)* These corroborating sources justify citing the ANMJ commentary as a practice-sector

perspective, while the empirical evidence base is drawn from the peer-reviewed and official literature cited throughout this section.

2.1 Cost & Impact of Diabetes

In 2011, diabetes affected 366 million people globally, a figure that was projected to rise significantly in the following decades (Sim et al., 2017). As of 2021, this number has grown dramatically, with diabetes now impacting over 537 million people worldwide (IDF 2021). This represents about 10.5% of the global adult population. The number of people with diabetes is expected to rise significantly, reaching an estimated 643 million by 2030 and 783 million by 2045 (IDF 2021).

In Australia, the cost of diabetes has seen a significant increased from $3.0 billion in 2018–19 to $3.4 billion in 2020–21 (AIHW 2024), about a 13% rise. That represents approximately 2.3% of the total expenditure in both years (AIHW 2024). As of 2021, the annual economic burden of diabetes on the Australian healthcare system is estimated to be around $3.4 billion (Health and Welfare, 2024). This amount represents roughly 1.5 % of Australia's total health expenditure in 2021–22 ($221 billion), ranking diabetes seventh among chronic-disease spending; by comparison, cardiovascular disease accounted for $11.8 billion (5.3 %) and musculoskeletal conditions $12.5 billion (5.7 %) (AIHW 2023). This figure highlights the substantial financial strain that diabetes imposes on the healthcare system, with projections suggesting that costs could soar to $45 billion per year by 2050 due to rising prevalence and associated complications (Health and Welfare, 2024). These numbers indicate that a considerable portion of Australia's health expenditure is dedicated to managing diabetes, reinforcing the importance of developing better strategies for its management and prevention.

The financial costs of diabetes are one social impact of diabetes. However, there are others, some of which will be highlighted during this research. These include but are not limited to, lifestyle constraints, dietary constraints and work prospect constraints etc. because of diabetes. It is also important to note that this research in no way implies that it is working towards a solution to the total elimination of diabetes. It simply explores ways of better dealing with and managing the disease.

In the United States, the economic burden of diabetes is even more staggering. In 2022, diabetes-related costs reached $413 billion, which includes direct medical costs and indirect costs from

lost productivity (Parker et al., 2023).This represents a significant increase from previous years, underscoring the rapidly growing financial impact of diabetes in the U.S.

In Europe, the economic impact of diabetes is also significant. A study by the International Diabetes Federation (IDF) reported that the total healthcare expenditure on diabetes in Europe was approximately €166 billion in 2019 (Williams et al., 2020). This includes direct medical costs such as hospital admissions, medications, and outpatient care, as well as indirect costs related to loss of productivity and early retirement. In China, the cost of diabetes has also surged. The total annual cost of diabetes was estimated at $167.5 billion as of the most recent assessments (Liu et al., 2023). This figure reflects the significant financial burden diabetes places on the healthcare system in China, driven by the high and rising prevalence of the disease

In Africa, the cost of diabetes is also substantial, although data is less comprehensive. A study in South Africa estimated that the total cost of diabetes in 2015 was approximately $2.7 billion, including both direct and indirect costs (Mutyambizi et al., 2018). A more recent study in South Africa estimated that the total cost of diabetes as of 2021 is approximately $3.5 billion, which includes both direct costs (such as hospital care and medication) and indirect costs (like lost productivity and disability (Basu et al., 2021).The economic burden of diabetes across Africa is likely still underestimated due to limited healthcare resources, underreporting, and challenges in gathering accurate data in many regions. Research shows that, as the prevalence of diabetes continues to rise, particularly in urban areas, the financial strain on healthcare systems in African countries, including South Africa, is expected to escalate further (Basu et al., 2021). The economic burden of diabetes in Africa is likely underestimated due to limited healthcare resources and underreporting.

Because national estimates use differing methods and price bases, the figures below are interpreted as within-country burden; cross-country comparisons are avoided unless expressed on a common denominator such as a percentage of total health expenditure. Country-specific estimates are informative for domestic policy; cross-country claims require normalised denominators and are therefore not made in this thesis.

These global estimates highlight the immense economic impact of diabetes and the urgent need for effective management and intervention strategies. Addressing the social determinants of health and improving access to care are critical components in reducing the financial burden of diabetes worldwide. Indeed, the financial costs of diabetes are only one social impact of diabetes. There are others which include lifestyle constraints, dietary constraints and employment prospect of sufferers. It can also reduce educational attainment potential (Hill et al., 2013). This research

does not seek a solution to the total elimination of diabetes. It will rather facilitate impact and public health risk analysis which will in turn pave the way for policymakers to manage resources to better deal with and manage the disease.

The implementation of a robust KBS could significantly alleviate the financial burden of diabetes by improving disease management and prevention strategies. By identifying high-risk populations and optimising resource allocation, the KBS could help reduce the incidence of diabetes-related complications, which are often the most costly aspect of the disease (Williams et al., 2020). In Australia, where the healthcare system is already under significant financial strain (AIHW, 2024), the cost savings from a KBS could be substantial. These savings could be redirected towards other critical areas of healthcare, such as mental health services, aged care, and preventive health programs. For instance, improving mental health services could help address the increasing rates of depression and anxiety, which are often co-morbid with chronic diseases like diabetes (Morgan et al., 2021).

Additionally, enhancing aged care services is crucial as the population ages. Redirecting funds saved from better diabetes management could improve the quality of care for the elderly, ensuring they receive timely and appropriate medical attention (Lee et al., 2018). Preventive health programs focusing on lifestyle modifications, such as promoting physical activity and healthy eating, could further reduce the prevalence of chronic diseases, creating a healthier population and reducing long-term healthcare costs (Bull et al., 2020).

Overall, the implementation of a KBS for diabetes management not only has the potential to reduce the direct costs associated with the disease but also offers broader economic benefits by improving overall public health and reallocating resources to other essential healthcare services.

2.2 Social Determinants

There have been many articles outlining how social determinants play a role in the onset of Type II diabetes. These include Hill, Nielsen & Fox (2013), Sauliune and Kalediene (2015) and Schwerdtle (2016) to name a few. However, the question of how and which social determinants combine to expedite or contribute to the onset of diabetes type 2 is open for conjecture. Social determinants such as income, education, housing and nutritious food have been suggested as possible candidates (Hill et al., 2013). However, it is difficult to pin the causal chain. As earlier discussed, Type 2 diabetes itself can cause missed employment opportunities, decrease

productivity and reduce educational attainment potential. The prevailing view is that those social determinants should be addressed at policy levels (Schwerdtle, 2016). Schwerdtle (2016) suggests a strong emphasis on education in conjunction with strategies to promote healthier lifestyle, incorporating more exercises and healthier food etc. Schwerdtle (2016) also suggests that nurses are well placed to advocate for these changes to occur.

Hill, Nielsen & Fox (2013) propose a sociobiological cycle of diabetes (Figure 2.1). This cyclic process both results in and contributes to adverse outcomes. If some of these issues could be addressed, it may lead to a slowing of the progression rate of diabetes. However, policymakers and their advisors don't have the time to go and extract this kind of data from the various sources to construct some useful information. Therefore, it stands to reason that a system that can manipulate the various data from the various sources to produce useful and accurate information would be a handy tool for policymakers and their advisors.

Hence, this thesis purposes of a Decision Support System (DSS) built on an incrementally developed KBS.

As has been stated so far, there is extensive literature on social determinants and Type 2 diabetes; the remaining need is to operationalise these determinants in computable, interpretable decision-support systems, as suggested by Hill, Nielsen & Fox (2013) and Schwerdtle (2016). It therefore stands to reason that the KBS that would need to be developed incrementally with a capacity for future growth to accommodate more social determinants associated with type 2 that are established in the future.

Figure 2.1 shows that responses to type 2 diabetes are predominantly clinical (biological and psychological). It also points to the need for stronger social responses, which motivates this thesis's emphasis on socio-demographic determinants.

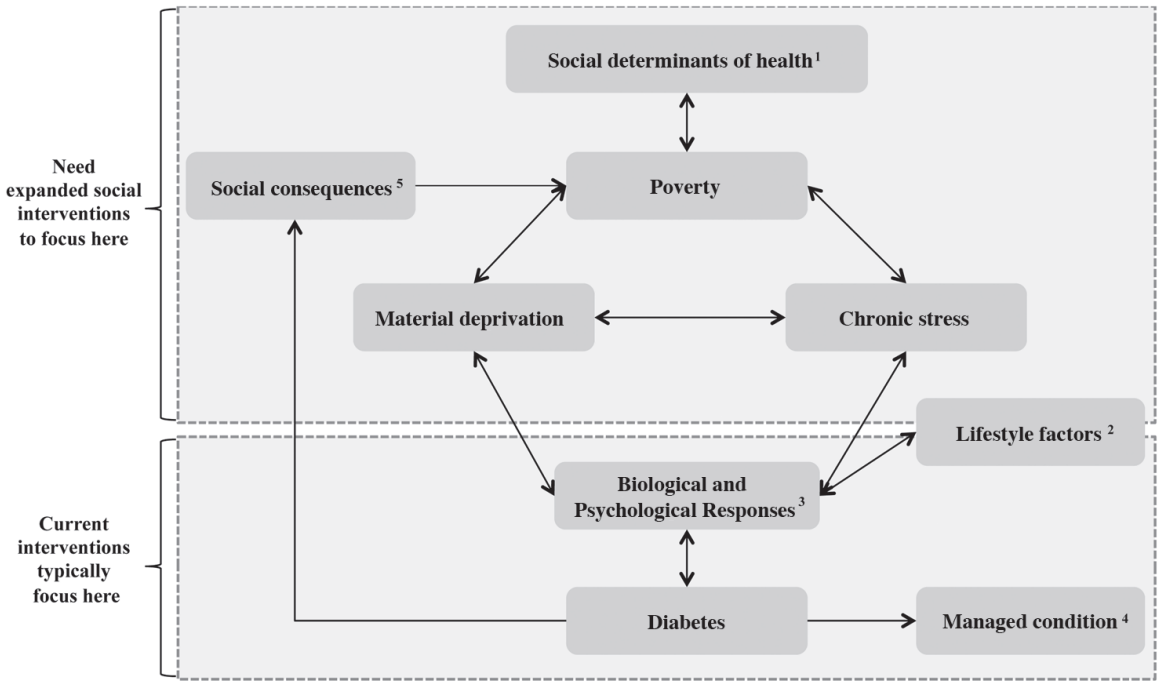

Figure 2. 1 The sociobiological cycle of diabetes. Source: Hill, Nielsen & Fox (2013)

Recent studies continue to support the significant role that social determinants play in the development and management of type 2 diabetes. According to the World Health Organization (WHO), social determinants such as income inequality, education, and access to healthcare are critical factors influencing the prevalence of diabetes (Hill-Briggs & Fitzpatrick 2023). Income inequality can lead to disparities in access to nutritious food and healthcare services, which are essential for diabetes prevention and management.

In the United States, research has shown that individuals from a lower socioeconomic status are more likely to develop type 2 diabetes (Walker et al., 2021). This is due to a combination of factors including limited access to healthy foods, lack of safe spaces for physical activity, and higher levels of stress. Policies aimed at reducing these disparities, such as improving access to affordable healthy foods and creating safe environments for exercise, could help reduce the incidence of diabetes.

In Europe, large multi-country cohort evidence shows a consistent inverse association between educational attainment and incident type 2 diabetes. In the EPIC-InterAct case-cohort (8 countries), higher education predicted lower diabetes risk, and recent Mendelian-randomisation within the same consortium supports a causal protective effect of education on diabetes incidence. Mechanistically, education appears to operate partly via health literacy and related behavioural

pathways documented in European cohorts. *(EPIC-InterAct; JECH 2025; Steele et al., 2017).* In India, rapid urbanisation and lifestyle change are linked to higher diabetes prevalence in urban than rural settings, with substantial state-level heterogeneity. Earlier ICMR–INDIAB rounds (15 states) reported higher urban risk; more recent nationally representative estimates (covering all states and union territories) confirm a high and uneven burden and document major gaps across the care continuum. These findings are specific to India and are not generalised to Asia as a whole. *(Anjana et al., 2017; ICMR–INDIAB national report 2023; JAMA Intern Med 2023—care continuum).*

In sub-Saharan Africa, limited access to diagnostics, medicines and chronic-care follow-up continues to exacerbate diabetes burden. Recent evidence shows that integrated chronic-care models (combining HIV, diabetes and hypertension care) can improve retention without compromising HIV outcomes, supporting calls to integrate NCD services into existing platforms; scoping and systematic reviews report similar feasibility and service benefits. *(Lancet 2023 randomised trial; PLOS One 2023 scoping review; BMJ Glob/Family Med & Comm Health 2021–2022 reviews).*

These findings underscore the importance of addressing social determinants in the fight against diabetes. By focusing on the broader social and economic context in which individuals live, policymakers can develop more effective strategies to prevent and manage diabetes. This holistic approach not only improves health outcomes but also reduces the long-term economic burden of the disease.

So far, the discussion has focused on establishing and utilising the social determinants associated with type 2 diabetes. This gives rise to the question that once the social determinants have been established and utilised to develop a KBS (described in the next section), how is the output used?

Ramaprasad et al. (2016) discuss the development of an ontological mapping of Australia's National Health Programs. The health programs discussed are mapped onto an ontology with five dimensions. The five dimensions are policy-scope, policy-focus, outcomes, type of care and population served (Ramaprasad et al., 2016). Since there is an emphasis on policy, it would make sense that a tool is required to justify such policy, hence, the purpose of this research and the development of a KBS. Ramaprasad et al. (2016) discuss traditional policy analysis from various perspectives. These perspectives include policy makers, town planners, medical insurance companies, government bodies, doctors and allied staff etc. Ramaprasad et al. (2016) also discusses the domain in which these perspectives operate. These domains include communicable diseases, maternal and child health, and geographical location (Ramaprasad et

al., 2016). The geographical location domain gives some credibility to social determinants being associated with patient location (WHO 2025). However, the article does not discuss how to obtain this information and present it in a usable manner. Hence, the proposed KBS would be developed incrementally. The proposed KBS would fill this gap. The incrementally developed KBS would ensure that it maintains currency as new social determinants associated with type 2 diabetes are established. From a public health informatics perspective, the challenge is not establishing the social determinants (Hill, Nielsen & Fox, 2013), but representing and integrating them in computable, governance-aware form for fair and transparent decision support—precisely what the KBS architecture enables.

2.3 Knowledge-Based Systems (KBS)

As discussed in the earlier sections of this chapter, there is ample research to support the idea of utilising social determinants to address and manage diabetes type 2 around the globe. This section pursues the question of using these social determinants to aid in the actual development of a knowledge base system to support those management decisions. Such a system can be deployed in assisting policy makers how best to maximise the use of available resources. However, there is a need to use these determinants to aid in the development of a KBS. The KBS could then be used in assisting policy makers how best to maximise the use of available resources.

KBSs have previously been used in healthcare, as reported by Bindof, Peterson and Curtain (2014) Their systems required minimal training and could identify approximately 90% medication-related patient problems with an error reduction of 2.02 per patient (Bindoff et al., 2014b). Sim et al. (2017) conducted a pilot study where they developed a clinical decision support system for diabetes care. Although this pilot study developing a DSS is from a clinical perspective, many concepts could be used to develop a DSS from a population health perspective. For example, figure 2 below shows a series of box plots various findings from the Sim et al. (2017) study. In section C. it discusses the "ability to identify the long-term trends of results of the markers in the glycaemic panel" (Sim et al. 2017). In this research, this could be adapted to look at the long-term trend current type 2 diabetes patients using the various social determinants. Examples of such determinants could include dietary habits, education level, place of birth and current place of residence etc.

The effectiveness of Decision Support Systems (DSS) is inherently tied to the robustness and comprehensiveness of the Knowledge-Based Systems (KBS) they rely upon. A DSS is designed to support decision-making processes by providing relevant information and recommendations based on data analysis. However, the value of these recommendations depends significantly on the underlying KBS that supplies the necessary knowledge and contextual understanding. Without a well-structured KBS, a DSS would lack the foundational knowledge required to generate accurate and reliable outputs.

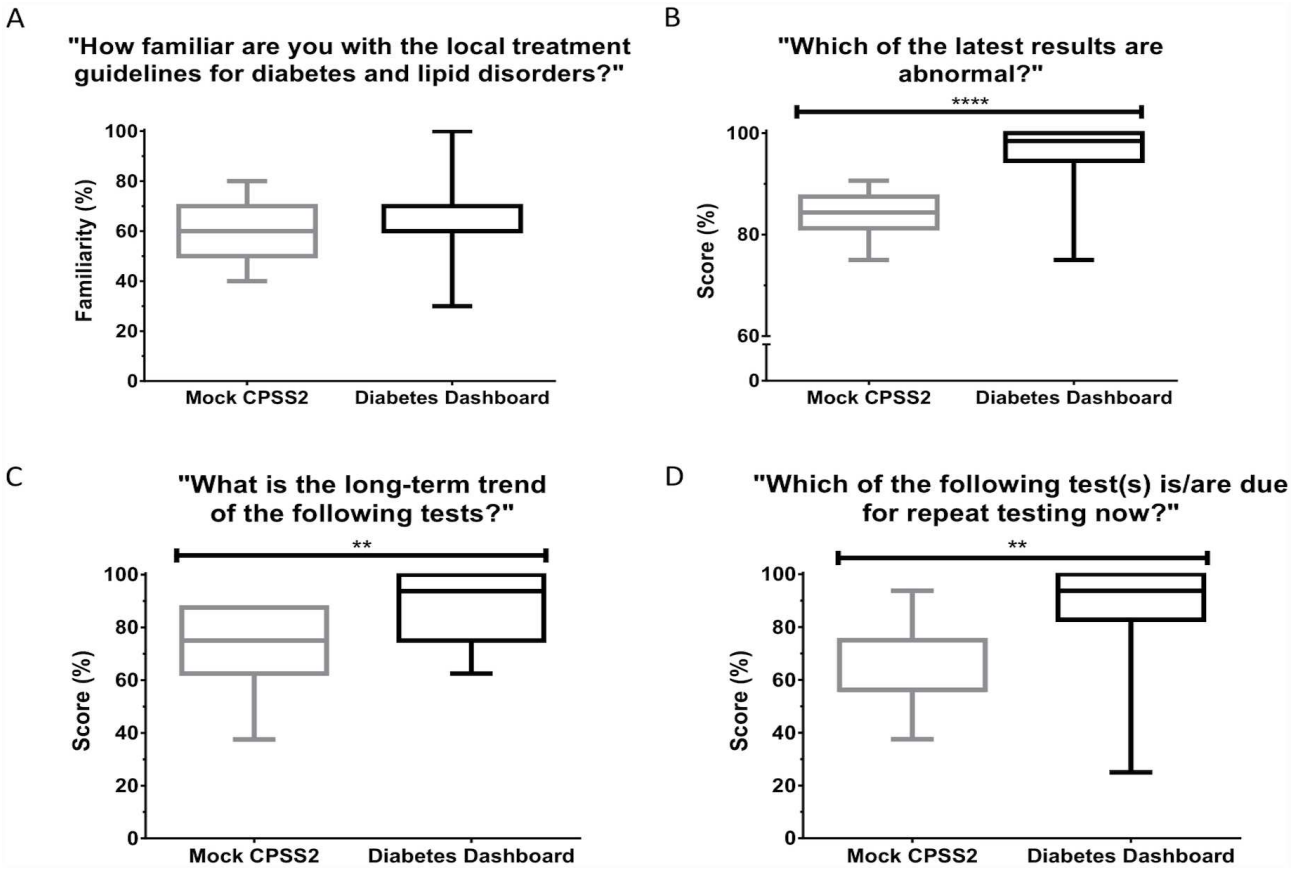

Figure 2. 2 Box plots showing the participants' Source: Sim et al. (2017)

Figure 2.2 indicates substantial dispersion across groups: medians differ, interquartile ranges are wide, and several outliers are present. This pattern suggests real heterogeneity rather than measurement noise. For this thesis, such variability justifies using an interpretable, rule-based approach that sets attribute-specific thresholds and can be refined incrementally (RDR), rather than relying on a single global cut-off.

Figure 2.2 presents four box plots based on survey responses assessing participants' performance and confidence in interpreting diabetes-related data. Plot (A) shows perceived familiarity with diabetes-related guidelines. Plot (B) illustrates participants' ability to identify abnormal values in recent glycaemic panel results. Plot (C) evaluates their ability to recognize long-term trends in glycaemic markers. Finally, plot (D) assesses their judgment on whether HbA1c or LDL tests required retesting. Each plot corresponds to a specific survey question, and significance levels are denoted by asterisks (for $p<0.05$; **** for $p<0.0001$) (Sim et al., 2017).

KBSs serve as repositories of domain-specific knowledge, incorporating expert insights, empirical data, and evidence-based guidelines. They enable KBSs to interpret complex datasets, identify patterns, and derive meaningful conclusions. In the healthcare sector, for instance, a KBS can integrate vast amounts of medical research, clinical guidelines, and patient data, which it then uses to assist healthcare professionals in making informed decisions about diagnosis, treatment, and patient management (Hurtubise et al., 2016).

The interdependence between DSSs and KBSs can be illustrated through various examples. In clinical decision support, KBSs provide the structured knowledge that helps DSSs evaluate patient symptoms, medical histories, and laboratory results. This evaluation process is crucial for diagnosing conditions and recommending appropriate interventions. Without the foundational knowledge provided by a KBS, the DSS would be unable to process this information effectively, leading to potential errors and suboptimal patient care (Greenes et al., 2018).

Moreover, the continuous development and updating of KBSs are essential for maintaining the relevance and accuracy of quality management that rely on them. As new medical research and clinical guidelines emerge, the KBS must be updated to reflect these advancements. This ensures that quality management remains a reliable tool for healthcare professionals, providing the most current and evidence-based recommendations (Khamisi et al., 2019).

In the context of public health, the relationship between KBSs and DSSs is equally critical. Public health DSSs utilize KBSs to analyse epidemiological data, monitor disease outbreaks, and develop prevention strategies. For example, during the COVID-19 pandemic, KBSs containing up-to-date information on virus transmission, symptoms, and treatment options were integral to the functioning of DSSs used by public health authorities to manage the crisis (Benke et al., 2020).

Similarly, a study by Hamedan et al. (2020) developed a DSS for chronic kidney disease management. This DSS relied on a comprehensive KBS that integrated clinical guidelines, patient data, and expert knowledge. The system demonstrated significant improvements in

patient outcomes, highlighting the critical role of KBS in enhancing the effectiveness of DSS. Without the KBS, the DSS would not have been able to provide personalised treatment recommendations accurately (Hamedan et al., 2020).

Another example is the research conducted by Rani et al. (2021) on a DSS for managing heart disease. This DSS was underpinned by a robust KBS that included vast amounts of cardiology research, patient histories, and treatment protocols. The KBS enabled the DSS to analyse complex cardiac data and offer precise diagnostic and treatment suggestions. The study concluded that the success of the DSS was heavily dependent on the quality and comprehensiveness of the KBS (Rani et al., 2021)

Additionally, a study by Walsh et al. (2019) focused on a DSS for managing cancer treatment. The DSS utilised a KBS that integrated oncological research, clinical trial data, and patient records. This integration allowed the DSS to offer personalised treatment plans and predict patient responses to various therapies. The researchers emphasised that without the underlying KBS, the DSS would lack the necessary depth of knowledge to make accurate and effective recommendations (Walsh et al., 2019).

Furthermore, the integration of KBSs in DSSs enhances the system's ability to handle complex decision-making scenarios. This integration allows DSSs to simulate various scenarios, predict outcomes, and evaluate the potential impact of different intervention strategies. In the case of diabetes management, a DSS equipped with a robust KBS can analyse patient data to predict the likelihood of complications, suggest personalised treatment plans, and recommend lifestyle changes to mitigate risk factors. RDRs also provide the opportunity for the subject matter expert (SME) to inject their own knowledge and experience into the knowledge base (Compton and Kang, 2021).

The above discussions and associated articles support the development of a DSS for the improved management of type 2 diabetes. The next question is how to develop this DSS? More specifically, which tool(s) best suit this development? This is where RDR come into play. The next section examines how RDRs could be used for this purpose along with supporting evidence.

2.4 Ripple-Down Rules (RDR)

RDR is a knowledge acquisition methodology designed for the incremental development and maintenance of knowledge-based systems. Developed by Compton and Jansen in the late 1980s,

RDRs address the challenges associated with the initial development and subsequent refinement of knowledge bases (Compton & Jansen, 1988). The primary advantage of RDRs lies in its simplicity and efficiency, allowing domain experts to add or modify rules directly without requiring the intervention of knowledge engineers. This results in a more flexible and adaptable system capable of evolving over time.

Compton and Richards (2000) define RDRs as the incremental development of a KBS. The reason for developing RDRs was so that experts did not have to supply comprehensive reasoning for their decisions in particular situations (Compton and Richards, 2000). That is, the decision hasn't been justified in any way, apart from the experts' opinion. A system developed using RDRs would provide such a justification (Compton & Jansen, 1990; Compton & Richards, 2000; Beydoun & Hoffman, 2013).

Apart from the justifications, RDRs also allow writing rules for exception cases. That is, cases which may not necessarily conform to the standard KBS developed. Using various types of machine language (ML) techniques, these cases would be put down as exceptions to the norm. However, RDRs provide a tool to write rules to accommodate these so-called exception cases. Because exception lists and per-rule audit logs accumulate with each misclassified case, they become increasingly informative over time—surfacing recurring failure patterns and guiding targeted rule refinement as the KBS evolves.

RDRs also allow an expert in a particular field to use their time more effectively. It does this by allowing a less experienced person to make an initial diagnosis without involving the expert. For example, consider the classic contact-lens case described by Compton & Jansen (1990). In this example, the task is to determine whether a person requires contact lenses and, if so, whether hard or soft lenses are appropriate. Ordinarily, an optometrist would need to make such a decision. That is because they would need to know what terminology such as "presbyopia" actually means. However, using RDRs in a KBS allows anybody to make that kind of decision and have a system's support behind that decision. This would also allow the optometrist more time to concentrate on other pressing issues such as retinal photography screening services and ophthalmology clinics. However, the optometrist (the expert in the field) needs to develop and update the RDRs used in the KBS.

Normally, constructing a KBS involves the RDRs being developed on the basis of existing data rather than an expert's knowledge (Hyeon et al., 2016). This is referred to as Induct RDR (Hyeon et al., 2016). However, there is still a need for an expert's input. Hyeon et al. (2016) discuss the various techniques and methodologies used to produce an Induct RDR KBS. These

issues will become salient as larger, multi-site datasets and external validation cohorts become available; at present, there is sufficient evidence that RDRs are suitable for medical KBS construction, which motivates their use here. This is because the exception list accumulates across cohorts, each resolved exception either generalises into a new rule or records a boundary condition, so coverage and accuracy improve on future datasets and new regions (Compton, Kim & Kang 2014).

In an RDR KBS, rules are added only in the context of their desired application (Compton, Kim & Kang 2014). Rules are added to satisfy a case for which the original sequence of rules failed, excluding cases covered by its predecessor rule (Beydoun and Hoffmann, 2013; Compton, Kim & Kang, 2014). Rules are never removed or modified because the corrected case is actually in the newest added rule (Beydoun and Hoffmann, 2013; Compton, Kim & Kang, 2014). Due to the way conditions of new rules are added, a correction made by the expert is guaranteed to be valid (Beydoun and Hoffmann, 2013; Compton, Kim & Kang, 2014). Should an expert disagree with a knowledge base conclusion, then the knowledge base has failed and requires modification, hence, the proposed monitoring during development concept. This modification can be carried out directly by the expert due to the simplicity of its modification (Beydoun and Hoffmann, 2013; Compton, Kim & Kang, 2014). There are two main features of RDRs which lead to the simplicity of its modification. Firstly, the cause of failure is automatically determined due to the knowledge base's tree-like structure. That is, a new rule is added to the leaf node and is attached to the last visited rule prior to the knowledge base's failure (Beydoun and Hoffmann, 2013; Compton, Kim & Kang, 2014). Secondly, the knowledge base's framework ensures that every newly added rule is consistent with its corresponding new case, without creating any inconsistencies with previously classified cases. That is, each new rule added is justified for a case classified by the expert (Beydoun and Hoffmann, 2013; Compton, Kim & Kang, 2014).

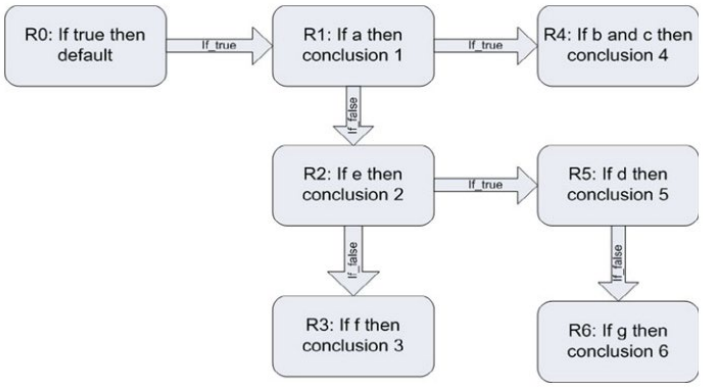

Figure 2. 3 A classification RDR tree. A case to be classified starts at the root default node and ripple down to a leaf node. Source: Beydoun and Hoffmann (2013)

Accordingly, we now move from concept to implementation. As our required data tends to come in subsets where a uniform distribution is likely, i.e. assume two different people randomly selected in any given postcode will have the same probability of having type 2 diabetes, hence, there is coverage but, very limited density and volume (ABS 2021 SEIFA Technical Paper). The coverage is provided by the expert knowledge of the individual(s) developing each RDR (see Figure 2.4).

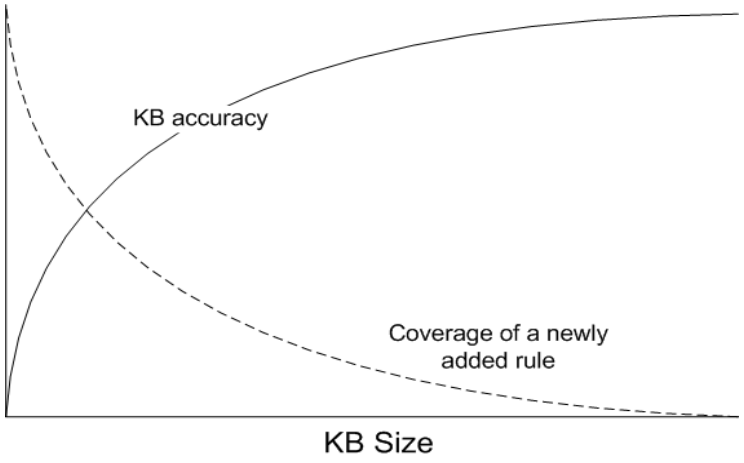

Figure 2. 4 RDR Convergence; added accuracy is diminished with size. The number of instances that individual rules classify drops quickly as the knowledge base converges. Source: Beydoun and Hoffmann (2000, 2001)

Predictive analysis techniques normally allow for discrepancies in conclusions based on datasets without offering any real explanation on the reason. For example, if you had a dataset said indicated a person over 50 years of age, born in the Middle East with a family history of type 2 diabetes and the conclusion of that person is a diabetic and in the same dataset, along comes a person with the same social determinants and the conclusion is that that person is not a diabetic. Most predictive analysis techniques would put that down to discrepancies in the dataset and often ignore it. However, the very nature of RDRs would force the subject matter expert to examine other social determinants to establish why this discrepancy occurred and develop a new rule to account for it.

Recent research continues to highlight the effectiveness of RDRs in healthcare. A study by Hyeon et al. (2016) & Byeong (2018) demonstrated the application of RDRs in developing a clinical decision support system which accounts for the complexities of diseases and various symptoms. Essentially, extracting an expert's knowledge into a KBS. The system achieved an impressive improvement in accuracy rate to 70%, emphasising the utility of RDRs in handling complex medical data and providing reliable recommendations. This study underscores the adaptability of RDRs in accommodating new medical knowledge and improving diagnostic processes (Hyeon et al., 2016).

In another study, Clinical Dc and Laukkanen (2017)) utilised RDRs to create a DSS for managing chronic kidney disease. The system integrated patient data, clinical guidelines, and expert knowledge to offer personalised treatment recommendations. The researchers found that the RDR-based system significantly improved patient management by providing timely and accurate advice, thus reducing the likelihood of complications and hospitalizations (Kunutsor and Laukkanen, 2017). As discussed above, these DSSs depend on a maintained knowledge base; without one, their recommendations cannot be updated or audited.

Moreover, the use of RDRs in developing systems for managing diabetes has also been explored. A study by Hussain et al. (2021) focused on creating an RDR-based decision support system in assisting evidence-based decision making for complex problems. This system incorporated a comprehensive knowledge base of the various complex problems and patient data. The findings indicated that the system enhanced the decision-making capabilities of healthcare providers by offering evidence-based recommendations tailored to individual patient needs. This approach not only improved clinical outcomes but also streamlined the management of various health conditions (Hussain et al., 2021).

The flexibility and incremental nature of RDRs make them particularly suited for dynamic and evolving fields like healthcare. RDRs allow for continuous updates and refinements to the knowledge base, ensuring that the system remains current with the latest medical research and clinical guidelines. This adaptability is crucial in healthcare, where new treatments and diagnostic techniques are constantly being developed (Compton et al., 2018).

In addition, the integration of RDRs with other advanced technologies, such as machine learning and natural language processing, has further enhanced their capabilities. A study by Lee et al. (2019) explored the combination of RDRs and machine learning algorithms to develop a hybrid DSS for cardiovascular disease management. The system leveraged the strengths of both methodologies, resulting in improved diagnostic accuracy and treatment planning. This hybrid approach exemplifies the potential of RDRs to be integrated with cutting-edge technologies and provide robust support for complex medical decisions (Lee et al., 2019).

Furthermore, RDRs have been employed in the development of telemedicine applications, which have gained prominence during and since the COVID-19 pandemic. However, even prior to COVID-19, a study by Han et al. (2015) focused on using RDRs to create a telehealth platform for remote patient monitoring and management. The system allowed healthcare providers to continuously monitor patients' health status and offer timely interventions based on real-time data. The use of RDRs ensured that the knowledge base could be easily updated with new information, enhancing the system's responsiveness and effectiveness (Han et al., 2015). In the past decade, rule-based and hybrid CDSSs have supported triage, guideline adherence, and telemedicine workflows in routine care (Greenes et al., 2017; Sim et al., 2017; Hamedan et al., 2020).

Overall, the continued application and development of RDRs in healthcare highlight their significant role in enhancing DSSs. By providing a flexible and adaptable framework for knowledge acquisition and management, RDRs enable healthcare providers to make more informed and accurate decisions, ultimately improving patient care and outcomes.

2.5 Summary of the Literature Review

This literature review examined factors influencing type 2 diabetes, emphasizing social determinants, economic impacts, and the role of decision support systems (DSS) and Ripple

Down Rules (RDR). The chapter highlights the necessity of a robust knowledge-based system (KBS) to manage these complexities effectively.

This thesis has explored diabetes's economic and social impacts, relevant health policies, and the implementation of a KBS using Ripple Down Rules (RDR). The reviewed literature shows these elements have been well-studied individually but require integrated research to address Type 2 diabetes holistically.

This is evident in the selected articles listed in the synthesis matrix in Table 2.1. Only five articles relating to previous research (columns) have been selected to illustrate the point. However, this is not to say that this table could not be expanded to include more previous research work. Indeed, more aspects (rows) could also be considered in future research.

	(Diabetes in Australia 2015)	(Schwerdtle, 2016)	(Hyeon et al., 2016)	(Sim et al., 2017)	(Ramaprasad et al., 2016)
Cost & Impact	√				
Social determinants		√			
RDRs			√		
Policy		√			√
KBs & DSSs				√	

Table 2. 1 . Relevance of selected references within their own domain

The purpose of this current research is to combine all the above aspects (rows) in the development of a KBS using RDRs to assist in policy making aimed at reducing the cost and social aspect of type 2 diabetes. This would involve using all the concepts and techniques discussed in this research thesis and more in the development of a user-friendly KBS that assists policy makers in allocating the appropriate resources more efficiently. The KBS would be developed using RDRs as its algorithm. The RDRs themselves would be based on existing data gathered from experts in the field.

In so doing, this research would be continuing from where some of the research left off, hence essentially filling a gap in the existing research.

Recent studies highlight the escalating global burden of type 2 diabetes, which continues to rise in prevalence and economic impact. The World Health Organization (WHO) emphasises that addressing the social determinants of health, such as income inequality, education, and access to healthcare, is crucial in combating this disease (Bull et al., 2020). In Australia, the financial burden of diabetes has been increasing, with recent estimates suggesting a cost of approximately $15.3 billion in 2018 (Lee et al., 2018). Similar trends are observed in other regions, such as the United States, Europe, Asia, and Africa, where the economic impact of diabetes is substantial (Association, A.D, 2018; Williams et al., 2020; Xu et al., 2018; Mutyambizi et al., 2018).

The need for effective management strategies is evident, and DSSs have been shown to play a pivotal role in healthcare by optimising resource use and enhancing decision-making processes. However, the effectiveness of DSSs is intrinsically linked to the robustness of the underlying KBS. Without a well-structured KBS, DSSs lack the foundational knowledge required to provide accurate and reliable recommendations (Hurtubise et al. 2016; Greenes et al. 2018; Khamisi et al. 2019)

RDRs offer a practical solution to the challenges associated with developing and maintaining a KBS. They enable incremental knowledge acquisition and continuous refinement, ensuring that the system remains current with the latest medical research and clinical guidelines. This adaptability is particularly beneficial in healthcare, where new treatments and diagnostic techniques are constantly emerging (Compton et al. 2014; An et al. 2023). Recent applications of RDRs in healthcare demonstrate their effectiveness in managing complex medical data and improving patient outcomes. Studies have shown that RDR-based systems significantly enhance the decision-making capabilities of healthcare providers, offering personalised treatment recommendations and reducing the likelihood of complications (Kunutsor & Laukkanen 2017; Hussain et al. 2021).

The integration of RDRs with other advanced technologies, such as machine learning and natural language processing, further enhances their capabilities. For instance, hybrid DSSs that combine RDRs and machine learning have been developed for cardiovascular disease management, resulting in improved diagnostic accuracy and treatment planning (Lee et al., 2017). Similarly, the use of RDRs in telemedicine platforms for remote patient monitoring and management has shown great promise, particularly during the COVID-19 pandemic (Han et al., 2015).

This chapter has provided the groundwork, hence this thesis now advances into the methodological realm, where the theories and concepts previously discussed will be applied and tested. The next chapter details the research methodology, a critical component that will outline the systemic approach utilised in developing a user-friendly (DSS) powered by RDRs.

Social determinants like income, education, housing, and nutrition are pivotal in Type 2 diabetes prevalence and management. Addressing these determinants at the policy level, as suggested by Schwerdtle (2016), is essential for developing effective interventions. Research indicates that education plays a crucial role in promoting healthier lifestyles and preventing diabetes (Hill et al. 2013; Sauliune & Kalediene 2015).

The incremental development of a type 2 diabetes KBS using RDRs provides a flexible and adaptive framework for capturing and integrating new knowledge about social determinants and their impact on diabetes. This approach ensures that the KBS evolves continuously, incorporating the latest research findings and clinical guidelines. By leveraging RDRs, the proposed KBS can offer personalised recommendations and support the proactive management of diabetes, ultimately reducing the economic and social burden of the disease.

This methodological exploration will traverse the structured steps of identifying pertinent data, formulating rules based on expert insights, and constructing a KBS aimed at efficient resource allocation and policymaking to address the economic and societal burdens of type 2 diabetes. The ensuing chapter shifts from theoretical synthesis to a pragmatic construction, setting the stage for the innovative application of RDRs in synthesising a knowledge base that not only aligns with but also amplifies the findings from the literature. By intertwining the established research with new methodologies, the thesis seeks to bridge existing gaps and pave the way for future explorations in the field of diabetes management and public health policy.

In recent years, there has been a renewed interest in hybrid KBS integrating symbolic reasoning with data-driven techniques to address interpretability and adaptability challenges (Compton & Kang, 2021; Alrige et al., 2023). Advances in explainable AI (XAI) have also influenced the design of modern KBS, aiming to provide transparent decision support in healthcare and other critical domains (Holzinger et al., 2019).

In conclusion, the literature review has reaffirmed the need for a comprehensive KBS for managing type 2 diabetes, highlighting the benefits of using RDRs for its development. The incremental and adaptive nature of RDRs makes them particularly suited for the dynamic and complex field of healthcare. As new social determinants and medical insights emerge, the KBS

can be continuously updated to provide accurate and evidence-based recommendations. This approach not only enhances the decision-making capabilities of healthcare providers but also contributes to better patient outcomes and more efficient use of healthcare resources.

Accordingly, a clear gap remains: an interpretable, incrementally extensible KBS that operationalises socio-demographic determinants for population-health decision support in data-sparse settings; the following chapters address this gap. While RDR is established in clinical domains, its extension to population-health with socio-demographic inputs and explicit governance remains underexplored; the next chapter presents that implementation and evaluation.

Chapter 3: Methodology

This chapter is divided into seven sections and delineates the structured methodology underpinning the research, guided by the objective to elucidate the relationship between social determinants and the onset of type 2 diabetes, and to explore the feasibility of establishing a comprehensive knowledge-based system (KBS) to address this challenge. Section 3.1 introduces the research questions, laying the groundwork for the study by probing the link between social determinants and type 2 diabetes and seeking to identify specific influential factors. Section 3.2 delves into problem identification, examining existing literature to pinpoint gaps in the current understanding of these social determinants, thereby setting the stage for the KBS's design and incremental development. The structural blueprint of the proposed KBS, which aims to incorporate and adapt to new determinants over time, leveraging Ripple Down Rules for its development, is outlined in Section 3.3. Section 3.4 discusses the design science research methodology (DSRM), chosen for its suitability in crafting artifacts that address identified problems, specifically in the context of developing a type 2 diabetes KBS. The critical phases of evaluation and validation, where the KBS's efficacy, reliability, and performance are assessed using real-life data and expert insights, are detailed in Section 3.5. This phase is pivotal in validating the system's capability to employ social determinants effectively in predicting type 2 diabetes. Section 3.6 emphasises the importance of meticulous documentation and transparent reporting of the research process, findings, and the KBS development and evaluation outcomes. Finally, Section 3.7 summarises the methodology chapter, encapsulating the research's structured approach from the initial problem identification to the evaluation of the developed system, reaffirming the study's commitment to enhancing type 2 diabetes management through innovative, knowledge-based solutions.

3.1 Research Questions

This research addresses the following three research questions:

1. Is there a relationship between social determinants and the onset of type 2 diabetes?

2. If so, can some of these social determinants be established?

3. Can the aforementioned social determinants be used to incrementally develop a knowledge-based system?

These research questions serve as the foundation for the research methodology and guide the subsequent phases of the study. The problem identification phase (Section 3.2) involves investigating the existing knowledge and identifying the research gap related to social determinants and type 2 diabetes. The design phase (Section 3.3) focuses on the creation of the KBS, while the conceptual framework phase (Section 3.4) lays the groundwork for the incremental development of the KBS.

The evaluation and validation phase (Section 3.5) is a critical component of the research methodology. It involves assessing the performance, effectiveness, and reliability of the developed KBS. The evaluation process includes analysing real-life patient data, cross-matching it with social determinants associated with specific geographic regions and comparing the KBS's outcomes with known diagnoses. The validation process ensures the credibility and generalizability of the KBS, confirming its ability to accurately determine the likelihood of a person being a type 2 diabetic in any given region. The involvement of subject matter experts and healthcare professionals adds further credibility to the evaluation and validation process.

Overall, Chapter 3 presents the research methodology employed in this thesis, which encompasses four key phases: problem identification, design of a knowledge-based system, initiation of the conceptual framework, and evaluation and validation of the concept. By following this methodology, the research aims to address the research questions and contribute to the development of an effective KBS for type 2 diabetes. The evaluation and validation phase play a crucial role in ensuring the accuracy and reliability of the KBS in identifying the social determinants associated with the disease.

3.2 Overview of the Design Science Research Methodology (DSRM)

The DSRM provides a structured framework for constructing and evaluating artefacts that interact with the domain problem to solve specific problems or enhance the context. In this chapter, the five phases of the design cycle in DSR, as described by Wieringa (2014), and the seven guidelines proposed by Hevner et al. (2004) for conducting DSR in information systems as used in this thesis are explained. The application of DSR is particularly suitable for this research as it aligns with the solution-oriented approach required for developing a type 2 diabetes KBS. How the seven guidelines of DSRM are applied and their relevance to this research will be clearly articulated.

The DSRM focuses on the creation of various artefacts, as indicated in Table 3.1. These artefacts include constructs, models, frameworks, architectures, design principles, methods, instantiations, and design theories (Vaishnavi & Kuechler, 2004). These artifacts are designed to address specific problems and improve the practical applications in the given context. The research process involves the construction, modelling, and instantiation of these artifacts to generate new knowledge and solutions.

Artefact	Description
Constructs	The symbols or vocabulary to communicate a research domain
Models	The abstractions that express the relationships between constructs
Methods	Sets of steps (e.g. algorithms, process) that provide a guide to performing the tasks
Frameworks	The conceptual guides which serve as a guidance
Architectures	High level structures that underlying the designed systems
Instantiations	The proof of concepts of an artefact in its environment
Design Principles	The main concepts and principles that guide the design of the artefact
Design Theory	Sets of statements that facilitate understanding of the guidance to achieve a certain objective

Table 3. 1 Artefacts of Design Science Research (Vaishnavi & Kuechler 2004)

DSR is widely used in fields such as engineering and computer science and is also prevalent in other fields such as information systems. The approach is mainly concerned with problem-solving and improving the practical applications of the developed artefacts. In DSR, the developed artefacts are designed to interact with the domain problem to solve the problems or improve the context. The process of design science encompasses the construction of new knowledge through an original design of artefacts, which involves constructing, modelling, and instantiating new concepts or ideas. These artefacts can take different forms, such as constructs, models, frameworks, architectures, design principles, methods, instantiations, and design theories. DSR approach is suitable for the implementation of this research because it is field-problem driven and solution-oriented, making it ideal for practical problem-solving with continuous improvement. Additionally, this approach is well-suited for the development of a type 2 diabetes KBS because it requires solution-oriented knowledge that can be used to solve

complex and relevant problems. It is utilised to develop a new method that enables the rapid evaluation of incrementally developing a type 2 diabetes KBS.

The KBS has a focus on developing a diagnostic tool that can help clinicians with the early detection and diagnosis of type 2 diabetes. The proposed method is designed to improve the efficiency and accuracy of the diagnosis process. To implement the design science approach, the research is divided into five phases: problem investigation, treatment design, treatment validation, treatment implementation, and evaluation. In the problem investigation phase, the research explains what the phenomenon is and why it needs to be investigated and improved. In the treatment design phase, the researchers design one or several artefacts to solve the problem. In the treatment validation phase, the researchers explain if a design can solve a problem. In the treatment implementation phase, the researchers treat the problem with one of the design artefacts. Finally, in the evaluation phase, the researchers evaluate if the treatment can solve the problem, what needs to be improved, and whether the cycle should start from the beginning.

Moreover, this research also follows the seven guidelines proposed by Hevner et al. (2004) to conduct DSR in the information systems field. These guidelines are design as an artefact, problem relevance, design evaluation, research contributions, research rigor, design as a search process, and communication of research (see Table 3.2). The proposed diagnostic KBS will be developed using these guidelines, and each guideline's application in this research is described in detail.

No	Guideline	Description	Application in this research
1)	Design as an Artefact	DSR needs to produce an applicable artefact	The development of artefacts that includes constructing, model, method and architecture

2)	Problem Relevance	Artefact in DSR is developed to provide important solutions that are relevant to human problems	An incremental knowledge acquisition method will be adopted to construct a diagnostic KBS
3)	Design Evaluation	The artefact needs to be demonstrated through rigorous evaluation procedures in terms of quality and efficacy	The design artefact will be evaluated against actual primary data
4)	Research Contributions	DSR must provide a clear contribution in terms of the design or methodologies.	The design artefact will contribute a new understanding of evaluating a developed KBS.
5)	Research Rigor	The design artefact must be rigorously constructed, implemented, and evaluated	The research rigour is achieved by testing and evaluating the design artefact to determine the best solution with the reasons to improve the design. It will be developed by a subject matter expert.
6)	Design as a Search Process	The expected artefact is achieved from an effective searching process while satisfying the problem	The design artefact has its own methodology that will be applied to improve the knowledge of the KBS.
7)	Communication of Research	The output of DSR should be effectively presented to technology-oriented and management-oriented audiences	The communication of this research was presented through conference and journal publications (Omar et al., 2019) and (Omar et al., 2022).

Table 3.2 DSR Guidelines (Hevner et al. 2004, p. 83)

The implementation of DSR in this research is organised in four phases.

These phases were created in accordance with the DSR guidelines, as shown in Table 3.2. As illustrated in Figure 3.1, the first phase focuses on identifying the knowledge gap from relevant literature. This is the implementation of design as an artefact and problem relevance. The second phase is the realisation of design evaluation, designing the generic knowledge base for the type 2 diabetes KBS. In Phase 3, the design and development process that addresses the research question is examined. This phase is the realisation of design evaluation and research contribution. Finally, Phase 4 demonstrates and evaluates the proposed prototype.

This is the implementation of research rigor and design as a search process. According to the principles of DSR, following evaluations, prior phases are revisited to refine the model and the knowledge base development process.

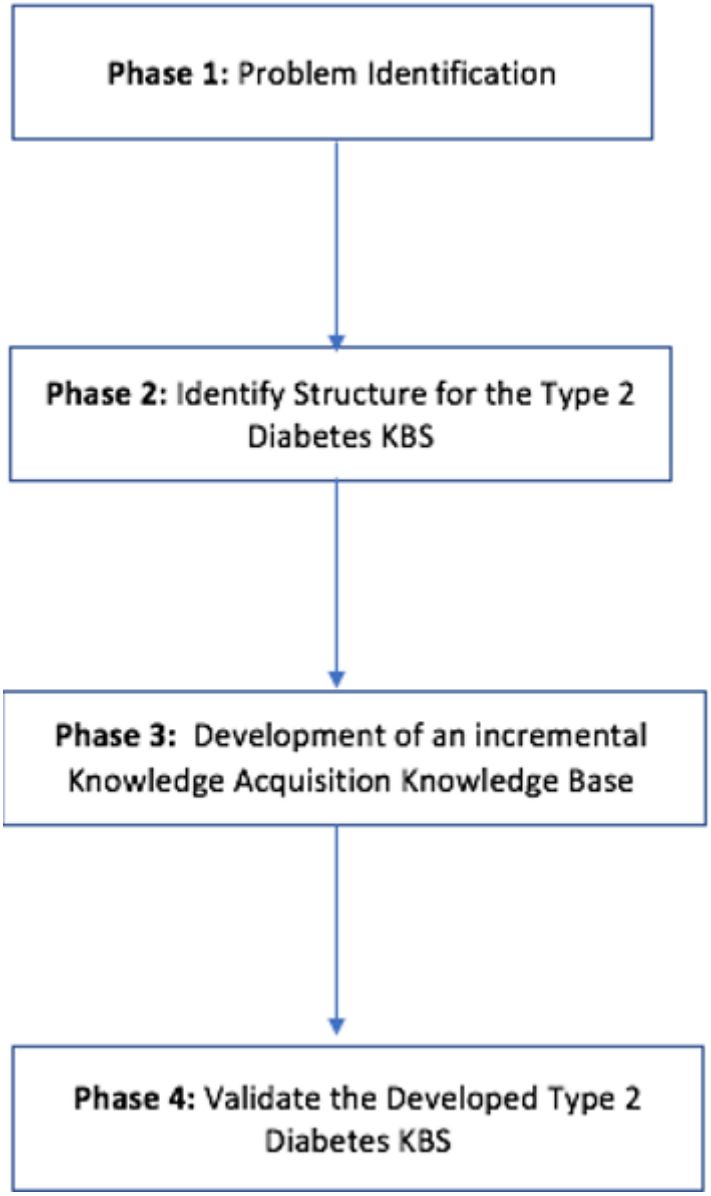

Figure 3. 1 Research Phases

The DSRM is selected as the research approach in this study to develop a new method for the rapid evaluation of developing KBS. The proposed method is designed to improve the efficiency and accuracy of the diagnosis process for type 2 diabetes. The research follows the four phases of design cycle, and seven guidelines proposed by Hevner et al. (2004) to conduct DSR in the information systems field.

3.3 Phase 1: Problem Identification

This section describes the initial phase of the DSRM, which focusses on problem identification. The purpose of this stage is to investigate existing knowledge to raise the main problem and explore potential solutions.

As discussed in section 1.1, this research answers the following research questions:

1. Is there a relationship between social determinants and the onset of diabetes type 2?
2. If so, can some of these social determinants be established?
3. Can the aforementioned social determinants be used to incrementally develop a knowledge-based system?

This research aims to answer three research questions related to the relationship between social determinants and the onset of type 2 diabetes, establishing some of these social determinants, and using them to develop a KBS incrementally. In DSR, this phase is also referred to as problem investigation, where the purpose is to understand the underpinning problem and explore a better designed solution. According to Wieringa (2014), there are four categories to investigate the problem: problem-driven, goal-driven, solution-driven, and impact-driven investigation. These processes of problem identification are similar to those suggested by Peffers et al. (2007). As shown in Figure 3.2, there are four possible entry points in DSRM: problem-centred initiation, objective-centred solution, design and development-centred initiation, and client/context initiation. The problem identification process may start at any step and move outward.

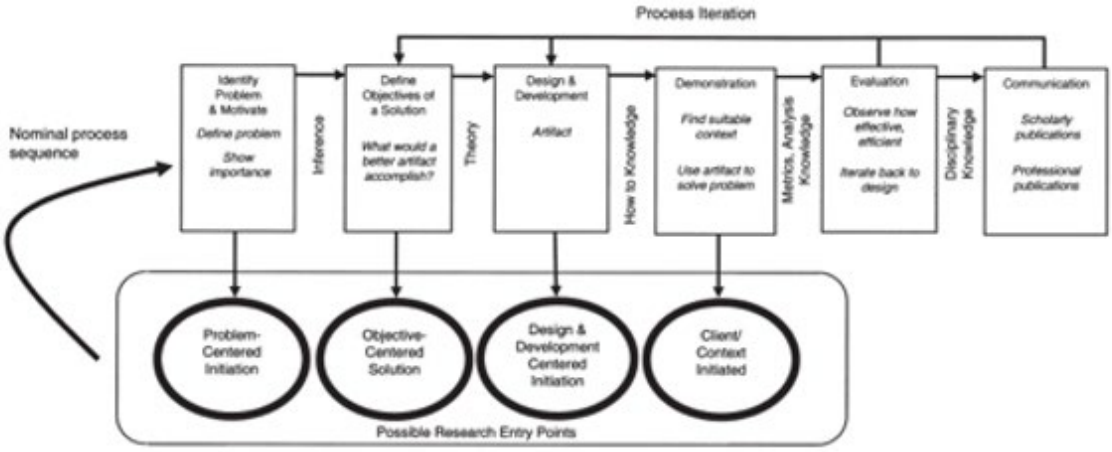

Figure 3. 2 DSRM Process Model (Peffers et al. 2007)

The problem is framed within the Design Science Research (DSR) method, entering at the design-and-development stage. This phase involves reviewing previously published studies related to knowledge acquisition and KBSs, specifically related to evaluation, validation, verification, issues, and opportunities. The artefact is an incrementally developed rule-based KBS for Type 2 diabetes. After developing a proof-of-concept-level prototype, the proposed KBS was implemented using actual primary data accumulated from various medical centres around the Albury-Wodonga regions on the NSW-VIC border. The dataset used to develop the type 2 diabetes KBS is broken down into three datasets: the training dataset, the KBS development dataset, and the production dataset. Finally, it was highlighted that the three datasets were formed by randomly selecting cases to place into each of the datasets, removing any bias of any types of cases and contributing to the validation of the system.

The dataset used to develop the type 2 diabetes KBS was initially structured into three subsets: the training, development, and production datasets, as discussed in the previous paragraph. However, the practical implementation and evaluation primarily utilised two datasets—the training dataset for incremental rule development and the production dataset to rigorously validate the KBS on unseen cases

Consistent with Figure 3.2, this study adopts the design-and-development–centred entry point of the DSRM (Peffers et al., 2007). The research artefact is an incrementally developed rule-based KBS built via RDR knowledge acquisition, selected for its flexibility and interpretability.

To determine the conditions to terminate the KBS development process, the evaluation field was surveyed, and we focused on dynamic evaluation by monitoring the knowledge acquisition process (Beydoun & Hoffmann 2013). Despite this promising philosophy, more investigation needs to be undertaken as the approach has never been applied in practice. The prototype's capability of the proposed KBS was confirmed after developing a proof-of-concept-level prototype.

Chapters 1 and 2 of this thesis detail the output of the activity in phase 1. Chapter 1 provided the rationale behind this research, ranging from the description of the problem to the solution of this research and Chapter 2 described the investigation problems that existed in the literature.

3.4 Phase 2: Identifying the Structure of the KBS

The research work aims to develop a type 2 diabetes KBS using the DSRM. The second phase of the DSR involves identifying the structure of the type 2 diabetes KBS. The purpose of the KBS is to represent the domain knowledge of a person with type 2 diabetes in a social environment in a generic way that can be used to determine the likelihood of a person being a type 2 diabetic, regardless of their social background.

To construct the KBS, a comprehensive guideline has been developed, as illustrated in Figure 3.3, which comprises six steps. In Step 1, the researcher collected data from appropriate sources to develop the KBS. In Step 2, the researcher selected relevant attributes from the collected dataset. Step 3 involves the extraction of concepts related to a type 2 diabetic from the identified attributes, and Step 4 involves the selection of general concepts that are used in the models. Step 5 is the classification of selected concepts based on their function in the KBS. Finally, Step 6 is the validation of the initial KBS to measure its generality, expressiveness, and completeness.

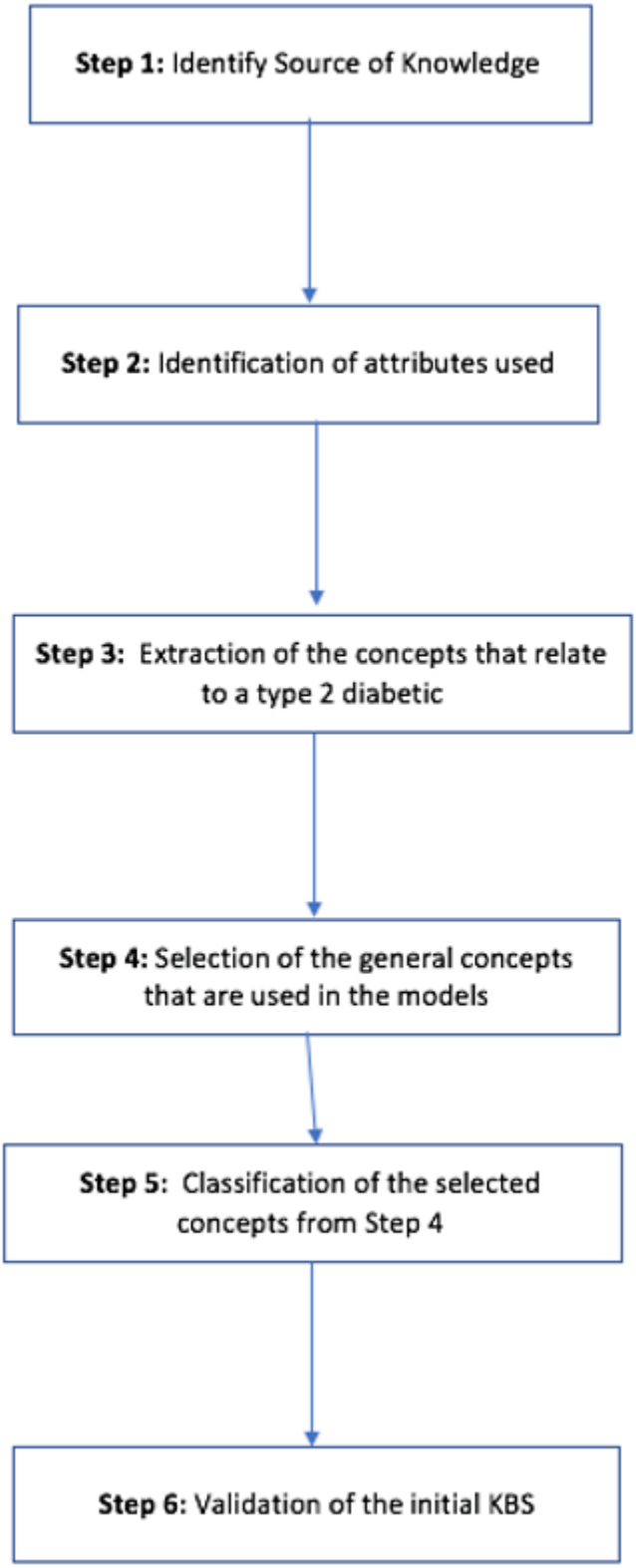

Figure 3. 3 Overall Guidelines of a type 2 diabetes KBS

Figure 3.3 summarises the six-step guideline; the steps are briefly outlined below.

Step 1: Identification of the source of data to be used in the development of the KBS

The main activity in this step is to collect data from the appropriate sources. To obtain the best results, the source of knowledge is selected from several medical centres in the Albury – Wodonga region of the NSW – VIC border, Australia. This region was selected for pragmatic access to de-identified data and because its cross-border, mixed urban–rural profile provides varied case mix; findings are interpreted as a prototype within-region demonstration rather than a claim of national representativeness.

Step 2: Identification of attributes used

Using the dataset obtained in the previous step, the attributes deemed relevant for this research are selected. The attributes selected for this research and the reasons they were selected are discussed further in Chapter 5.

Step 3: Extraction of the concepts that relate to a type 2 diabetic

The extraction process involves the identification of concepts among models that have high feasibility of being included in the initial KBS model. The extraction process is carried out by inspecting all the attributes, then identifying the important concepts to be used as model data. The extracted concepts represent the characteristics of a type 2 diabetic.

Step 4: Selection of the general concepts that are used in the models

In this step, the list of concepts derived from the previous step was analysed and refined. The first step in selecting a patient concept is to filter the number of occurrences and generality of the attribute(s). After a while, the KBS developer starts to see patterns of certain attributes being repeated with type 2 diabetes patients. These attributes are closely examined and considered in future RDRs. This is discussed in more detail in Chapter 5.

Step 5: Classification of the selected concepts from the previous step

In this step, the selected concepts are grouped into categories according to their function in the KBS. As stated in Step 2, the classification is selected from the review of actual clinical data based on patients with type 2 diabetes.

Step 6: Validation of the initial KBS

The last step is validation to measure the quality of the KBS by investigating the generality, expressiveness and completeness of the proposed KBS. For this purpose, a series of experiments was conducted and evaluated. These experiments are discussed in Phase 4 of this chapter and will be further discussed in more detail in Chapter 5.

The output of Phase 2 of the overall development process of the type 2 diabetes KBS is described in detail in Chapter 5 of the thesis. The research work discussed above builds on previous work by Beydoun et al. (2009) in developing a generic type 2 diabetes KBS by eliminating the last step. The methodology proposed in this research will contribute to the development of a generic type 2 diabetes KBS that can be used to identify the likelihood of a person being a type 2 diabetic, regardless of their social background.

The methodology proposed in this research work is important because it addresses the issue of identifying the structure of the KBS, which is a critical component in the development of a type 2 diabetes KBS. The approach is based on the DSR paradigm, which provides a framework for the development of information systems that are relevant to specific domains. The proposed methodology is therefore relevant not only to the development of a type 2 diabetes KBS but also to other domains that require the development of a KBS. The validation of the KBS in Step 6 of the methodology ensures that the KBS is of high quality and is useful in decision-making processes related to type 2 diabetes.

The second phase of the DSR involves identifying the structure of the type 2 diabetes KBS. The proposed methodology is based on the DSR paradigm and comprises six steps that are used to develop a generic type 2 diabetes KBS that can be used to determine the likelihood of a person being a type 2 diabetic, regardless of their social background. The methodology proposed in this research work is relevant to the development of a KBS in other domains and ensures that the KBS is of high quality and useful in decision-making processes.

3.5 Phase 3: Development of an Incremental Acquisition Knowledge Base

Phase 3 of the DSR process involves the design and development of the primary artefact to address the problem identified in Phase 1. For this PhD thesis, the main artefact is an instantiation of a type 2 diabetes KBS. The prototype of the KBS is built from scratch using the RDR development tool based on macros in Excel. The knowledge base of the system is incrementally built from the test cases supplied into the system, and the system refines the test cases and builds a rule tree, based on these cases.

Two requirements need to be met for the development process. First, the system should facilitate knowledge extraction and the structuring of knowledge from the source. The system allows for direct rule creation, but not correction to existing rules, as this could negatively impact the existing knowledge in the knowledge base. Hence, the subject matter expert (SME) plays a crucial role in manually refining rules and/or creating new rules. Second, the system should enable the knowledge acquisition process to be monitored, which is another role for the SME. The artefacts' output, including the model, framework, and architecture, is presented in Chapter 5.

Phase 3 produced the research artefact: a prototype, RDR-based, rule-driven KBS instantiated for Type 2 diabetes. The outputs feed Chapter 5, where we evaluate rule growth, performance and interpretability against the stated research aims.

3.6 Phase 4: Validation of the Incrementally Developed KB

Phase 4 of the DSR methodology involves the validation of the developed KBS. In this phase, the main goal is to ensure that the developed system is fit for its intended purpose and meets the specified requirements. This phase is critical as it ensures that the system developed in the previous phase (Phase 3) meets the expected quality standards and performs as intended.

Phase 4 of the DSR methodology involves the evaluation and validation of the incrementally developed KBS for type 2 diabetes. This phase aims to assess the effectiveness and accuracy of the KBS in identifying the likelihood of a person being a type 2 diabetic, with a particular emphasis on the utilization of real-life patient data and the incorporation of social determinants associated with specific geographic regions.

In addition, the socio-demographic KBS will be introduced in Chapter 5 was also evaluated on the production (unseen) dataset and compared with geographic KBS, as detailed in Chapter 6 (Section 6.2). This assessment specifically measures the system's ability to generalize to new data, highlighting significant improvements in accuracy and transferability over the geographic-based KB.

Furthermore, to comprehensively evaluate the effectiveness of the RDR-based KBS approach, a comparative analysis using a machine learning (ML) algorithm (specifically, the Weka J48 decision tree generator) was conducted. This ML-based approach provided a performance benchmark, allowing the strengths and limitations of the expert driven incremental RDR approach to be clearly demonstrated, as presented in detail in Chapter 6.

3.6.1 Evaluation of the KBS

The evaluation process focuses on measuring the performance of the KBS and its ability to provide reliable results. To evaluate the KBS, a comprehensive analysis of real-life patient data is conducted. The initial dataset is collected from various medical centres in the Albury-Wodonga region of the NSW-VIC border, Australia, as previously mentioned. This dataset serves as a valuable resource for validating the KBS, as it represents the actual patient population from the targeted geographic region.

During the evaluation, the KBS is tested using this real-life patient data, and the results are compared against known diagnoses and outcomes. The accuracy, precision, recall, and other relevant evaluation metrics are calculated to determine the performance of the KBS in correctly identifying type 2 diabetes cases. Additionally, the KBS is assessed in terms of its scalability, efficiency, and usability to ensure its practical viability in a healthcare setting.

3.6.2 Validation of the KBS

The validation process focuses on verifying the credibility and generalizability of the incrementally developed KBS. A key aspect of the validation is the involvement of the SME who has been actively contributing to the development of the KBS. The SME's knowledge, expertise, and insights are crucial in ensuring that the KBS accurately represents the domain and effectively captures the relevant concepts and rules.

Furthermore, the validation process incorporates the integration of social determinants associated with the geographic regions where the patient data was sourced. Initially, the patient

data included postcodes as a proxy for geographic information (ABS 2021 SEIFA Technical Paper). However, to enhance the KBS's global applicability, the postcodes were replaced with social determinants, such as a person ethnicity, highest education level achieved and employment status and type etc. This allows the incrementally developed KBS to accurately determine the type of patient likely to be a type 2 diabetic in any location around the globe, based on the known social determinants associated with that specific region.

In addition to using actual patient data, the developed system was also evaluated by domain experts. The experts evaluated the system's knowledge base in terms of its accuracy, consistency, completeness, and usability. The results of the evaluation indicated that the system's knowledge base was accurate, consistent, and complete, and the system was easy to use.

Figure 3.4 illustrates the guidelines for evaluating the process. Figure 3.4 also shows the SME's input into the incrementally developed KBS.

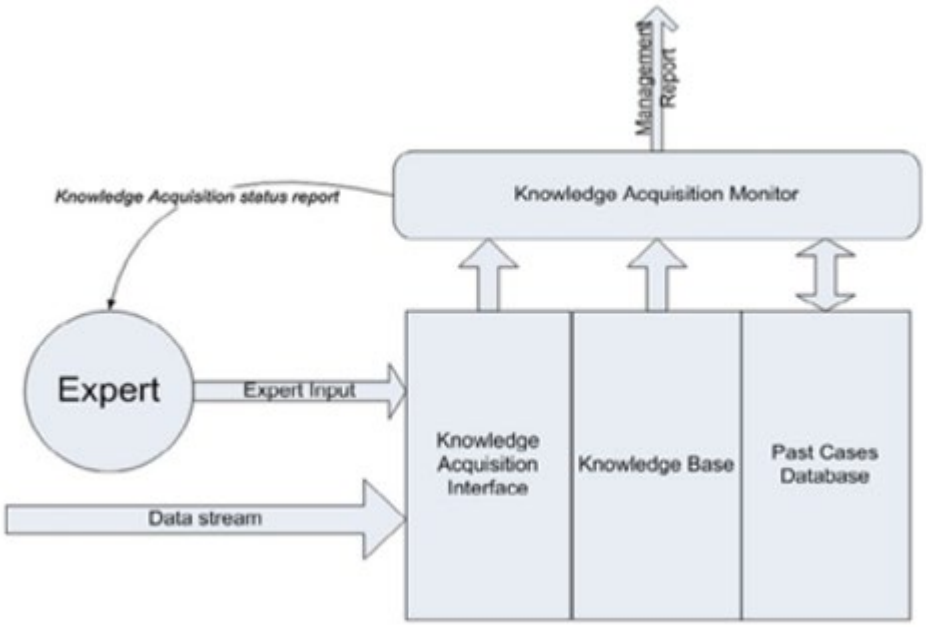

Figure 3. 4 A dynamic montioring process supplements the knowledge acquistion process.

Figure 3.4 illustrates a dynamic monitoring process that supports knowledge acquisition by providing real-time feedback. It incorporates a statistical monitoring component that informs experts about the quality of their most recent interaction and offers management-level insights into the overall progress of the knowledge acquisition project (Beydoun & Hoffman, 2013).

As discussed by Beydoun & Hoffman (2013), the actual knowledge acquisition, and the effectiveness of the newly added rules are evaluated using statistical analysis on the data stream as well as analysing the structure of the KBS. That is, the evaluation is concurrent to the knowledge acquisition cycle (Beydoun and Hoffmann, 2013) as shown in Figure 3.4.

3.6.3 Documentation and Reporting

The findings of the evaluation and validation phase, including the performance metrics, insights from the SME, and feedback from stakeholders, are documented and reported in Chapter 6 of the thesis. The chapter provides a comprehensive analysis of the KBS's performance, its strengths, limitations, and areas for potential improvement. Additionally, the chapter highlights the significance of incorporating real-life patient data and social determinants in the development and validation process, enhancing the credibility and generalizability of the KBS.

Overall, Phase 4 focuses on the evaluation and validation of the incrementally developed KBS for type 2 diabetes. The evaluation phase involves assessing the KBS's performance using real-life patient data, while the validation phase emphasises the involvement of the SME and the integration of social determinants for enhanced credibility and generalizability. The documentation and reporting of the evaluation and validation findings contribute to the overall understanding and advancement of KBS development in healthcare, supporting informed decision-making and improved patient outcomes.

The results of the validation phase are presented in Chapter 6, and the limitations of the study and future research directions are discussed in Chapter 7.

3.7 Chapter Summary

This chapter presents the research methodology used in this thesis, which aims to develop a type 2 diabetes KBS using an incremental knowledge acquisition approach through the DSR framework. This research is structured in four iterative phases, where the problem identification and artefact development phases were discussed in Chapters 1-3.

The problem investigated in this research fits into the category of the design and development-centred approach, where the aim is to enhance and further develop existing type 2 diabetes knowledge. The first phase was problem identification, which included reviewing existing

literature to identify the research gap. Phases 2 and 3 refer to the artefact development based on the problem identified in the first phase. Phase 2 involved the development of a generic diabetes knowledge model, while Phase 3 involved the development of the type 2 diabetes KBS using an incremental knowledge acquisition approach.

Phase 4, which is the focus of this chapter, covers the evaluation of the artefacts developed in Phases 2 and 3. The purpose of the evaluation process is to estimate the effectiveness of the developed knowledge base. The evaluation is an essential phase in the DSR methodology, as it provides feedback to refine the artefacts and ensure they meet the requirements of the stakeholders. The evaluation process is interleaved with the knowledge acquisition and development process to ensure the knowledge base is always up to date.

The evaluation involves two types of validation: face validation and construct validation. Face validation involves evaluating the developed artefact's perceived usefulness, relevance, and ease of use, while construct validation assesses whether the developed artefact measures what it intends to measure. The evaluation process involves domain experts and end-users to ensure the developed artefact meets the requirements and expectations of the stakeholders.

Building on the methodological foundation of the previous chapters, the next chapter presents the data. It specifies data sources, acquisition procedures, quality checks, and preprocessing steps used to convert raw records into the analytic datasets for the type 2 diabetes KBS.

The narrative covers the multifaceted processes undertaken to ensure that the data used not only aligns with the stringent requirements of the DSRM but also stands up to the scrutiny of robustness and reliability. The forthcoming discourse on data explores the hurdles faced in gathering relevant datasets and the meticulous efforts employed to tailor this data, making it fit for purpose. In doing so, it will set the stage for demonstrating the real-world applicability and impact of the KBS, a testament to the incremental and iterative approach that characterises this research.

As we pivot from the theoretical and procedural to the tangible aspects of research, the next chapter provides both an exposé on the practical challenges of data manipulation in health informatics and a validation of the methods that underpin the transformative potential of a type 2 diabetes KBS. This segue serves as a bridge to the empirical endeavours that substantiate the theoretical advancements detailed thus far, leading us closer to realising a comprehensive tool for policymakers and healthcare professionals in the battle against type 2 diabetes.

In conclusion, the development of the type 2 diabetes KBS using an incremental knowledge acquisition approach through the DSR methodology is an innovative approach to enhancing the existing knowledge in the field. The evaluation process is crucial in ensuring the developed artefact meets the requirements and expectations of the stakeholders. The concluding chapter of this thesis will present the conclusion and potential future work.

Chapter 4: Integrating the Geographic Dimension in Knowledge-Based Development

As discussed in previous chapters, integrating social determinants and health indicators can provide a holistic and potentially more accurate assessment of type 2 diabetes risk. The knowledge-based system (KBS) developed in this thesis in essence aims to bridge the gap between clinical data and real-world socio-economic factors towards the risk assessment of type 2 diabetes. The focus of this chapter is on integrating social determinants to enhance the KBS's applicability across diverse contexts, bridging the gap between clinical data and real-world socio-economic factors. Notably, these socio-economic factors are often closely correlated with geographic location, as the environment in which individuals live can significantly influence their access to resources, healthcare, education, and employment opportunities. The location can also reflect the income bracket and overall can be a proxy wealth indicator (Reid et al. 2024). This interplay underscores the importance of considering geographic context when examining the social determinants of health. Beginning with geographic attributes, this chapter presents the interleaving process of the social determinants and health indicators. This will become the knowledge elicitation environment that drives the incremental construction and evaluation of the knowledge base with a human expert engaged in the process. The chapter also delineates the systematic processes involved in the acquisition, preparation, and manipulation of datasets, crucial for constructing a KBS that uses socio-economic attributes to assess type 2 diabetes risk. It is divided into seven sections. Section 4.1 addresses the challenges of sourcing suitable data that straddles socio-technical and medical features and how these challenges are tackled. Section 4.2 focuses on the techniques used to tailor the dataset for the research to ensure that both realms of knowledge are connected. This section, for instance, elaborates on the transition from geographically specific data, which reflects the environmental and locational context influencing health outcomes, to a more generalised dataset incorporating social determinants. Section 4.3 details the critical phase of data preparation and the conceptual framework of the KBS for type 2 diabetes, aligning with the second phase of the design science research methodology (DSRM), as discussed in Chapter 3, that is, defining the objectives of the developed KBS. Section 4.4 provides the detail and statistics of the first attempt to develop a KBS using raw geographic attributes at 90.2% accuracy rate for the training dataset. This section also illustrates the first artefact of this research to be validated and refined later in this thesis. Section 4.5 improves the accuracy rate of the KBS described in section 4.4 by increasing accuracy rate of the training dataset to 100%. Section 4.6 discusses the production dataset accuracy when rated against the

training dataset at the point the training set reached 100% accuracy during rule induction. Finally, Section 4.7 summarises the findings of this chapter, setting the stage for Chapter 5.

Attribute	Description
Case Id	Patient case number. Integer number used to identify the case.
Patient ID	Code to protect patient confidentiality
Analysis_Id	Code to record a particular consultation
Date Attended	Date the patient attend a GP's office or a medical centre.
Times Attended	Times a patient attend GP's office
Withdrawn Screening	Patient deciding to withdraw medical screen. (Yes/No)
Patient Age	Patient's age.
Patient Street	Patient street name. (4-digit code for patient confidentiality)
Patient Town	Town/Suburb of patient's residence
Patient Postcode	Postcode of patient's residence
GP Street	GP's office street name.
GP Town	Town/Suburb of GP.
GP Postcode	Postcode of GP.
Diagnostic DM (years)	Years diagnosed with type 2 diabetes. (Years)
CVD Status	Cardiovascular Disease status (Yes/No)
Diagnostic CVD (years)	How long a patient has been diagnosed with cardiovascular disease. (Years)
HT Status	Hypertension status of a patient. (Yes / No)
Diagnostic HT (years)	Years diagnosed with hypertension (Years)
Alcohol	If a patient consumes alcohol. (Yes/No)
Family History DM	If a patient has a patient has a family history of diabetes. (Yes / No)
Family History CVD	If a patient has a family history of cardiovascular disease (Yes/No)
PHQ9	Measure of depression. (range between 5 to 20)
Diet	If a patient adheres to a particular diet to manage type 2 diabetes. (Yes/No)

Attribute	Description
Case Id	Patient case number. Integer number used to identify the case.
Exercise Duration (hrs/week)	Exercise duration per week. (Hours).
Exercise Intensity	How intense a patient exercises. (Range between None, Low to High)
Last visit to GP	Last time a patient visited a GP. (Months)
Last visit to diabetes educator	Last visit to a diabetes educator. (Months)
Frequency yr GP	How often a patient visits a GP per year.
Frequency (yr) diabetes educator	How often a patient visits a diabetes educator. (per year)
Waist Circumference	Waist circumference (Cm)
Height	Patient's height. (Cm)
Weight	Patient's weight. (Kg)
BMI	Patient's BMI value.
Target	Expected DM result as per medical records. That is, diabetic or not. This was the expected target for each ripple down rule written. (Yes / No).
Conclusion	The KBS conclusion based on the rule

Table 4. 1 Attributes used in the development of the geographic KBS.

4.1. Identifying Medical and Social Data for Interleaving

The central hypothesis of this thesis is that socio-technical attributes, encompassing both social and technical factors, are of immense utility in predicting type 2 diabetes at both the population and individual levels. By incorporating these attributes into the synthesis of the KBS, the resultant system can leverage a comprehensive set of variables that influence diabetes risk towards improved risk assessment. These attributes include, but are not limited to, socio-economic status, education level, housing conditions, and access to healthcare. Many of these socio-economic factors are deeply influenced by geographic location. Environmental and regional disparities often shape access to resources and eventual health outcomes. The integration of such diverse

data points, grounded in geographic and socio-economic contexts, allows for a more nuanced and accurate identification of risk factors.

Multiple sources from the Australian Bureau of Statistics, Diabetes Australia, Diabetes NSW, and NSW Health were examined to obtain relevant data for the construction of a KBS that integrates both health indicators and geographically linked socio-economic factors (Sun et al., 2022; Diabetes around the world in 2021 2024). Notwithstanding the breadth of available demographic data, senior stakeholders confirmed that integrated medical–demographic datasets were not available within NSW Health at that time (Taylor 2019, interview). This constraint was independently corroborated by Professor Vincent Wong, Head of Diabetes at Liverpool Hospital (Wong 2016, interview), who emphasised practical difficulties in assembling comprehensive SDoH-linked clinical datasets. Despite the abundance of demographic and health data, a significant challenge was the lack of integrated datasets combining these two dimensions. It was evident that such datasets were not routinely managed or collected within the purview of NSW Health, posing an obstacle to a holistic understanding of type 2 diabetes risk factors. Alas, despite an abundance of demographic data, the integration with medical data was missing. It was evident that integrated data combining geographically related socio-demographic and clinical health indicators was not routinely managed or collected within the purview of NSW Health. This gap, specifically, the absence of integrated medical/demographic datasets, highlighted the challenges of developing a comprehensive KBS for type 2 diabetes. Incorporating demographic inquiries into Australia's National Health Survey in the future is of course suggested.

Whilst the data combining socioeconomic and health outcomes was not available in NSW, insights into the role of social determinants in type 2 diabetes were gathered within the context of public health data practices at NSW Health. It was observed that socio-demographic data collection was not routinely captured in patient records. Nonetheless, the potential value of incorporating such data was recognised, particularly for designing targeted educational initiatives aimed at mitigating the social and financial impacts of type 2 diabetes. At the time of this study, capture of socio-demographic variables in NSW administrative datasets was limited; while some collections are expanding coverage, it remains incomplete. Additionally, HealthStats NSW offers statistical insights into health inequalities and determinants of health using socio-demographic data (HealthStats NSW, 2024). This shift reflects a broader recognition of the critical role socio-demographic factors play in public health and provides a valuable context for future research, even though it falls outside the immediate scope of this chapter. Chapter 5 will explore the use of socio-demographic data to incrementally develop a type 2 diabetes KBS.

The current scope of socio-demographic data collection by NSW Health includes variables such as age, sex, country of birth, indigenous status, and geographic location (NSW Health APDC Data Dictionary, 2024). These elements are documented in datasets such as the NSW Admitted Patient Data Collection (APDC), which aims to capture essential demographic attributes for patients admitted to hospitals across the state (NSW Health APDC Data Dictionary, 2024).

For clarity, the KBS in this study uses de-identified records and does not require disclosure of individual income or employment status at point of care; socio-demographic context is represented via area-level measures or approved administrative linkages (see section 4.2; HREC H2006-42). Any future operational deployment would follow data-minimisation and consent requirements under ethics and privacy governance. Additionally, HealthStats NSW (2024) provides an extensive overview of health indicators, utilising socio-demographic data to inform public health policies and monitor health inequalities across different population groups (HealthStats NSW, 2024). Despite this progress, there are still notable limitations in the breadth of socio-demographic data collected. While basic demographic attributes are included, more detailed socio-economic indicators such as household income, employment status, educational attainment, and housing conditions are not routinely integrated with the health outcomes data. This gap restricts the ability of NSW Health to fully assess the impact of social determinants on health outcomes, particularly in chronic diseases like type 2 diabetes. Comprehensive socio-demographic data collection is crucial for identifying at-risk populations and tailoring public health interventions effectively.

The separation of socio-demographic data from clinical health records has remained a limitation in the development of integrated health models. Without a unified dataset, critical interactions between social determinants and clinical indicators may be overlooked, limiting the predictive power of a KBS for type 2 diabetes. Future efforts by NSW Health to bridge this gap, such as expanding data collection practices and incorporating socio-demographic variables into clinical databases, could significantly enhance public health strategies and contribute to reducing the burden of chronic diseases. After an exhaustive two-year search, a breakthrough was achieved with the discovery of a dataset from Albury-Wodonga, a region spanning the New South Wales and Victorian border in Australia. This dataset, compiled as part of a diabetes complications research initiative at Charles Sturt University, provides a unique combination of medical statistics and demographic information, including patient residence, age, and BMI. Unlike other datasets, which are largely fragmented or lack integration, this dataset offers a more comprehensive foundation for constructing the prototype diabetes DKBS and evaluating the feasibility of

incremental knowledge base development. Preprocessing was undertaken to tailor the dataset to the objectives of this thesis, as described in the next section.

4.2. Pre-processing of Diabetic Health Data

Use of the de-identified Albury–Wodonga dataset was approved by Charles Sturt University Human Research Ethics Committee (HREC reference H2006-42). The committee granted a waiver of individual patient consent under National Statement section 2.3.10 because no identifiable information was provided to the researchers.

For orientation, the dataset and variables are summarised in section 4.1 and Table 4.1; this section details the preprocessing undertaken prior to RDR rule induction. The original dataset consisted of 2809 records and included individuals diagnosed with both type 1 and type 2 diabetes. To align the dataset with this research's focus on type 2 diabetes, all type 1 diabetes data was systematically removed. The dataset contained suburb and street names, which were anonymised by replacing street names with unique numbering codes known only to the researcher. This process protected patient privacy while allowing geographic analysis of type 2 diabetes clusters.

The dataset does not reflect the normal distribution of type 2 diabetes broadly. It is biased towards a positive datapoint i.e. type 2 diabetic individuals. This inherent imbalance necessitates robust statistical techniques and analytical approaches to derive meaningful insights into the socio-economic unique factors contributing to the onset of type 2 diabetes. Accordingly, results are not interpreted as population prevalence; they are used to guide incremental rule formation and are reported with appropriate caution given the class imbalance. As discussed previously, these socio-economic factors are often deeply intertwined with geographic location, as the environment in which individuals reside can shape their access to resources, healthcare services, and opportunities, thereby influencing health outcomes. It is important to examine the dynamics of type 2 diabetes onset within the context of social determinants and at the same time, recognise the dataset's intrinsic distribution and how it compares to the real-world prevalence of diabetes within the population. After pre-processing, the dataset displays a distribution of approximately 70% non-diabetic and 30% diabetic cases. There is a bias in the data towards positive cases as the real-world type 2 diabetes prevalence of 10% of the world's population. That is, one in ten people around the world have diabetes (IDF 2024). This clearly underscores the urgency of

understanding the social determinants intertwined with geographical location that contribute to diabetes, for the purpose of this research.

As the focus of this research is type 2 diabetes, not type 1 diabetes, all type 1 cases were removed from the dataset. This eliminated 71 cases from the dataset, leaving a total of 2738 cases. This dataset forms the cornerstone of the subsequent analyses and findings presented in this thesis. The number of attributes also reduced from 129 to 135, focusing on relevant socio-demographic factors influencing type 2 diabetes (rather than clinical/pathological). Attributes with more than 90% missing values were also eliminated, and those unrelated to socio-demographic factors were also excluded. This is a screening heuristic to remove near-empty fields; sensitivity checks at 85% and 95% yielded the same retained set. This eliminated 94 attributes, leaving 35 attributes. In fact, some attributes had no values in them at all, hence it made no sense to use these attributes as there were insufficient values from which to draw any valid conclusion(s). These values were missing from the data when the data was first collected. For instance, the *smoking* attribute is rendered redundant as it contained no data and consequently was excised from the attribute set. Similarly, attributes like *bladder problem* are also removed due to very little to no data availability. The attributes *nausea* and *epilepsy*, upon clinical consultation, were also deemed unrelated to social determinants and subsequently excluded. Two additional attributes, *case ID* and *patient ID*, serving as participant case identifiers, were left in the dataset. However, due to their lack of relevance, either clinically or demographically, they were not used in the construction of any RDRs. This left a subset of 35 pertinent attributes available for the construction process of RDRs. Table 4.1 lists show these attributes along with a description of each.

During the data pre-processing phase, attributes with more than 90% missing values were dropped as a data-quality screen; we did not fix this from prior literature robustness checks at 85% and 95% produced the same retained set. This threshold was chosen to ensure the reliability of the constructed KBS and to minimize potential biases arising from excessive data gaps. The same criterion was applied both to the original dataset and to the subsequent, updated version received approximately eighteen months later, which contained a lower proportion of missing values. By consistently applying this rule, only variables with adequate data coverage were included in the KBS development process. Although this approach may limit the breadth of included attributes, it was necessary to maintain analytical robustness and avoid unreliable inferences based on highly incomplete variables.

The final step in preprocessing the geographic dataset involves dividing it into three subsets: training, production, and test datasets. These subsets are randomly partitioned to facilitate the incremental development and validation of the geographic KBS. This division ensures that the system is rigorously tested for reliability and generalizability within the scope of geographic attributes. The integration of socio-demographic datasets and their role in KBS development is discussed in Chapter 5. The training dataset, comprising 500 records, is used to develop the KBS, specifically for constructing the first prototype of the system. The production dataset, comprising 1200 records, is used to test the performance of the KB created by the training dataset. A test dataset, containing the remaining 1038 records, is reserved as a backup in case further refinements are needed.

To summarise, several data pre-processing steps were conducted to align it with the research objectives. First, the pre-processing step ensured the data is focussed solely on type 2 diabetes cases. The second step involved rationalising clinical attributes. Given the focus on socio-demographic factors, these clinical attributes were mostly removed. However, certain clinical attributes with socio-demographic significance were retained for KBS development. Despite their clinical nature, these attributes were deemed relevant and instrumental, leading to their strategic retention for further analysis and utilization within the KBS development process. The final dataset comprised 2738 records with 35 attributes (Table 4.1). The last pre-processing step was to divide the data into three subsets: 500 cases for the training dataset, 1038 cases for the test dataset, and 1200 cases for the production dataset. The test dataset, initially for refining the KBS, was accurate enough for direct use on the production dataset, suggesting its potential for future research phases due to its independent and randomised composition.

In what follows, the abovementioned data is used to incrementally develop a first knowledge base to identify type 2 diabetes cases. This first KBS uses attributes from Table 4.1 and a modified RDR incremental development process. The first KBS does not focus on socio-economic factors but will later serve as benchmark. Socio-economic factors will be used in a subsequent KBS, to be detailed in Chapter 5. The comparison will specifically answer the research question examining the impact of socio-technical factors on the resultant knowledge base. However, in the next section, the customised knowledge acquisition process is first presented.

4.3 KBS Development Process for Type 2 Diabetes KBS

In this section, we explore the crucial steps involved in constructing the KBS for type 2 diabetes, which focuses on identifying key attributes from Appendix 1 to ensure that the order of presentation of cases yields the best outcome. The process begins with the identification of relevant geographic data sources, transitioning from region-specific datasets to broader datasets enriched with geographic indicators. This approach lays the groundwork for a robust framework that captures the unique environmental influences on type 2 diabetes risk while acknowledging the socio-demographic factors often linked to certain geographic regions.

Step 1: Identification of Relevant Attributes

Selecting relevant attributes is crucial for influencing type 2 diabetes knowledge acquisition. An in-depth analysis of these attributes involves statistical methods such as correlation analysis and factor analysis to understand their relationship with type 2 diabetes. These methods provide insights into how various attributes interact and contribute to diabetes risk, ensuring a comprehensive understanding of the underlying patterns. This is a static step that is performed at the beginning of the knowledge acquisition process.

The selection of socio-demographic attributes for this study was driven primarily by data availability and the inherent limitations of accessible datasets. While an ideal model would incorporate a comprehensive range of variables—including ethnicity, occupation, and other social determinants—the reality of data collection practices in Australia, particularly at the time this research was initiated, meant that only a subset of relevant attributes could be obtained.

Specifically, attributes such as median household income, education level, family size, and basic geographic identifiers were chosen because they were consistently reported and publicly accessible. Other important variables, such as ethnicity, were not included because such data were not collected or made available by NSW Health or other relevant agencies during the data collection period, despite direct inquiries to these bodies. In fact, NSW Health explicitly stated that certain socio-demographic data were not collected at that time, although more comprehensive data collection practices have since emerged. Data collection period: 25 November 2002 to 2 June 2014.

These constraints reflect broader challenges in population health research, where data completeness and granularity are often limited by privacy concerns, policy, or historical data practices. As a result, while the findings of this thesis are relevant to the available data and can

inform targeted public health interventions, there are limitations regarding the generalizability and fairness of the results, particularly across diverse population groups where unmeasured factors such as ethnicity may play a significant role.

Future research should prioritize collaboration with public health authorities to advocate for the collection of additional socio-demographic variables, including ethnicity, to improve the robustness, generalizability, and equity of predictive models. Efforts are ongoing to engage with agencies such as NSW Health and propose the inclusion of relevant questions in national health surveys such as the Australian National Health Survey (NHS).

Step 2: Extraction of Data Patterns Relating to Type 2 Diabetes

Patterns in the data are analysed to identify any clear association with type 2 diabetes outcomes. Data mining techniques, such as clustering and association rule mining, can be used in this step to uncover hidden patterns and relationships within the data. These techniques can reveal complex interactions that might not be apparent, providing deeper insights into the data.

This step overlaps with the incremental knowledge acquisition process and continues during the KBS development as a standing and ongoing activity. For example, suppose certain postcodes and suburbs frequently appear in the rules. In this case, they are flagged to indicate potential clusters of type 2 diabetes cases linked to geographic factors such as limited access to healthcare or environmental conditions. These rules are constructed based on observed geographic trends, with diabetic and non-diabetic outcomes determined by these patterns. Rule refinements are applied to address instances where the initial cornerstone cases did not align, ensuring a comprehensive and accurate representation and possible associations.

Step 3: Identification of Key Attribute Patterns for Rule Development

During the initial development of the KBS, exploratory data analysis (EDA) is conducted to identify geographic attributes that exhibit meaningful patterns in relation to type 2 diabetes classification. This process involves analysing the distribution of suburbs, postcodes, and other location-based factors to assess their predictive value. The insights gained from this analysis informed the selection of key attributes that would maximize case coverage while ensuring classification accuracy, serving as the foundation for rule construction in the KBS.

In the RDR framework, an expert's decision-making is influenced by the context, which refers to the set of cases considered when constructing new rules. The selection of attributes is guided by

the observed frequency of specific attributes within diabetic and non-diabetic cases. By analysing the context in which cases are misclassified, the expert determines which attributes are most informative for rule construction. For example, if multiple cases from a specific postcode consistently exhibited diabetes, the expert might prioritize that attribute in constructing a distinguishing rule. This iterative approach ensures that new rules effectively address misclassifications while maintaining generalizability.

Additionally, data interrogation techniques such as filtering and conditional formatting are applied to further refine attribute selection. Attributes with strong predictive power are identified using frequency analysis and statistical comparisons across case groups. This process streamlines rule construction, allowing for the incremental refinement of the KBS without excessive rule redundancy.

By systematically identifying key attribute distributions before rule addition, the rule development process is more efficient. As the KBS expands, the focus shifts from generalizable geographic trends to more refined exceptions and case-specific conditions, improving the system's adaptability. The expert's ability to interpret case-specific variations is clearly a pivotal role in refining rule construction. By actively engaging with case data, the expert ensures that each new rule addressed meaningful patterns rather than relying on static generalizations. For example, if the KBS identifies a correlation between specific socio-economic attributes (e.g., low-income areas, limited healthcare access) and high diabetes risk, these insights can inform rule development by guiding condition selection. This process ensures that rules prioritize statistically significant relationships, increasing the system's ability to generalize findings effectively. Rather than solely cataloguing patterns, the approach strengthens the interpretability of results, supporting both model refinement and practical intervention strategies. Certain variable outputs were filtered to look for common variables with diabetic or not diabetic outputs to quantify the relationships between the selected attributes and diabetes risk. Attributes included, but were not limited to postcodes, suburb names, patient age, exercise duration and patient BMI etc. These techniques allowed for a detailed understanding of how different factors interact and influence health outcomes.

The refinement process also involves sensitivity analyses to assess the robustness of the KBS. This step ensured that the KBS models are stable and reliable, capable of providing consistent risk assessments across different populations and settings. By prioritising the most relevant and impactful attributes, the KBS is designed to deliver accurate and actionable insights for diabetes prevention and management (Fleiss, 2011).

Step 4: Categorization of Key Attributes for KBS Development

Although not initially formalised as a distinct step, the selection of attribute categories played a significant role in refining the rule construction process for the geographic KBS. As rules were iteratively developed, it became evident that different attributes contributed to case classification in varying ways. Attributes naturally fell into four broad categories: location-based attributes (e.g., suburb, postcode), medical history attributes (e.g., family history of diabetes, hypertension), lifestyle attributes (e.g., physical activity levels, diet), and current health status attributes (e.g., BMI, age, blood pressure). While early rules primarily relied on straightforward geographic indicators, later stages of development required a more deliberate selection process, where the choice of attributes was guided by their potential to increase case coverage and improve classification accuracy. By analysing the existing rule set and considering which category of attributes would best resolve misclassified cases, the expert was able to refine the KBS systematically. This evolving approach, though not explicitly documented as a separate step at the time, proved instrumental in ensuring that the knowledge base adapted effectively to the dataset's complexities.

Furthermore, in the hands of a subject matter expert (SME), this categorization process could be highly beneficial. SMEs, with their deeper understanding and domain expertise, are better equipped to determine which attribute category should be prioritised when constructing new RDRs. This is particularly relevant in cases requiring refinement rules to correct misclassified cases, where expert judgment can guide the selection of the most influential attributes to enhance classification accuracy.

After identifying key attribute patterns, the next step involved categorising these attributes to enhance their integration into the rule-based framework. The selected attributes were grouped based on their predictive significance, enabling a structured approach to rule formulation.

Attributes were broadly classified into geographic factors (e.g., postcodes, suburbs) and socio-demographic factors (e.g., average household size, income levels). This classification allowed for a more systematic approach in rule construction. Geographic factors were often more direct indicators, with some postcodes exhibiting higher diabetes prevalence. Socio-demographic factors, however, introduced an additional layer of complexity, requiring more nuanced rule conditions.

The classification process also facilitated context-aware rule refinement, a critical aspect of the RDR methodology. By categorising attributes, the expert could determine which attribute combinations were most effective at distinguishing between diabetic and non-diabetic cases. This

approach prevented redundancy in rule construction and ensured that each rule contributed meaningfully to the system's overall performance.

Ultimately, the structured classification of key attributes streamlined the rule development process, improving both efficiency and predictive accuracy. By aligning rule construction with observed attribute distributions, the KBS was able to adapt dynamically to patterns in the dataset, strengthening its overall reliability.

The classification process involves organising the selected concepts into coherent categories that reflect their function within the KBS. The aforementioned factors collectively could influence an individual's ability to afford healthy food, access healthcare, and maintain a stable living environment (Stringhini et al., 2017). Geographic attributes, such as patient suburb name and patient postcode, were categorised based on their potential impact on healthcare access, availability of resources, and neighbourhood-specific health trends. Socio-demographic attributes, such as education level, income, and housing stability, were included to account for their influence on health literacy, lifestyle choices, and environmental stressors (Cutler and Lleras-Muney, 2010). Housing conditions, for instance, affect living stability, quality of life, and neighbourhood safety, all of which contribute to stress levels and accessibility to healthy environments (Shaw, 2004). By grouping similar attributes, the classification process ensures that the KBS can systematically evaluate and interpret data, providing comprehensive insights into the multi-dimensional nature of type 2 diabetes risk. This structured approach enhances the KBS's ability to identify and address the various factors contributing to type 2 diabetes risk, supporting targeted interventions and informed decision making.

The classification process also involved the development of a hierarchical framework to represent the different levels of influence of each geographic attribute on its possible association with socio-enhanced attribute(s). For example, a geographic attribute such as postcode could be linked to socio-demographic economic indicators, which could then be subdivided into primary factors (e.g. income level, employment status) and secondary factors (e.g. access to financial resources, economic stability). This hierarchical structure provides a comprehensive understanding of how multiple factors interact and contribute to diabetes risk, enabling more precise risk assessments and targeted interventions. The role of socio-demographic factors in influencing type 2 diabetes will be explored in depth in Chapter *5*, particularly in relation to their integration into the KBS framework.

Additionally, the classification process was guided by established theoretical models and frameworks from public health and social sciences. For instance, the Social Determinants of

Health framework provided a conceptual foundation for understanding how socio-economic factors contribute to health disparities and influence type 2 diabetes risk (Cutler & Muney 2006). In addition, the socio-ecological model structured the classification features across individual, interpersonal, community and policy layers (Hill, Nielsen & Fox 2013). By incorporating these theoretical perspectives, the KBS is grounded not only in empirical data but also in well-established models, enhancing its scientific rigor, interpretability, and practical applicability.

The classification process was further refined to ensure that the selected attributes were both practical and relevant to diverse population needs. By incorporating insights from multiple perspectives, the framework was adjusted to address varying priorities and challenges across different socio-geographic contexts. This iterative refinement process enhanced the applicability and effectiveness of the KBS, enabling it to provide meaningful support for diabetes prevention and management across a wide range of settings.

4.4. Results for Geographically Specific Dataset (training dataset at 90.2%)

For the first geographic knowledge base, the first step involves selecting relevant attributes from Appendix 1 based on their geographic significance in influencing type 2 diabetes. These include attributes such as street, suburb name, postcode, patient age and family history of type 2 diabetes etc (ABS 2021 SEIFA Technical Paper). For instance, geographic factors were broken down into street name, suburb, and postcode. Each component was analysed in a bid to see if there were any clusters of type 2 diabetes areas. This would be indicated by the frequency of suburb, street and/or postcodes appearing the rules used to construct the KBS.

The development process then begins by examining the first case in the geographically specific *training* dataset. The examination of the case involved reviewing the attributes and selecting a particular combination of values that leads to a particular conclusion, matching the actual conclusion for that particular case, that is, either *diabetic* or *not diabetic*. Then a rule is written for this case, and this rule is run across the entire dataset. The accuracy rate was then examined as well as selecting the first case in the dataset where the recently added rule did not hold true. The attributes for that case were then compared to the most recently written rule for the last cornerstone case, looking for variations in attribute values between these two cases. A new rule was then written to correct this variation in the classification output. This new rule then became the latest cornerstone case. This process was repeated until an accuracy rate of 90.2% was obtained. A 90.2% stopping point was selected because accuracy gains plateaued beyond this

level while rule count and overfitting risk increased; subsequent refinements produced marginal improvements better assessed on the production dataset.

As per step 2 described earlier, some socio-demographic elements, such as income levels and educational attainment, were considered indirectly as they often correlate with specific geographic attributes. This subtle inclusion acknowledges the intertwined nature of geographic and socio-demographic factors, without shifting the focus away from the primary geographic analysis. Concepts, such as regional healthcare access, proximity to medical facilities, and environmental factors like urban versus rural living, are integrated into the KBS to ensure it accurately reflects the influence of geographic determinants on health outcomes. Each geographic attribute was analysed to understand its specific impact on type 2 diabetes risk. In addition to geographic influences, some clinical factors like body mass index (BMI), blood pressure, and previous diagnoses of prediabetes were incorporated into the geographic KBS. The analysis also considered how regional disparities in healthcare access might affect the management of these clinical indicators. For instance, areas with limited access to medical facilities showed a higher incidence of poorly managed clinical risk factors, underscoring the need to account for both geographic and socio-demographic disparities in the KBS. This combined analysis ensured a holistic approach, enhancing the system's ability to provide accurate risk assessments. Socio-demographic factors will be discussed in depth in Chapter 5.

The attribute and data interrogation techniques included filtering the raw data, using formulae and conditional colouring etc, to see which attribute(s) match a greater number of case outcomes. This approach was used with a variety of different attribute(s) ranging from one to all thirty-five attributes, considering that the more attributes used, the more case specific the rule becomes. The maximum number of attributes used for a single rule in this experiment was eight. The minimum number of attributes used for a single rule was one. Table 4.2 shows the results of the steps discussed to obtain an accuracy rate of 90.2% using 79 rules.

Case Id	Attribute(s) Used	Rule	Rule Conclusion	Cumulative Accuracy (%)	Order Added	RDR No	If True Go To	If False Go To
4	Pat. P Code	IF PPC = 2640	Not Diabetic	33	1	1	2	30
1151	Pat. St. Code	IF PSTC=3656	Diabetic	69.2	11	2	exit	3
1169	Pat. Street & BMI	IF PS=9603 AND BMI>=30	Diabetic	70.4	16	3	exit	4
1192	Gender & Pat. Street	IF Gender=F AND PS=1453	Diabetic	71.4	19	4	exit	5
1227	Pat. St	IF PS=6236	Diabetic	75.8	28	5	exit	6
1230	Pat. St	IF PS=2100	Diabetic	76	29	6	exit	7
1234	Pat. St	IF PS=3130	Diabetic	76.6	31	7	exit	8
1247	Pat. St & last visit to Dr	IF PS=9940 AND LVTDr>0	Diabetic	77.8	35	8	exit	9
1252	Pat. St & Gender	IF PS=6408 AND Gender=F	Diabetic	78	36	9	10	11
840	Gender, Age & Pat. St	IF PS=6408 AND Gender=F AND Age>66	Not Diabetic	78.2	37	10	exit	exit
1275	Pat. St	IF PS=9104	Diabetic	78.4	38	11	exit	12
1283	Pat. St	IF PS=9104	Diabetic	79	40	12	13	15
675	Pat. St & GP St	IF PS=6440 AND GP St=2644	Not Diabetic	79.2	41	13	14	exit
1923	Age, Pat. St & Pat. P Code	IF Age>70 AND PS=6440 AND PPC=2640	Diabetic	88.2	70	14	exit	exit
1307	Pat P Code, Age & BMI	IF PPC=2640 AND Age>73	Diabetic	80.2	46	15	16	19

		AND BMI>28						
952	Gender, Pat P Code, Age & PHQ9	IF Gender=F AND PPC=2640 AND Age<75 AND PHQ9=0	Not Diabetic	80.6	47	16	exit	17
1633	Pat. St	IF PS=6144	Not Diabetic	85.4	63	17	exit	18
1984	Fam. Hist. DM, Pat. St & Pat. P Code	IF FHDM= No AND PS=2120 AND PPC=2640	Not Diabetic	88.8	72	18	exit	exit
1477	Pat. St	IF PS=8535	Diabetic	81.6	51	19	exit	20
1479	Pat. St & HT Stat.	IF PS=7006 AND HT=Yes	Diabetic	81.8	52	20	exit	21
1554	Pat. St	IF PS=6797	Diabetic	83	55	21	exit	22
1583	Pat. St & GP St	IF PS=8425 AND GPST=1453	Diabetic	85.2	62	22	exit	23
1716	Gender, Pat P St, Age & HT Stat	IF Gender= M AND PST=9004 AND Age=67 AND HTSTAT =Yes	Diabetic	87	65	23	exit	24
1830	Pat. St & Pat. P Code	IF PS=6052 AND PPC=2640	Diabetic	87.4	67	24	exit	25
1868	Pat. St & Pat. P Code	IF PS=2609 AND PPC=2640	Diabetic	87.4	68	25	exit	26

1881	Age, Pat. St & Pat. P Code	IF Age>70 AND PS=9808 AND PPC=2640	Diabetic	87.8	69	26	exit	27
2066	Pat. St & Pat. P Code	IF PS=6627 AND PPC=2640	Diabetic	89.8	77	27	28	29
1963	Pat. St, Pat. P Code & Alcohol	IF PS=6627 AND PPC=2640 AND Alch.=No	Not Diabetic	90	78	28	exit	exit
2068	Gender, Age, Pat P Code, CVD_stat, HT Stat, Alchol, Waist Circum. & BMI	If Gender=F AND Age>=69 AND PPC=2640 AND CVDSTAT=Yes AND HTSTAT=Yes AND Alch=Yes AND WST CIRM>=100 AND BMI>=26	Diabetic	90.2	79	29	exit	exit
6	Pat. P Code	IF PPC = 2641	Not Diabetic	48	2	30	31	46
1168	Pat. Town	IF PTwn=Norris Park, Lavington	Diabetic	70.2	15	31	exit	32
1186	Pat. Town	IF PTwn=Norris Park	Diabetic	71	18	32	exit	33
1193	Withdrn. Scrn & Hist DM	IF PWTHSCN=Yes AND HISTDM=Yes	Diabetic	74.4	22	33	34	35
968	Pat. Str.	IF PS=8201	Not Diabetic	74.6	23	34	exit	exit

1282	Pat. St & Cardio Vasc.Stat	IF PS=6939 AND CVD=Yes	Diabetic	78.8	39	35	exit	36
1286	Pat. St & GP St	IF PS=3219 AND GP St=4342	Diabetic	79.4	42	36	exit	37
1487	Pat. St	IF PS=3241	Diabetic	82.2	53	37	exit	38
1494	Pat. St	IF PS=6797	Diabetic	82.6	54	38	exit	39
1556	Pat. St	IF PS=1154	Diabetic	83.6	56	39	exit	40
1568	Pat. St	IF PS=5169	Diabetic	84	58	40	exit	41
1570	Pat. St	IF PS=5169	Diabetic	84.2	59	41	exit	42
1793	Gender, Age & Pat. Street	IF Gender= M AND PST=2491 AND	Diabetic	87.2	66	42	exit	43
1947	Pat. St & Pat. P Code	IF PS=1813 AND PPC=264 1	Diabetic	88.6	71	43	exit	44
2012	Gender, Age, Pat. St & Pat. P Code	IF Gender= M AND Age=76 PS=6773 AND PPC=264 1	Diabetic	89.2	74	44	exit	45
2017	Pat. St & Pat. P Code	IF PS=6773 AND PPC=264 1	Diabetic	89.4	75	45	exit	exit
16	Age	IF Age< = 55	Not Diabetic	54.2	3	46	47	55
1134	Pat. St. Code	IF PSTC=59 80	Diabetic	68.6	9	47	exit	48
1139	Pat. St. Code	IF PSTC=34 68	Diabetic	69	10	48	exit	49
1155	Pat. St. Code	IF PSTC=76 14	Diabetic	69.4	12	49	exit	50

1157	Pat. Post Code	IF PPC=2644	Diabetic	69.6	13	50	exit	51
1160	Pat. Post Code	IF PPC=3677	Diabetic	70	14	51	exit	52
1172	Pat. Street	IF PS=3960	Diabetic	70.6	17	52	exit	53
1472	Gender & Pat. Street	IF Gender=F AND PST=3867	Diabetic	80.8	48	53	exit	54
1473	Pat. St	IF PS=4792	Diabetic	81	49	54	exit	Exit
390	Pat. P Code	IF PPC = 2642	Not Diabetic	54.8	4	55	56	57
1217	Age & History DM	IF Age>=63 AND HistDM= Yes	Diabetic	75	25	56	exit	Exit
416	Pat. P Code	IF PPC = 3690	Not Diabetic	64	5	57	58	70
1732	PHQ9	IF PHQ9>0	Diabetic	74.2	21	58	exit	59
1216	Gender, Age & Pat. St	IF Gender = F AND Age>=60 AND Pat St= 5755	Diabetic	75.2	26	59	exit	60
1233	Pat. St	IF PS=3509	Diabetic	76.4	30	60	exit	61
1245	Pat. St	IF PS=3130	Diabetic	77.4	34	61	exit	62
1303	Pat. St	IF PS=8014	Diabetic	79.8	44	62	63	64
941	Pat. St & Age	IF PS=8014 AND Age<=75	Not Diabetic	80	45	63	exit	Exit
1476	Pat. St	IF PS=3336	Diabetic	81.2	50	64	exit	65
1562	Pat. St & Pat. P Code	IF PS=2028 AND PPC=3690	Diabetic	83.8	57	65	exit	66
1573	Pat. St	IF PS=7322	Diabetic	84.8	60	66	exit	67
1580	Pat. St	IF PS=7322	Diabetic	85	61	67	exit	68
1994	Pat. St & Pat. P Code	IF PS=2210 AND	Diabetic	89	73	68	exit	69

		PPC=3690						
2078	Withdrawn Screening, Gender Age, & Pat. P Code	IF WthdrnScrn=Unknown AND Gender=F AND Age>=79 AND PPC=2640	Diabetic	89.6	76	69	exit	Exit
420	Pat. P Code	IF PPC = 3688	Not Diabetic	64.4	6	70	exit	71
486	Pat. P Code	IF PPC = 2660	Not Diabetic	64.8	7	71	exit	72
503	HT_Status	IF HT_Status=No	Not Diabetic	68.4	8	72	73	78
1197	GP Street	IF GP St=8872	Diabetic	74.8	24	73	exit	74
1225	Pat P Code	IF PPC= 2611	Diabetic	75.4	27	74	exit	75
1240	Pat P Code	IF PPC= 3700	Diabetic	76.8	32	75	76	77
1042	Gender & Pat. St	IF Gender= M AND PS=3867	Not Diabetic	77	33	76	exit	exit
1289	Pat. St & BMI	IF PS=7745 AND BMI>=24.5	Diabetic	79.6	43	77	exit	exit
1423	PHQ9	IF PHQ9=0	Not Diabetic	74	20	78	79	exit
1662	Pat. P Code	IF PPC=3700	Diabetic	86.8	64	79	exit	exit

Table 4. 2 Rule construction table for the geographic training KBS using the geographically specific dataset (90.2% accuracy).

As discussed earlier, the rule construction process involved multiple iterations, each contributing to the incremental refinement of the KBS. Table 4.2 shows that, as rules were added, accuracy generally increased from 33% after the first rule to 90.2% after the 79th rule, with occasional plateaus or small reversals.

The attributes used in rule formulation ranged from patient postcodes (PPC) and patient street codes (PSTC) to more complex combinations involving socio-demographic factors such as age, gender, family history of diabetes, and other health indicators like body mass index (BMI) and cardiovascular status (CVD) (Reid et al. 2024). This iterative rule refinement process exemplified the system's capacity to incorporate both generalizable patterns and nuanced individual cases.

The results of the training phase underscore the system's potential for real-time KBS, particularly in settings where geographic-demographic variables play a significant role. However, it is equally important to examine the system's performance when exposed to previously unseen data. Table 4.1 summarises the system's performance on both the training and production datasets, highlighting KBS's ability to generalize its predictions. The training dataset is the dataset used to develop the KBS, whilst the production dataset is the unseen dataset used to test the developed KBS.

	Accuracy	Specificity	Sensitivity
Training Dataset	90.2%	96.9%	73.6%
Production Dataset	84.5%	92.3%	46%

Table 4. 3 Results obtained during the experiment for the training and production datasets using the geographically specific dataset. Source: Omar et al. (2022).

The accuracy, specificity, and sensitivity metrics presented in Table 4.1 demonstrate that while the KBS performed exceptionally well on the training dataset, achieving 90.2% accuracy, its performance on the production dataset, which comprised of unseen cases, was lower, with an accuracy rate of 84.5%. This drop in accuracy was accompanied by a more pronounced decline in sensitivity (73.6% on the training dataset compared to 46% on the production dataset). The reduction in sensitivity suggests that the system was less effective in identifying true positive cases (i.e., correctly diagnosing diabetic patients) when applied to the production dataset, likely due to the inherent variability and diversity of new, unseen data. Noteworthy here is to also highlight the fact that social determinants are not the only factors that contribute to type 2 diabetes. There are many other clinical factors that can contribute to the onset of type 2 diabetes that are not considered in this research work. For example, one, such factor is the hereditary nature of type 2 diabetes within families.

Despite this, the specificity of the system remained high in both datasets, with 96.9% on the training dataset and 92.3% on the production dataset. This indicates that the system was

consistently effective at identifying non-diabetic cases, reflecting the robustness of the rules developed for capturing patterns related to the absence of diabetes.

The relationship between each rule added and KBS's accuracy was also represented graphically for the reader to see this relationship immediately. Figure 4.1 shows the relationship between the number of rules against the cumulative accuracy rate of each rule, that is, the increase in accuracy as each rule is added. The graph also shows the number of attributes (Operands) used in the construction of each rule added (Omar et al., 2022).

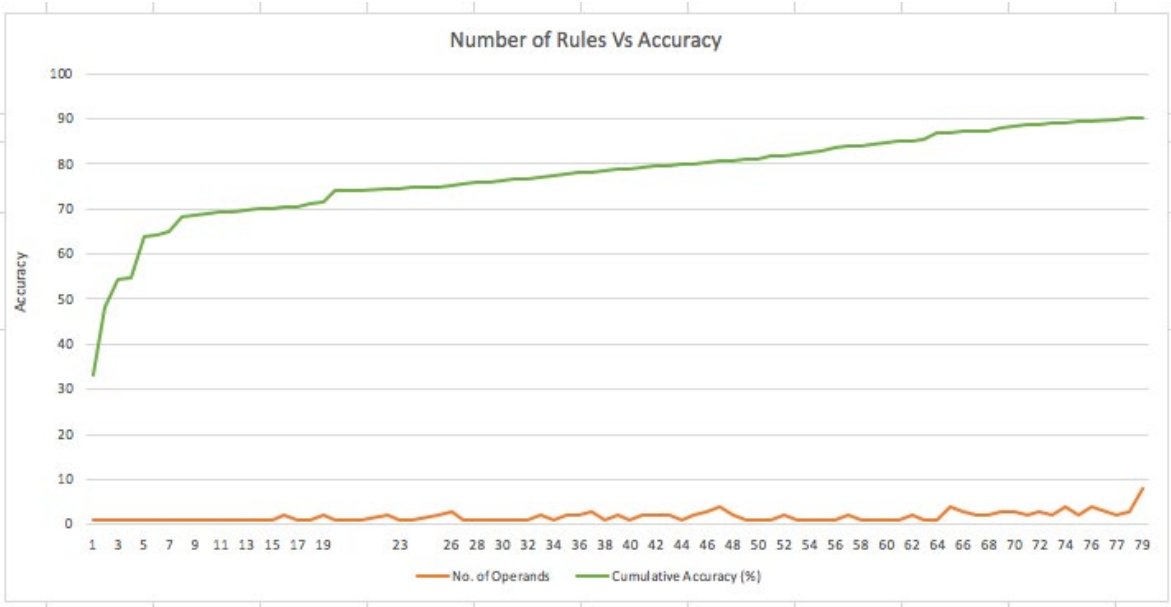

Figure 4. 1 Number of rules against accuracy as each rule is added, including number of attributes used for each rule. Source: Omar et al. (2022).

As Table 4.2 indicates, a total of 79 rules were written to obtain an accuracy rate of 90.2%. Figure 4.1 also indicates that I could have ceased writing new rules at approximately 45 rules and still obtain an 80% accuracy rate. It is also further evidence of the benefits of developing the KB during the data acquisition process, that is, incremental development. This approach is particularly useful in situations where socio-demographic data on type 2 diabetes are scarce (Omar et al., 2022).

The results obtained from both the training dataset and the production dataset highlight the KBS's strong performance in recognising patterns and making accurate predictions. The iterative refinement process facilitated the system's adaptability, enabling it to evolve with the introduction of new rules while maintaining a high level of accuracy. The observed differences between the

training and production datasets also underscore the challenges of generalization in KBSs, particularly in medical domains where unseen data can present unexpected variability.

The developed KBS demonstrated commendable performance, particularly given the limited and fragmented data available for its construction. Despite the inherent challenges posed by the scarcity of comprehensive datasets, the KBS achieved notable statistical outcomes, with high accuracy levels on both the training and production datasets. These results highlight the efficacy of the incremental KBS methodology employed during this research, underscoring its potential for developing robust systems even in data-constrained environments. This proof of concept establishes credibility for the methodologies and techniques utilised, validating the incremental approach as a viable strategy for constructing such systems.

However, it is important to acknowledge the region-specific nature of the developed KBS. Much of its functionality relies on locational attributes such as suburb names and postcodes, limiting its applicability to the Albury-Wodonga region. While this specificity was sufficient to meet the objectives of the current research, it underscores the need to transition towards a socio-demographic focus for broader applicability. The goal of this research is to establish socio-demographic factors that influence type 2 diabetes rather than geographic factors. Achieving this will require the development of a new KBS based on the principles and techniques demonstrated in this research, leveraging more universally applicable socio-demographic data.

4.5. Results for Geographically Specific Dataset (Training KBS at 100%)

Building upon the findings of the previous section, where the geographic KBS achieved a 90.2% accuracy rate on the training dataset using 79 rules, covering 451 out of the total 500 cases, the next phase involved evaluating its generalizability on unseen data. This step was crucial in assessing the robustness of the system beyond the dataset it was trained on, ensuring its applicability in real-world scenarios. A decision was made to further refine the system to achieve 100% accuracy on the geographic training dataset, which required the addition of 35 rules, bringing the total to 114 rules and covering all 500 cases. This approach aimed to evaluate the impact of complete accuracy on the training dataset when applied to the production dataset, which comprised unseen cases. By doing so, the research sought to gain deeper insights into the trade-offs between achieving perfect performance on a controlled dataset and the generalizability of the KBS when exposed to new, diverse data. The subsequent sections outline this extended

experimentation process, detailing the methodology employed, the results obtained, and an analysis of the outcomes.

As done in the previous section, the development process begins by examining the first case in the socio-demographic training dataset. The examination of the case involved reviewing the attributes and selecting a particular combination of values that leads to a specific conclusion, matching the actual conclusion for that case, that is, either diabetic or not diabetic. Once identified, a rule was written for this case and then run across the entire dataset. The accuracy rate was subsequently examined, along with the identification of the first case in the dataset where the recently added rule did not hold true. The attributes for that case were then compared to the most recently written rule for the last cornerstone case, looking for variations in attribute values between these two cases. A new rule was then written to address this variation in the classification output. This new rule then became the latest cornerstone case.

Training continued until the rule base reached 100% accuracy on the training set, requiring 114 rules (vs 79 rules to reach 90.2%). To isolate misclassifications we used simple data interrogation (filters, basic formulae, conditional formatting) to spot attribute combinations; single-rule attribute counts ranged from 1 to 8.

As discussed earlier, the rule construction process involved multiple iterations, each contributing to the incremental refinement of the KBS. Appendix C illustrates how the rules were progressively developed to increase the system's predictive accuracy, moving from an initial accuracy rate of 33% after the first rule to achieving a final accuracy of 100% after the 114th rule. The attributes used in rule formulation ranged from geodemographic factors such as patient postcode and patient suburb, to more nuanced combinations involving age, gender, family history of diabetes, and health indicators like BMI and cardiovascular status (CVD). This iterative refinement process exemplified the system's ability to integrate both generalizable patterns and detailed individual cases.

The results of this training phase underscore the system's potential for incremental development, particularly in domains where socio-demographic factors play a critical role in determining health outcomes. However, it remains essential to evaluate the system's performance on previously unseen data to gauge its generalizability. Table 4.4 summarises the KBS's performance on both the training and production datasets, highlighting its capacity to generalize predictions beyond the training phase. The training dataset refers to the data used to develop the KBS, while the production dataset represents an unseen dataset used to test its predictive performance.

	Accuracy	Specificity	Sensitivity
Training Dataset	100%	100%	100%
Production Dataset	63%	83.3%	39.3%

Table 4. 4 Results obtained during experiment for the training and production datasets using the geographically specific dataset (100% accuracy on the training KBS).

The accuracy, specificity, and sensitivity metrics presented in Table 4.2 demonstrate that while the KBS achieved perfect performance on the training dataset, with 100% accuracy, specificity, and sensitivity, its performance on the production dataset, comprising unseen cases, was markedly lower, with an accuracy rate of 63%. This significant drop in accuracy was accompanied by a steep decline in sensitivity (100% on the training dataset compared to 39.3% on the production dataset). The reduction in sensitivity indicates that the system struggled to identify true positive cases (i.e., correctly diagnosing diabetic patients) when applied to the production dataset. This discrepancy highlights the challenges posed by the inherent variability and diversity of unseen data, particularly when transitioning from a controlled training environment to real-world scenarios.

It is important to note that geodemographic factors, while valuable, are not the sole determinants of type 2 diabetes. Many other clinical and hereditary factors significantly contribute to the disease's onset but were not within the scope of this research. For instance, familial predispositions to type 2 diabetes play a critical role that is not captured by the geodemographic attributes used in this experiment.

Despite the decline in overall accuracy and sensitivity on the production dataset, the specificity of the system remained relatively robust, with 83.3% on the production dataset. This indicates that the KBS was still effective at identifying non-diabetic cases, demonstrating the strength of the rules developed for recognising patterns associated with the absence of diabetes.

The accuracy, specificity, and sensitivity metrics presented in Table 4.2 illustrate the nuanced performance of the geodemographic KBS. Sensitivity refers to the system's ability to correctly identify diabetic cases, or true positives, while minimising the number of diabetic cases it fails to identify, or false negatives. In contrast, specificity measures the system's ability to correctly identify non-diabetic cases, or true negatives, while minimising the number of non-diabetic cases incorrectly labelled as diabetic, or false positives.

In the context of type 2 diabetes detection, sensitivity is particularly important because it relates to the system's ability to detect individuals who are at risk. Missing a diabetic case (false negative)

could have severe consequences, as untreated diabetes can lead to complications such as cardiovascular disease, kidney failure, and vision loss. On the other hand, specificity reflects the system's capacity to correctly reassure non-diabetic individuals. While false positives (incorrectly labelling someone as diabetic) may cause unnecessary anxiety and follow-up tests, they are less harmful than false negatives. In layman's terms, it is better to err on the side of caution by identifying a person as diabetic when they are not, than to miss a case where someone genuinely has diabetes.

For the training dataset, the KBS achieved perfect sensitivity and specificity, correctly identifying all diabetic and non-diabetic cases. However, the performance on the unseen production dataset tells a different story. Sensitivity on the production dataset dropped to 39.3%, indicating that the system failed to identify a significant number of true diabetic cases. This decline highlights the challenge of generalising rules developed on a training dataset to new, unseen data. The specificity, while lower than in the training dataset, remained relatively robust at 83.3%, showing the system's effectiveness at identifying non-diabetic cases.

One contributing factor to the decline in sensitivity may be the distribution of the original dataset. The dataset was skewed, with approximately 30% diabetic and 70% non-diabetic cases. This imbalance likely influenced the rule development process, as the system had more opportunities to learn patterns associated with non-diabetic cases. Consequently, the KBS might be better at avoiding false positives than avoiding false negatives, which is reflected in the higher specificity relative to sensitivity on the production dataset.

These results underscore the importance of balancing sensitivity and specificity in the development of a decision-making system like this KBS. The skewed dataset and the inherent variability of socio-demographic factors in the production dataset highlight the need for additional refinements to improve sensitivity. For instance, integrating supplementary attributes, such as clinical or hereditary factors, could enhance the system's ability to identify true positives more effectively.

This trade-off between sensitivity and specificity reflects the inherent challenges posed by the skewed dataset (70% non-diabetics, 30% diabetics) and the complexity of geodemographic factors in predicting type 2 diabetes. While the KBS shows lower sensitivity on the production dataset, this limitation should be viewed in the context of the system's proof-of-concept nature. The primary aim of this research was to demonstrate the feasibility of integrating socio-demographic factors into a KBS for type 2 diabetes prediction, not to create a clinically deployable diagnostic tool at this stage.

Moreover, the observed specificity of 83.3% on the production dataset highlights the system's robustness in identifying non-diabetic individuals, which is critical in reducing unnecessary interventions. The low sensitivity, while a limitation, emphasises the need for future iterations of the KBS to incorporate additional clinical and lifestyle variables, as well as improved dataset balance, to enhance its capacity to detect diabetic cases. This iteration sets the stage for future refinements that can strike a better balance between sensitivity and specificity.

By considering both specificity and sensitivity, the geodemographic KBS provides valuable insights into the trade-offs involved in rule-based system design. While the system performs exceptionally well in controlled training environments, its limitations on unseen data suggest areas for further research and development to enhance its real-world applicability.

In summary, these results underscore the incremental and iterative nature of KBS development, demonstrating its potential while also identifying areas for targeted improvement. Future work will address the sensitivity gap through more balanced datasets, enhanced feature selection, and potential integration of clinical indicators.

The iterative refinement process was visualised in Figure 4.2, which depicts the relationship between the number of rules added and the cumulative accuracy achieved. The graph illustrates that achieving 100% accuracy on the training dataset required 114 rules, and it reflects the increasing complexity of the rules needed to cover edge cases and improve precision incrementally. The addition of these rules highlights both the adaptability of the system, and the trade-offs involved in balancing training accuracy with generalizability.

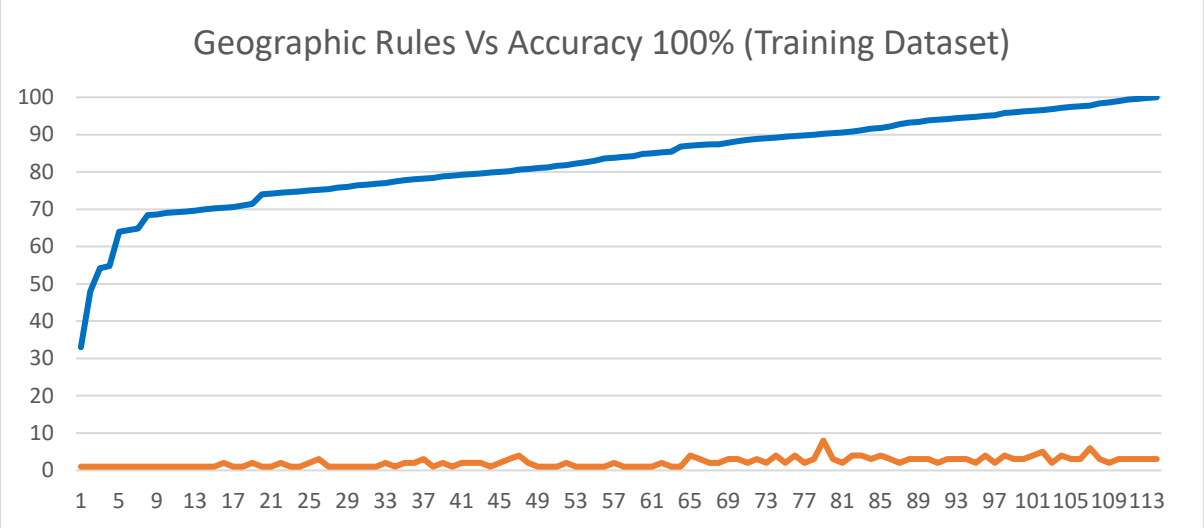

Figure 4. 2 Number of rules against accuracy as each rule is added, including number of attributes used for each rule

As Appendix C indicates, a total of 114 rules were written to achieve an accuracy rate of 100% on the training dataset. Figure 4.2 also shows that incremental accuracy gains were most significant during the earlier phases of rule construction, with approximately 50 rules achieving an accuracy rate close to 80%. This reinforces the advantages of incremental development during the KBS construction process, allowing for efficient progress even when faced with limited or fragmented data. Incremental development ensures that improvements are targeted, avoiding unnecessary complexity while maximising the system's ability to generalize.

The results from both the training dataset and the production dataset highlight the KBS's ability to adapt to known patterns and provide accurate predictions. However, achieving 100% accuracy on the training dataset came with trade-offs, as seen in the performance on the production dataset. The iterative refinement process enabled the KBS to handle diverse cases within the training environment, but the drop in performance when applied to unseen data underscores the complexities of generalization in knowledge-based systems. The stark decline in sensitivity on the production dataset, despite a respectable level of specificity, suggests challenges in accurately identifying true positives in unseen cases.

The developed KBS demonstrated impressive accuracy under controlled conditions, reflecting the robustness of its incremental design methodology. Despite the challenges posed by the geodemographic dataset's inherent variability, the system achieved notable statistical outcomes, validating the principles of iterative rule refinement and incremental development. These findings underscore the potential of this methodology to construct decision support systems that remain adaptable even in data-constrained settings.

Nevertheless, it is essential to acknowledge the limitations associated with achieving 100% accuracy on the training dataset. Overfitting to the training data may have contributed to the observed discrepancies in performance on unseen data, particularly the reduced sensitivity. This highlights the need for a balanced approach to rule construction, ensuring that the KBS remains generalizable and applicable beyond the training environment. Additionally, while this experiment focused on geodemographic attributes, Chapter 5 explores integrating clinical and socio-demographic factors to further enhance the KBS's predictive power and broaden its applicability.

Given the region-specific nature of the socio-demographic inputs, immediate applicability beyond Albury–Wodonga is limited. The results nevertheless indicate that the approach can be extended to larger, more diverse datasets with appropriate local validation. The following sections

therefore (i) assess performance on an unseen production dataset (section6.2) and (ii) compare RDR with ML baselines (sections 6.3–6.4), with implications summarised in section 6.5.

4.6. Accuracy Performance on the Production Dataset (training dataset at 100%)

To ensure comparability, I applied exactly the same RDR training and evaluation procedure described in section 4.4; the only change here is the input dataset (production vs. training). I therefore present just the outcomes—accuracy and rule counts—with detailed values in Appendix D and trends in Figure 4.3.

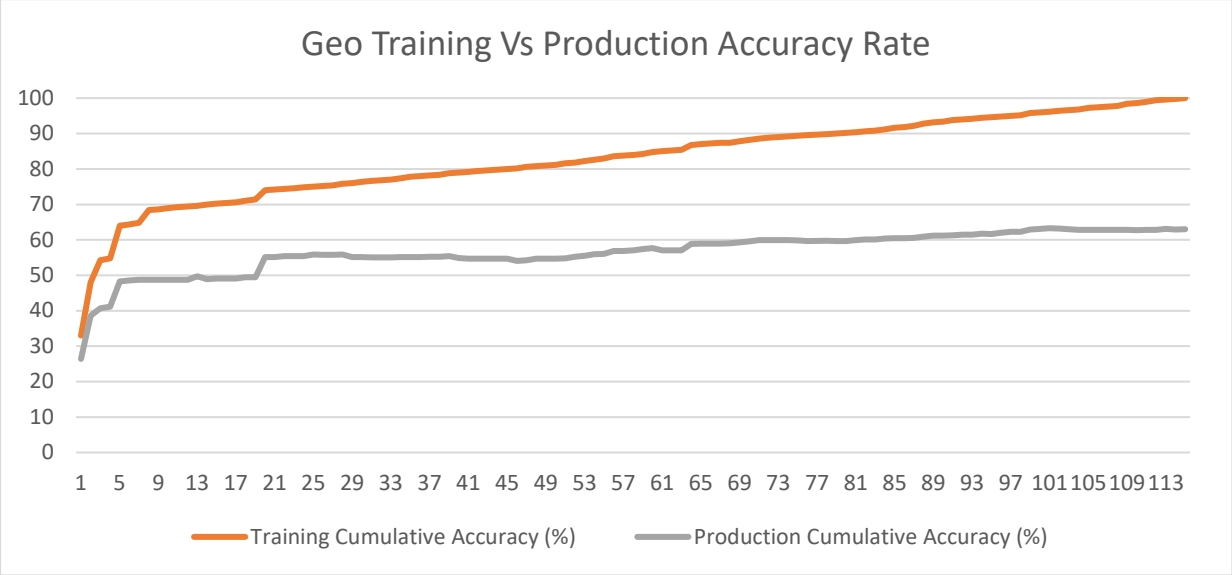

Figure 4. 3 Training against production dataset accuracy rate (Geographic KBS)

The results reveal a clear discrepancy between the performance of the KBS on the training dataset versus the production dataset. While the training dataset achieved perfect accuracy, the production dataset exhibited a gradual increase in accuracy before stabilising around 60% ± 5%. This tapering-off effect suggests that while the geographic KBS demonstrated some generalizability, it also encountered limitations when applied to unseen cases.

One explanation for the discrepancy between training and production performance is overfitting. The KBS was constructed to achieve 100% accuracy on the training dataset, meaning that the rules were highly optimised to fit the specific cases seen during training. However, these rules did not necessarily generalize well to the production dataset, which introduced novel cases that deviated from the training patterns.

Despite this, the KBS was not entirely ineffective on the production dataset. The stabilization of accuracy at approximately 60% indicates that the system retained some predictive capability, meaning that it captured some generalizable geodemographic patterns relevant to type 2 diabetes risk assessment. This outcome highlights the strengths and limitations of geodemographic attributes in KBS development. While geodemographic factors provide meaningful insights into disease risk, they are not definitive predictors on their own. Socio-demographic factors, explored in Chapter 5, will offer a broader perspective on improving the generalizability of the KBS.

The implications of these findings underscore the necessity of balancing training accuracy with real world applicability. Achieving 100% accuracy in a controlled dataset does not always translate to equivalent performance in practical applications. Future improvements may involve integrating socio-demographic and medical factors, refining rule conditions, or implementing hybrid methodologies that incorporate statistical validation alongside expert-driven rule formulation.

4.7 Chapter Summary

This chapter presented the processes involved in acquiring, preparing, and manipulating the datasets foundational to this research, specifically focusing on the geodemographic determinants of type 2 diabetes. By addressing the challenges in data sourcing, the meticulous modifications required, and the data manipulation techniques applied, this chapter established a robust framework for the subsequent construction and analysis of the KBS.

The data acquisition process revealed significant challenges, including the fragmented availability of geodemographic and medical data, as well as legal and ethical constraints related to data privacy. Innovative strategies, including extensive research into data sourcing and exploring methodologies employed by healthcare organizations, were implemented to address these barriers while maintaining the integrity and relevance of the acquired data.

In the data preparation phase, critical modifications were made to refine the dataset for this study. These included the exclusion of type 1 diabetes data to focus solely on type 2 diabetes, and the anonymization of personal identifiers to meet privacy standards while still allowing for detailed geodemographic analysis. The chapter also detailed how the dataset was adapted for the development of a geodemographic KBS, emphasising the importance of location-based attributes such as street names, suburbs, and postcodes. These steps ensured the datasets were tailored to meet the research objectives and provide a strong basis for future studies.

Section 4.4 reports on the results of an experiment to evaluate the initial geodemographic KBS framework. This experiment aimed to assess the KBS's ability to utilize geodemographic and clinical data to identify individuals at risk of type 2 diabetes. The focus was on analysing patterns in geodemographic attributes, including regional healthcare access and environmental factors, with some consideration of socio-demographic associations that aligned with certain geographic regions. Rule creation and refinement based on these geodemographic patterns demonstrated the system's capability in distinguishing between diabetic and non-diabetic cases effectively.

Section 4.5 detailed the iterative refinement process that allowed the geodemographic KBS to reach 100% accuracy on the training dataset. The iterative refinement of the KBS achieved a notable milestone with 100% accuracy on the training dataset, showcasing the system's capability to model complex geodemographic relationships with precision. However, the performance on the unseen production dataset highlighted the inherent challenges of generalization, with an accuracy rate of 63%. This underscores the variability and complexity of unseen data in geodemographic modelling.

Additionally, the results revealed a pronounced imbalance between specificity (83.3%) and sensitivity (39.3%) on the production dataset, indicating the system's strength in identifying non-diabetic cases but challenges in recognising diabetic cases accurately. The results reaffirm the value of incremental KBS development, particularly in contexts with limited geodemographic data availability, where rule-based systems can provide meaningful insights despite data constraints.

Section 4.6 detailed a comparison between the geodemographic KBS on the training and the production dataset (unseen dataset). The results obtained from the training dataset demonstrated the efficacy of incremental rule development, achieving 100% accuracy. However, testing on the production dataset revealed a significant reduction in accuracy, stabilising at around 60%. This suggests that while the geodemographic KBS captured meaningful patterns, its reliance on location-based attributes introduced generalization limitations. These insights set the foundation for Chapter 5, which explores the integration of socio-demographic attributes to enhance the adaptability and predictive accuracy of the KBS beyond purely geodemographic indicators.

The experiment results highlighted the effectiveness of integrating geodemographic determinants into the KBS framework, showing substantial improvements in predictive accuracy compared to using clinical data alone. This confirmed the importance of considering geodemographic factors, such as proximity to healthcare facilities and urban versus rural environments, in assessing type

2 diabetes risk. The findings also indicated areas for further refinement, which will be addressed in the experiments detailed in Chapter 5.

The chapter transitions into the experimental phase in Chapter 5, which discusses how the refined datasets are used to test various hypotheses on the influence of socio-enhanced demographic factors on type 2 diabetes risk. The experiments provide empirical evidence to support or refute these hypotheses, deepening our understanding of the disease. Additionally, these experiments confirm the credibility and proof of concept of the novel technique of incremental KBS development, which involves constructing the KBS during the knowledge acquisition phase, as highlighted by Beydoun and Hoffman (2013). This approach not only validates the methodology but also demonstrates its utility in handling complex datasets and refining knowledge systems in real time (Beydoun and Hoffmann, 2013). Furthermore, the insights gained will contribute to developing targeted public health policies and interventions, tailored to address the specific socio-enhanced demographic factors and environmental needs of affected populations.

While the chapter outlines significant progress in constructing the KBS for type 2 diabetes, it also identifies notable limitations. The current dataset's restricted inclusion of detailed socio-demographic attributes, such as household income and educational attainment, limits the system's capacity to fully capture the interplay between social determinants and health outcomes. Additionally, the lack of integration between socio-demographic and clinical health records presents a challenge for holistic analysis. Addressing these gaps in future research and data collection efforts will be crucial for enhancing the KBS's predictive accuracy and its utility in public health strategies.

Another significant limitation of the current KBS is its reliance on location-specific variables, such as suburb names, streets, and postcodes. While these geodemographic factors contributed to good predictive results within the context of the dataset, they render the KBS highly location dependent. Consequently, the system is most effective in the specific area where the data was collected but becomes less applicable in other regions where geodemographic characteristics differ. This limitation underscores the need for a more generalised approach to KBS development, one that reduces reliance on location-specific data.

To address this limitation, Chapter 5 focuses on the construction of KBS that prioritises socio-demographic attributes over geodemographic attributes. By leveraging universally relevant factors such as socio-demographic attributes, education levels, and housing stability, the next KBS will aim to deliver a system that can be applied across various regions where these determinants are

well understood. This transition marks a critical step toward creating a more robust and versatile tool for predicting type 2 diabetes risk, with broader implications for public health.

This new system aims to address the constraints imposed by the reliance on location-based attributes, such as suburb names and postcodes, which restrict the geodemographic KBS's applicability beyond the Albury-Wodonga region. Building on the principles and techniques demonstrated in this chapter, the socio-demographic KBS will leverage broader and more universally applicable data to further refine the assessment and management of type 2 diabetes risk.

While the geodemographic KBS demonstrated commendable accuracy and specificity, its development is not without limitations. Firstly, its applicability is restricted to the Albury-Wodonga region, as it relies heavily on region specific geographic attributes such as postcodes and suburb names. This specificity limits the system's ability to generalize beyond the dataset used in its construction. Additionally, the absence of socio-demographic factors in the dataset creates a critical gap, as these attributes often play a significant role in determining the risk of developing type 2 diabetes. For example, attributes such as median household income, education levels, and average family size provide a more nuanced understanding of health determinants that cannot be captured by geodemographic factors alone.

Another limitation arises from the challenges encountered during the construction process, including handling missing data, managing the balance between sensitivity and specificity, and addressing the inherent variability in unseen datasets. These challenges, though partially mitigated, highlight the need for a more robust and generalizable system that extends beyond geodemographic data alone.

To address these limitations, Chapter 5 focuses on the development of a socio-demographic KBS. By incorporating attributes that capture socio-economic and demographic factors, this approach aims to create a system with broader applicability, capable of transcending geodemographic constraints. Furthermore, it will build on the iterative methods and incremental development principles established in this chapter, adapting them to a new data paradigm. This next step in the research will further validate the efficacy of the KBS methodology while addressing the challenges and limitations identified.

Chapter 5: Socio-Demographic Knowledge-Based System

This chapter focuses on the development of a knowledge-based system (KBS) based on socio-demographic factors. By leveraging attributes such as average family size, median house price, average income, and highest education level attained within specific areas, this chapter outlines the methodology for constructing a system capable of utilising these socio-demographic determinants for predictive modelling. In the previous chapter, the development of a geodemographic -based KBS was detailed, demonstrating how location-specific factors such as postcodes and suburbs can assist in predicting type 2 diabetes risk. While this approach proved effective, it has inherent limitations, particularly in its applicability beyond the geodemographic region from which the data was collected. A geodemographic -based KBS is inherently region-specific, meaning it cannot be easily transferred to other locations without significant adaptation. That is, the geodemographic based KBS is inherently limited in its applicability to regions beyond the Albury-Wodonga area, as its rules are tightly coupled to the specific geodemographic data from that region. This localised focus restricts the generalizability of the system, making it less effective in addressing broader public health challenges.

To address this limitation, this chapter focuses on the incremental development of a socio-demographic KBS, replacing geodemographic indicators with socio-demographic attributes that offer a more generalizable and scalable approach. Instead of relying on location-specific identifiers, this system integrates factors such as median household income, average family size, education levels, and housing affordability socio-economic determinants that influence diabetes risk regardless of geodemographic location. By shifting from location-based indicators to socio-demographic patterns, the new KBS provides a framework that can be applied globally, making it a more flexible and transferable model for diabetes risk prediction.

Attribute	Description
Case Id	Patient case number. Integer number used to identify the case.
Patient ID	Code to protect patient confidentiality
Analysis_Id	Code to record a particular consultation
Date Attended	Date the patient attend a GP's office or a medical centre.
Times Attended	Times a patient attend GP's office
Withdrawn Screening	Patient deciding to withdraw medical screen. (Yes/No)
Age	Patient's age
Median Age	Median age of people in a particular region
Median Household Income	Median Household Income (Monthly)
Median Mortgage	Median Mortgage Payments (Monthly)
Median Rent	Median Rent (Monthly)
Average Household Size	Average family size (Parents & children)
Highest education level (Male)	Highest education level achieved by a male
Highest education level (Female)	Highest education level achieved by a female
Diagnostic DM (years)	Years diagnosed with type 2 diabetes. (Years)
CVD Status	Cardiovascular Disease status (Yes/No)
Diagnostic CVD (years)	How long a patient has been diagnosed with cardiovascular disease. (Years)
HT Status	Hypertension status of a patient. (Yes / No)
Diagnostic HT (years)	Years diagnosed with hypertension (Years)
Alcohol	If a patient consumes alcohol. (Yes/No)
Family History DM	If a patient has a patient has a family history of diabetes. (Yes / No)
Family History CVD	If a patient has a family history of cardiovascular disease (Yes/No)
PHQ9	Measure of depression. (range between 5 to 20)
Diet	If a patient adheres to a particular diet to manage type 2 diabetes. (Yes/No)
Exercise Duration (hrs/week)	Exercise duration per week. (Hours).
Exercise Intensity	How intense a patient exercises. (Range between None, Low to High)
Last visit to GP	Last time a patient visited a GP. (Months)
Last visit to diabetes educator	Last visit to a diabetes educator. (Months)
Frequency yr GP	How often a patient visits a GP per year.
Frequency (yr) diabetes educator	How often a patient visits a diabetes educator. (per year)
Waist Circumference	Waist circumference (Cm)
Height	Patient's height. (Cm)
Weight	Patient's weight. (Kg)
BMI	Patient's BMI value.
Target	Expected DM result as per medical records. That is, diabetic or not. This was the expected target for each ripple down rule written. (Yes / No).
Conclusion	The KBS conclusion based on the rule

Table 5. 1 Attributes used in the development of the socio-demographic KBS (90.2% accuracy)

This chapter follows the same structured approach outlined in Chapter 4, detailing dataset preparation, attribute selection, rule development, iterative refinement, and evaluation. However, it emphasises the distinct challenges and advantages associated with socio-demographic data, demonstrating how this approach expands the applicability and adaptability of the KBS beyond the limitations of geographic data. That is, the incremental construction approach remains consistent with that of the geographic-based KBS described in Chapter 4. The primary distinction lies in the replacement of location-based identifiers such as street names, postcodes, and suburbs with socio-demographic attributes associated with those regions. These attributes provide a richer understanding of the social and economic environment, allowing the KBS to evaluate the interplay between these factors and type 2 diabetes risk. Socio-demographic attributes were chosen as more holistic indicators of type 2 diabetes risk. Table 5.1 provides a detailed breakdown of the attributes incorporated in the socio-demographic KBS, highlighting key socio-economic indicators such as median household income, education levels, and family size, which were found to have a significant correlation with diabetes risk.

This chapter is divided into six sections. Section 5.1 provides an overview of the socio-demographic KBS and outlines the chapter's structure. Section 5.2 discusses the experimental platform used for integrating and processing socio-demographic data, highlighting its adaptability for this context. Section 5.3 delves into the tailored rule development process, detailing the methodologies and techniques employed to create a robust KBS based on socio-demographic factors. Section 5.4 focuses on real-time feedback mechanisms and their role in refining the system. Section 5.5 evaluates the application of the socio-demographic KBS on the production (unseen) dataset, discussing its performance and comparing it against the geographic KBS from Chapter 4. Finally, Section 5.6 summarises the chapter's key findings and sets the stage for the results presented in the next chapter.

5.1 A Platform for Knowledge Integration

The role of socio-demographic factors in health outcomes is well-documented, with studies (Hill et al. 2013; Shaw, 2004) highlighting the impact of income inequality and educational attainment on chronic disease prevalence. Unlike static geographic indicators, socio-demographic attributes offer deeper insights into the systemic barriers affecting Type 2 diabetes risk. By incorporating these attributes, the KBS moves beyond location-based associations to capture the broader social

determinants that influence health outcomes, allowing for more targeted and scalable interventions.

Building on Chapter 4, this chapter develops the socio-demographic KBS and notes only the method differences from the geographic model (variable selection, standardisation, and rule induction). We focus on how socio-demographic attributes are operationalised and evaluated, with full geographic details remaining in Chapter 4.The socio-demographic dataset, sourced primarily from the Australian Bureau of Statistics (ABS) via Albury-Wodonga council databases, forms the foundation of this KBS. Unlike the geographically constrained data in Chapter 4, socio-demographic attributes provide a broader lens through which health disparities can be analysed. The selected attributes, including median household income, educational attainment, and family size, capture systemic socio-economic patterns that influence diabetes risk. This shift ensures that the KBS is not only data-driven but also adaptable across different regions, overcoming the geographic limitations of the previous model.

The experimental platform remains integral to this development process, serving as the foundation for integrating, analysing, and refining socio-demographic data. This platform has been adapted to accommodate the nuanced nature of socio-demographic attributes, ensuring seamless data processing and rule generation. By enabling iterative refinements and real-time feedback, the platform supports the dynamic construction of a robust KBS tailored to socio-demographic determinants.

Adapting the platform for socio-demographic data posed unique challenges compared to its geographic counterpart. Unlike fixed attributes like street names and postcodes, socio-demographic factors often encompass variable, aggregated data points (e.g., median income ranges or average household size). To address this, the platform was enhanced to integrate datasets from multiple sources while standardising these attributes for consistency. Additional filtering techniques were introduced to ensure the accuracy and relevance of the processed data. A key observation during the socio-demographic KBS development was the extensive reliance on socio-economic factors for rule construction. This reliance became particularly evident in the latter stages of development when refining the system from 90.2% to 100% accuracy on the socio-demographic training dataset. In some cases, rules were written to cover specific cases, followed by counter-rules with opposite conclusions to account for subsequent cases. This iterative refinement effectively redirected the KBS traversal within the decision tree. When the system reached a leaf node, additional rules were introduced to redirect classification along alternative branches, allowing for greater adaptability in handling complex scenarios. This broader

applicability ensures that the socio-demographic KBS aligns with the research's fundamental goal: to develop a predictive model that is not bound by geographic limitations but is instead universally adaptable to different populations and settings.

The subsequent sections delve into the specifics of this platform, detailing its design, functionality, and application in constructing the KBS. Emphasis is placed on the integration of new data types and the tools utilised to ensure consistency and accuracy throughout the development process.

5.2 Tailored Rule Development

The tool developed by Compton and Kang (2021) was adapted in the construction of both the geographic KBS discussed in Chapter 4 and the socio-demographic KBS introduced in this chapter. While its application in the geographic KBS facilitated iterative rule development using location-based attributes such as postcodes and suburbs, its role in the socio-demographic KBS required modifications to accommodate broader, non-static attributes. This section expands upon its application in the socio-demographic context, outlining the adaptations necessary for integrating complex socio-economic data and discussing the challenges encountered.

In this thesis we use demographic variables to denote fixed, individual-level descriptors—age, sex and biological family history—whereas socio-demographic variables capture the social and economic context in which individuals live (e.g., household income, educational attainment, occupation and housing stability). Structurally, demographic attributes are stored as discrete, well-defined fields, while socio-demographic attributes are usually aggregated into ranges or composite indices that require normalisation, imputation and hierarchical encoding before rule construction. Recognising this asymmetry, we preprocess socio-demographic data into interval-based features and implement nested RDR conditions that can accommodate overlapping categories. Explicitly separating these two layers mitigates collinearity, improves rule interpretability and aligns the knowledge base with established models of the social determinants of diabetes risk outlined by Hill, Nielsen & Fox (2013)."

The Ripple-Down Rules (RDR) tool evolved from an SQL-based system into an Excel-based platform, providing greater flexibility for handling socio-demographic attributes such as median household income, education levels, and average family size. Unlike geographic attributes, which are static and location-bound, socio-demographic attributes are inherently adaptable and can be applied across different regions. This transition allowed for seamless data integration and improved scalability, enabling a more universal application of the KBS framework. As shown in

Figure 5.1, the RDR tool offers an intuitive interface that allows domain experts without extensive programming knowledge to contribute to the system's incremental development. This adaptability makes it an ideal choice for constructing a socio-demographic KBS that can be expanded and refined in response to evolving public health needs.

3	Patient_Post code	id	patient_ID	analysis_ID	Date_Attended	Times_Attended	Withdrawn_ Screening	Gender	Patient_Age	Median Age	Median Total Family Income (Monthly)	Median Mortgage Payments (Monthly)	Median Rent (Monthly)	Avg. Household Size	Highest Education (Male)	Highest Education (Female)	Diagnostic_ DM_years	CVD_Status	Diagnostic_ CVD_years	HT_Status
4	2640	4	ALLD071250	ALLD071250_	38169	2	Unknown	M	21	39	5252	1517	1000	2.4	Cert III & IV	Bachelor Degr	0	No	0	No
5	2641	6	ALLD071250	ALLD071250_	13/12/06	4	Unknown	M	23	40	5056	1300	860	2.2	Cert III & IV	Cert III & IV	0	No	0	No
6	2641	8	ALLD071250	ALLD071250_	40878	6	Unknown	M	27	40	5056	1300	860	2.2	Cert III & IV	Cert III & IV	0	No	0	No
7	2641	10	ALLE160441	ALLE160441_	15/7/03	1	Yes	F	29	40	5056	1300	860	2.2	Cert III & IV	Cert III & IV	0	No	0	No
8	2640	12	ALLE260846	ALLE260846_	26/6/05	2	Unknown	M							Cert III & IV	achelor Degre	0	No	0	No
9	5000	16	ALLE260846	ALLE260846_	4/12/12	7	Unknown	M							achelor Degre	achelor Degre	0	No	0	No
10	3690	29	AMES190356	AMES190356_	30/10/07	1	Unknown	F							Cert III & IV	Cert III & IV	0	No	0	No
11	2640	34	ANDJ140751	ANDJ140751_	5/12/06	4	yes	F							Cert III & IV	achelor Degre	0	No	0	No
12	2640	35	ANDJ201142	ANDJ201142_	29/6/05	1	Unknown	F							Cert III & IV	achelor Degre	0	No	0	No
13	2640	38	ARBR260540	ARBR260540_	16/8/06	3	Unknown	M							Cert III & IV	achelor Degre	0	No	0	No
14	2640	40	ARBR260540	ARBR260540_	17/6/14	5	Unknown	M							Cert III & IV	achelor Degre	0	No	0	No
15	3749	47	ARTI250334	ARTI250334_	14/9/12	7	Unknown	F							Cert III & IV	achelor Degre	0	No	0	No
16	3690	48	ARTJ240134	ARTJ240134_	9/7/04	2	Unknown	M							Cert III & IV	Cert III & IV	0	No	0	No
17	3747	49	ARTJ240134	ARTJ240134_	28/6/05	3	Unknown	M	39	49	5836	1428	980	2.3	Cert III & IV	achelor Degre	0	No	0	No
18	3690	60	ARUJ020429	ARUJ020429_	15/6/11	7	Unknown	F	41	37	5924	1387	1000	2.4	Cert III & IV	Cert III & IV	0	No	0	No
19	3749	62	ARUM010128	ARUM010128	9/12/02	1	Yes	F	41	48	6808	1560	1080	2.5	Cert III & IV	achelor Degre	0	No	0	No
20	2640	65	ATKJ010156	ATKJ010156_	16/7/03	1	Yes	M	41	39	5252	1517	1000	2.4	Cert III & IV	achelor Degre	0	No	0	No
21	2640	69	BABB110444	BABB110444_	6/2/12	4	Unknown	M	42	39	5252	1517	1000	2.4	Cert III & IV	achelor Degre	0	No	0	No
22	3747	85	BAKC120837	BAKC120837_	5/11/09	1	Unknown	M	42	49	5836	1428	980	2.3	Cert III & IV	achelor Degre	0	No	0	No
23	3747	88	BAKR170122	BAKR170122_	13/7/04	1	Yes	M	43	49	5836	1428	980	2.3	Cert III & IV	achelor Degre	0	No	0	No
24	3690	89	BALJ281122	BALJ281122_	17/9/04	1	Unknown	M	43	37	5924	1387	1000	2.4	Cert III & IV	Cert III & IV	0	No	0	No
25	2640	90	BALJ281122	BALJ281122_	27/11/06	3	Unknown	M	43	39	5252	1517	1000	2.4	Cert III & IV	achelor Degre	0	No	0	No
26	2640	91	BALJ281122	BALJ281122_	21/8/07	4	Unknown	M	43	39	5252	1517	1000	2.4	Cert III & IV	achelor Degre	0	No	0	No
27	3690	98	BARE280757	BARE280757_	19/3/08	4	Unknown	F	43	37	5924	1387	1000	2.4	Cert III & IV	Cert III & IV	0	No	0	No
28	3688	99	BARE280757	BARE280757_	28/10/09	5	Unknown	F	43	41	6868	1387	1020	2.6	Cert III & IV	Cert III & IV	0	No	0	No

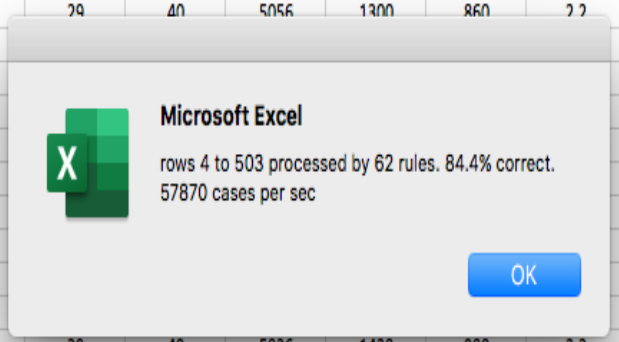

Figure 5. 1 The Excel rule builder output results, showing the accuracy rate, cases processed and speed of the processing.

A significant feature of the tool is its ability to provide real-time statistical feedback on a rule performance. After each rule is added or modified, the system generates a summary detailing the number of cases covered, the accuracy rate, and execution time. Given section 5.2.1 notes the higher variability of socio-demographic attributes, this real-time feedback expedites threshold-setting during rule refinement. As shown in Figure 5.1, this statistical summary enables users to assess the effectiveness of newly introduced rules, ensuring that the KBS adapts dynamically to evolving patterns in socio-demographic data. The tool highlights cases that do not conform to the most recently written rule. Unlike geographic attributes, which tend to follow more rigid patterns, socio-demographic data can present overlapping or ambiguous cases requiring iterative refinement. As depicted in Figure 5.2, cases highlighted in pink represent those that failed to align with the newly developed rule, signalling areas where further refinements or additional conditions may be necessary. This immediate visual feedback allows the KBS developer to quickly pinpoint inconsistencies and refine the system, accordingly, making the rule construction process more efficient and targeted.

Patient_Postcode	id	patient_ID	analysis_ID	Date_Attended	Times_Attended	Withdrawn_Screening	Gender	Patient_Age	Median Age	Median Total Family Income (Monthly)	Median Mortgage Payments (Monthly)	Median Rent (Monthly)	Avg. Household Size	Highest Educatio (Male)
3690	1522	GILR311058	GILR311058_1	12/7/04	1	Unknown	F	64	37	5924	1387	1000	2.4	Cert III &
3690	1536	JOHL130860	JOHL130860_	29/4/04	1	Unknown	F	71	37	5924	1387	1000	2.4	Cert III &
2641	1544	MARL090239	MARL090239_	1/4/05	2	Unknown	F	76	40	5056	1300	860	2.2	Cert III &
2640	1554	POPD030860	POPD030860_	13/7/04	1	yes	F	45	39	5252	1517	1000	2.4	Cert III &
2640	1560	ROOJ211037	ROOJ211037_	29/4/04	2	not interested	M	58	39	5252	1517	1000	2.4	Cert III &
2641	1561	RUMJ101058	RUMJ101058_	7/8/14	1	Unknown	F	60	40	5056	1300	860	2.2	Cert III &
3690	1562	RYAK080229	RYAK080229_	20/11/07	2	Unknown	F	60	37	5924	1387	1000	2.4	Cert III &
2641	1568	StEJ190439	StEJ190439_3	5/4/05	3	Unknown	F	66	40	5056	1300	860	2.2	Cert III &
2641	1570	StRE021028	StRE021028_1	4/4/05	3	Unknown	F	67	40	5056	1300	860	2.2	Cert III &
3690	1573	SYMM260538	SYMM260538	5/11/07	4	Unknown	F	68	37	5924	1387	1000	2.4	Cert III &
3690	1580	WILP260533	WILP260533_	5/9/07	5	Unknown	F	79	37	5924	1387	1000	2.4	Cert III &
2640	1581	WYLD010736	WYLD010736_	4/4/07	1	Unknown	F	82	39	5252	1517	1000	2.4	Cert III &
2640	1582	ZANA230361	ZANA230361_	24/4/06	2	Unknown	F	85	39	5252	1517	1000	2.4	Cert III &
2640	1583	ZERP010226	ZERP010226_	5/9/07	3	Unknown	F	90	39	5252	1517	1000	2.4	Cert III &
2641	1595	CLIJ290138	CLIJ290138_7	17/2/12	7	Unknown	F	56	40	5056	1300	860	2.2	Cert III &
2642	1596	CONM010835	CONM010835	14/5/08	4	Unknown	F	56	42	6364	1300	740	2.7	Cert III &
3690	1598	COUJ020140	COUJ020140_	7/7/11	5	Unknown	F	57	37	5924	1387	1000	2.4	Cert III &
2640	1606	FORN030139	FORN030139_	9/7/04	1	Unknown	F	62	39	5252	1517	1000	2.4	Cert III &
3749	1609	GRIE070126	GRIE070126_	10/9/08	5	Unknown	F	65	48	6808	1560	1080	2.5	Cert III &
2640	1633	MILD300840	MILD300840_	16/1/03	1	Unknown	M	79	39	5252	1517	1000	2.4	Cert III &
2640	1634	MILD310750	MILD310750_	15/7/04	1	Unknown	F	79	39	5252	1517	1000	2.4	Cert III &
2640	1639	PARM010172	PARM010172_	17/9/03	1	Yes	M	90	39	5252	1517	1000	2.4	Cert III &
2641	1646	SCOI160928	SCOI160928_4	20/8/08	4	Unknown	F	63	40	5056	1300	860	2.2	Cert III &
3700	1662	WUEH230329	WUEH230329	9/7/04	1	Unknown	F	82	48	5548	1083	760	2.3	Cert III &
3747	1663	ALLR090437	ALLR090437_	28/6/05	3	Unknown	M	34	49	5836	1428	980	2.3	Cert III &
2641	1666	CAUJ120522	CAUJ120522_	18/10/06	3	Unknown	F	54	40	5056	1300	860	2.2	Cert III &
2641	1671	DOUN100257	DOUN100257	13/8/08	1	Unknown	M	59	40	5056	1300	860	2.2	Cert III &

Figure 5. 2 The Excel rule builder mismatch cases output, showing cases that do not conform to any rules in the KBS.

This visual representation not only identifies inconsistencies but also enables the user to immediately recognize where the most recent rule fails. By highlighting these deviations, the tool accelerates the debugging and improvement process, ensuring rapid and efficient refinement of the rule set. This feature is particularly valuable in a socio-demographic KBS, where broader data variability increases the likelihood of subtle misclassifications. Since socio-demographic factors are often aggregated (e.g., median household income, educational attainment), real-time detection of classification errors is critical in maintaining system reliability. The RDR tool's ability to dynamically highlight these inconsistencies significantly enhances both the accuracy and adaptability of the KBS. These capabilities proved particularly beneficial when adapting the tool to handle socio-demographic data. As geographic attributes such as "street name" and "postcode" were replaced with socio-demographic equivalents like "median household income" and "education level," the statistical outputs provided insights into how well these new variables captured patterns in the dataset. This iterative feedback process was essential in refining rule construction, ensuring that the rules remained both statistically valid and generalizable across different population groups. Unlike geographic indicators, which are inherently region-dependent, socio-demographic attributes enable the system to be applied more broadly, increasing the model's scalability beyond a specific geographic area. To support socio-demographic inputs, we standardised range-based fields and recalibrated rule predicates for aggregated/overlapping categories while preserving interpretability..

5.2.1 Challenges and Adaptations

This subsection details the concrete adaptations to the RDR tool required by socio-demographic variables. Unlike geographic identifiers (discrete and location-bound), socio-demographic variables are often ranges or aggregates, necessitating standardisation before rule induction.

Adapting the RDR tool for socio-demographic data required overcoming several unique challenges. Unlike geographic identifiers, which are discrete and location-bound, socio-demographic attributes often encompass ranges, averages, or aggregated values rather than fixed, easily distinguishable indicators. This introduced significant preprocessing requirements, as the data required normalization and standardization to ensure consistency across different population groups, particularly when integrating information from multiple regional datasets.

Furthermore, the system's logic was recalibrated to accommodate the nuanced nature of socio-demographic data. For example, rules based on income brackets, household sizes, or education levels needed greater flexibility compared to rigid geographic attributes like postcodes, which have a one-to-one correspondence to locations. These modifications ensured that the rules generated by the RDR tool remained interpretable while effectively capturing complex socio-demographic patterns. Additionally, adjustments were made to ensure that overlapping socio-demographic categories (e.g., income levels spanning multiple classifications) were handled dynamically, reducing classification ambiguities.

5.2.2 Integration with Visual Dashboards

To further enhance its functionality, the RDR tool was complemented with interactive visual dashboards. These dashboards enabled KBS developers to monitor the distribution of socio-demographic attributes and assess their impact on rule performance in real time. For example, graphical representations allowed users to quickly identify outliers or anomalies in the data, which could skew the KBS's predictions if left unaddressed.

Accordingly, preprocessing standardised brackets (e.g., income, household size) to ensure consistent rule predicates across datasets. The visual dashboards were particularly useful in this context, as they facilitated the exploration of these complex relationships, helping developers understand the indirect and multifaceted impact of socio-demographic factors on type 2 diabetes risk.

As seen in Figures 5.1 and 5.2, the dashboard provided clear, actionable insights into the relationship between specific socio-demographic factors and diabetes risk. This feature was instrumental in refining rules, ensuring that each new addition to the KBS was both robust and relevant to the data at hand.

5.2.3 Transition to Tailored Rule Development

The platform's ability to integrate socio-demographic data and provide iterative feedback laid the groundwork for the tailored rule development process detailed in the next section. The interpretive and explorative stance in the formulation of socio-demographic data required a more flexible and refined approach to rule construction. By leveraging the insights gained from the tools, the aim is to construct a highly adaptive KBS that combines computational precision with

generalisation capability. The iterative refinements enabled by real-time feedback ensured that rule development remained both responsive and aligned with the complexities of socio-demographic determinants.

A key feature of the RDR platform was its ability to generate real-time performance metrics, significantly enhancing the iterative refinement process of the KBS. Metrics such as rule accuracy, speed of execution, and dataset coverage can be dynamically updated after each rule iteration, providing immediate insights into system performance (Beydoun and Hoffmann, 2013). This real-time feedback streamlined the decision-making process for refining rules, ensuring continuous improvement in classification accuracy. Using these performance metrics facilitated direct comparisons between the deployment of socio-economic attributes and their absence in the knowledge base developed in Chapter 4. This cross-comparison will also evaluate the system's scalability and adaptability, transitioning from location-based attributes to socio-demographic determinants.

One of the platform's most valuable outputs was the rule construction table, which systematically documented the impact of each rule iteration on dataset coverage and accuracy. These tables served as critical references for assessing model refinements and pinpointing areas where additional rules were required. The structure and insights derived from these tables are further elaborated in Section 5.3, where the iterative rule development process is discussed in detail.

By integrating complex datasets, offering real-time performance tracking, and providing visualised performance insights, the experimental platform proved to be a powerful tool for socio-demographic KBS construction. The ability to bridge computational precision with expert-driven insights underscored the platform's versatility, not only for type 2 diabetes prediction but also for broader decision support applications requiring nuanced, data-driven models.

5.3: Tailored Rule Development (90.2% accuracy on the training dataset)

This analysis operationalises the socio-ecological model of diabetes risk (Hill, Nielsen & Fox, 2013) by mapping each rule/metric to its corresponding social determinant.

As foreshadowed, the development of the socio-demographic KBS required notable adaptations to the knowledge acquisition process to address the fluid nature of socio-demographic data. The process still began by reviewing the first case in the socio-demographic Training dataset, evaluating its attributes to construct an initial rule for classifying cases as diabetic or not diabetic.

This rule was then applied across the dataset to assess its effectiveness, with subsequent refinements introduced based on cases that remained misclassified. However, unlike the geographic KBS, which relied on location-based variables such as street names, postcodes, and suburbs, the socio-demographic dataset integrated broader contextual factors, including median household income, family size, and education levels. These attributes required additional processing to standardize and normalize their representation within the KBS framework. The shift from discrete geographic identifiers to aggregated, range-based socio-demographic factors introduced new challenges in defining precise rule conditions. For instance, income brackets and household sizes often overlapped, necessitating greater precision to prevent ambiguity in classification. Additionally, imputation techniques were applied to handle missing values, ensuring that incomplete data did not compromise system accuracy. Specifically, attributes with more than 90% missing values were discarded and replaced with closely related attributes that contained a higher percentage of complete data. This preprocessing step helped preserve dataset integrity while maintaining socio-demographic relevance, allowing the KBS to capture broader patterns of type 2 diabetes risk.

5.3.1 Iterative Rule Refinement

As in the geographic KBS, the first misclassified case was identified and examined in greater detail. The attributes of this case were compared to the most recently added cornerstone rule, with the goal of determining which variations in attribute values led to classification errors. For instance, a rule based solely on median household income might fail to account for family size, resulting in a misclassification. This prompted the creation of an additional refinement rule that incorporated both attributes to improve classification accuracy.

The iterative refinement process followed the structured approach outlined in Chapter 4 but required adjustments for the socio-demographic context. Unlike geographic rules, which relied on fixed discrete attributes such as postcodes and street names, socio-demographic rules needed to accommodate continuous and categorical variables. These included attributes like income brackets, education levels, and employment status, which often required range-based conditions to ensure accurate classification. The greater variability and overlap in socio-demographic factors necessitated a more flexible rule development approach compared to the more rigid structure of geographic-based rules.

5.3.2 Enhancing Rule Coverage

To maximize the effectiveness of each rule, data interrogation techniques were employed prior to rule construction. This included filtering the data such as isolating cases with education levels above a bachelor's degree or income brackets exceeding a specific threshold. Applying formulae such as calculating the ratio of household income to household size to identify socio-demographic trends influencing diabetes risk and using conditional formatting to visualize attribute patterns. This process involved several key steps:

- **Filtering techniques:** Data was filtered to isolate cases meeting specific socio-demographic thresholds. For example, records were filtered by income level (e.g., selecting cases where the median household income was below a predefined poverty threshold) or by education level (e.g., identifying individuals with education levels below a high school diploma) to observe correlations with diabetes prevalence.

- **Formulae and computed attributes:** Derived attributes were calculated to uncover hidden patterns in the data. For instance, the ratio of household income to family size was computed to determine financial strain per household. This formula allowed researchers to distinguish between low-income households with fewer dependents (lower risk) and those with many dependents (higher risk).

- **Conditional formatting:** Color-coding techniques were applied in Excel-based analysis to quickly visualize trends. For example, conditional formatting highlighted cases where low education levels (e.g., primary school or below) and large household sizes (e.g., five or more individuals) were both present, correlating with a higher diabetes risk. This enabled researchers to instantly spot patterns across large datasets.

These insights directly informed rule development, ensuring that new rules were broadly applicable yet precise enough to maintain high classification accuracy. For instance, if filtering revealed that a specific income bracket and family size were consistently linked to diabetes, a rule was developed using those attributes as key conditions. This targeted approach optimised rule effectiveness, reducing redundancy and improving accuracy.

In total, the rules in this experiment utilised between one and eight attributes per case, mirroring the structure of the geographic KBS. However, the socio-demographic dataset required more

granular analysis due to the variability introduced by aggregated socio-economic data. This iterative approach continued until the system achieved the target accuracy of 90.2%.

5.3.3 Iterative Refinement and Final Accuracy

As demonstrated in Figure 4.1, the geographic KBS achieved approximately 80% accuracy after the incorporation of around 45 rules, with additional rules contributing only incremental gains beyond this point. A similar trend was observed during the iterative development of the socio-demographic KBS. However, unlike the geographic KBS, which reached a final accuracy of 90.2% with 79 rules, the socio-demographic KBS required a total of 90 rules to achieve the same target accuracy.

This increase in the number of rules reflects the greater complexity of socio-demographic data, which often involves aggregated or overlapping attributes such as income ranges or education levels. These nuances necessitated the creation of more rules to capture subtle patterns and variations within the dataset.

Despite the added complexity, the socio-demographic KBS demonstrated improved generalization compared to the geographic KBS. This is evidenced by:

- **Broader applicability**: While geographic-based rules were often region-specific (e.g., a particular postcode linked to a diabetes risk factor), socio-demographic rules leveraged universal factors (e.g., income levels, education, and employment status), making them more transferable across different regions.
- **More adaptable decision boundaries**: The socio-demographic rules were less constrained by locality, allowing the system to generalize better when applied to populations beyond Albury-Wodonga. In contrast, geographic rules heavily relied on location-specific identifiers, limiting their usefulness outside the original dataset.
- **Performance across unseen data**: The results from applying both KBS models to their respective production datasets further support this generalization. While both models saw a drop in performance on unseen cases, the socio-demographic KBS exhibited stronger adaptability due to its reliance on socially relevant attributes rather than strict geographic identifiers.

In short, these results support the aim of a more generalisable RDR-based KBS for Type 2 diabetes in similar settings.

The rule construction table (Appendix B) summarises the attributes used, case coverage, and accuracy progression for each rule iteration. For instance, Rules 40 through 60 focused on refining classifications based on household income and education levels, which significantly contributed to the overall accuracy. Beyond Rule 90, additional refinements were implemented primarily to address edge cases and improve generalizability, ensuring the socio-demographic KBS remained robust across varied datasets.

Figure 5.3 illustrates the accuracy progression over iterations, showing steady improvements across the 133 rules required to reach 100% accuracy on the training dataset. Notably, the plateau effect, where additional rules yielded diminishing accuracy improvements, occurred much earlier in the socio-demographic KBS compared to the geographic KBS, reinforcing its ability to generalize patterns more efficiently. The decision to extend rule construction beyond the 79 rules used in the geographic KBS was driven by the need to account for the broader variability inherent in socio-demographic data. Ultimately, the socio-demographic KBS demonstrated its adaptability, achieving 100% accuracy on the training dataset while offering a more scalable and transferable predictive model than its geographic counterpart.

5.3.4 Iterative Refinement and Final Accuracy

The iterative refinement process for the socio-demographic KBS closely mirrored the methodology employed for the geographic KBS, as described in Chapter 4. However, while the geographic KBS achieved 90.2% accuracy with 79 rules, the socio-demographic KBS required 90 rules to reach the same accuracy level. As noted in section 5.2.1, socio-demographic fields required finer-grained thresholds and exceptions, which explains the higher rule count.

As shown in Figure 5.3, the accuracy progression followed a predictable pattern, with a rapid increase in early rule iterations, followed by a gradual leveling off as accuracy gains became incremental. However, unlike the geographic KBS, where 80% accuracy was achieved after 45 rules, the socio-demographic KBS reached a comparable threshold earlier, demonstrating a stronger generalization ability at earlier stages of rule development. The final refinement phase beyond 79 rules primarily focused on edge cases and ensuring robustness in classification, particularly in cases where socio-demographic factors overlapped in complex ways.

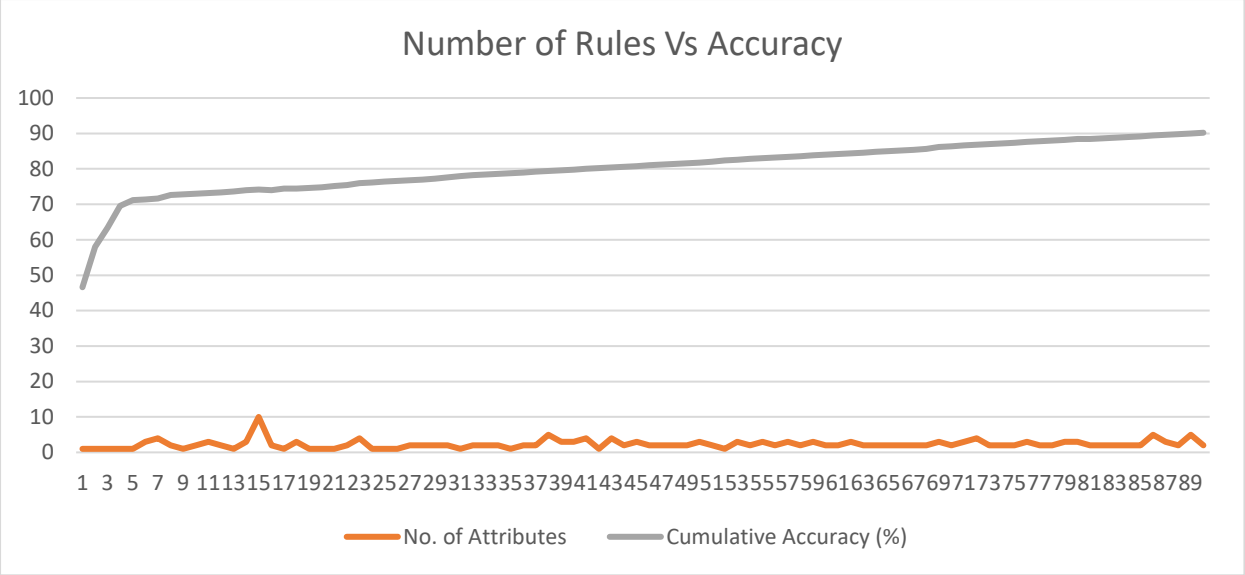

Figure 5. 3 The Excel rule builder mismatch cases output, showing cases that do not conform to any rules in the KBS.

Figure 5.3 illustrates that while early rules provided significant gains in accuracy, later rules primarily served to refine edge cases and improve generalizability. The first 50 rules accounted for approximately 85% of the total accuracy gains, whereas the last 40 rules contributed smaller, incremental improvements, refining the model's ability to classify borderline cases. By the time the 90th rule was added, accuracy had reached 90.2%, showing diminishing returns in further rule additions.

Critically, the production dataset results serve as an indicator of generalizability. While the geographic KBS (Chapter 4) experienced a 5.7% drop when applied to the production dataset (from 90.2% to 84.5% accuracy), the socio-demographic KBS showed a smaller decline, with accuracy only decreasing by 4.4% (from 90.2% to 85.8%). This suggests that socio-demographic factors provide a more generalised model, capable of maintaining predictive accuracy across different datasets. The iterative process was guided by real-time performance metrics, allowing for continuous monitoring and adjustment. By leveraging these metrics, the socio-demographic KBS demonstrated both adaptability and precision, effectively addressing the nuanced relationships within the dataset. These metrics not only provided immediate feedback on system performance but also informed key decisions regarding rule construction and refinement. Their role in the iterative process underscores the critical importance of real-time feedback, which is explored in greater detail in the following section.

5.3.5 Real-Time Feedback and Performance Metrics

The real-time feedback system played a critical role in the iterative refinement of the socio-demographic KBS. This system enabled the dynamic assessment of rule performance, allowing the identification of patterns, addressing misclassifications, and optimising rule coverage with greater efficiency. Their impact became increasingly significant as the KBS approached higher accuracy levels. The system provided essential performance indicators, which guided key decision-making in the knowledge acquisition process:

- Accuracy Rate:
  Tracked the system's ability to correctly classify cases as *diabetic* or *not diabetic*. This metric was recalculated after every rule iteration, ensuring continuous monitoring of system improvements.
- Execution Speed:
  Measured the processing time required to apply rules across the dataset, maintaining system efficiency as complexity increased.
- Dataset Coverage:
  Identified the proportion of cases successfully classified by the existing rules, highlighting areas requiring further refinement or additional rule development.

Using these real-time insights, it was possible to systematically refine rules, ensuring each contributed meaningfully to the KBS's overall performance. This dynamic approach reduced unnecessary rule proliferation, maintaining the balance between specificity and generalizability in classification outcomes. The real-time metrics were instrumental in guiding iterative refinements, allowing the researcher to pinpoint recurring misclassifications and areas requiring additional rule development. By continuously tracking accuracy trends, the system evolved to accommodate previously unclassified cases, ensuring broader dataset coverage. Beyond its role in optimising accuracy, the feedback system also provided empirical validation of the socio-demographic KBS's adaptability during its actual development. The ability to track, analyse, and respond to these variations in real time significantly enhanced the interpretability and applicability of the system.

By the end of the development process, the socio-demographic KBS not only matched the accuracy of the geographic KBS but also demonstrated its superior capacity to model complex, multidimensional health determinants. These findings underscore the practical advantages of

integrating socio-demographic insights into predictive modelling, paving the way for scalable, adaptable KBS solutions applicable beyond region-specific datasets. Most importantly, unlike the geographic KBS, which was bound by location-specific attributes, the socio-demographic KBS demonstrated greater generalizability, capable of identifying risk patterns across diverse populations.

5.3.6: Performance for Socio-Demographic KBS

The socio-demographic KBS clearly produced a high level of accuracy on the testing data, but the training had to be extended. In other words, the generalisation over the production data increased markedly but the knowledge acquisition process needed to be extended. For starters, the system reached an accuracy of 90.2% on the training dataset to enable the comparison with the performance of the geographic KBS. When applied to the production dataset, the accuracy of this intermediate KB dropped significantly to 75.8%. A closer examination of the metrics reveals that while specificity remained relatively high at 98% for the training dataset and 88% for the production dataset, sensitivity declined sharply from 71% to 17%, as illustrated in Table 5.1. This decline in sensitivity suggests that the system struggled to identify diabetic cases as effectively when dealing with unseen data. One possible explanation is the inherent variability in socio-demographic data across different populations, where unseen cases may exhibit unique socio-demographic profiles that were not sufficiently represented in the training dataset. By contrast to the geographic KBS, where errors clustered around postcode boundary effects and local spatial idiosyncrasies—the socio-demographic model is sensitive to population-composition shifts and the aggregation/granularity of attributes (e.g., income bands, education levels). Additionally, while the training phase ensured strong pattern recognition within the available dataset, it is possible that some rules were overly tailored to the specific distribution of attributes in the training data, leading to reduced generalizability. The high specificity, on the other hand, indicates that the system was still highly effective in correctly identifying non-diabetic cases. This trade-off highlights a key challenge in working with socio-demographic data; while these attributes provide valuable insights into diabetes risk factors, they may not fully encapsulate all medical and genetic contributors to the disease. Future iterations of the model could address this limitation by incorporating additional health indicators or refining rule generalization techniques to improve sensitivity without compromising specificity.

	Accuracy	Specificity	Sensitivity
Training Dataset	90.2%	98%	71%
Production Dataset	75.8%	88%	17%

Table 5. 2 The Excel rule builder mismatch cases output, showing cases that do not conform to any rules in the KBS.

These results underscore the challenges of generalising the socio-demographic KBS to unseen data. While the system achieved a strong accuracy of 90.2% on the training dataset, its ability to classify new cases in the production dataset was lower. This discrepancy suggests that the system effectively learned patterns specific to the training data but struggled to generalize those patterns to unseen cases. However, when the socio-demographic KBS was further refined to achieve 100% accuracy on the training dataset, the performance on the production dataset showed notable improvements. This suggests that a more comprehensive rule base led to better generalization, counteracting some of the limitations observed at the 90.2% accuracy level. Section 5.4 further explores these improvements, demonstrating how the refined KBS produced a stronger predictive model with enhanced generalizability.

5.3.7 Factors Contributing to Performance Variability

See section 5.3.6 for training Vs production results; this subsection outlines factors that may explain the observed discrepancies.

Several factors may explain the observed discrepancies between the training and production datasets:

1. Data Aggregation and Generalization
   Socio-demographic attributes often represent aggregated or range-based data, such as median household income or average education levels. Unlike geographic identifiers like street names and postcodes, which offer more discrete categorizations, socio-demographic attributes introduce a layer of generalization that can impact rule precision. For instance, two areas with similar median income levels but differing diabetes prevalence may lead to inconsistencies in classification.

2. **Dataset Variability**
   The production dataset may encompass demographic contexts, socio-economic distributions, or health determinants that were underrepresented in the training dataset. Although rule-based models such as RDR exhibit strong adaptability, discrepancies arise when previously unseen socio-demographic patterns emerge. This challenge is not unique to rule-based systems as other predictive models also struggle when dealing with out-of-distribution data, particularly when inputs are aggregated rather than case-specific.

3. **Sensitivity to Minority Patterns**
   The sharp decline in sensitivity on the production dataset suggests that the socio-demographic KBS struggled with detecting diabetic cases, particularly those representing edge cases or minority patterns. This may reflect an imbalance in case distribution within the training dataset, which resulted in rules favouring non-diabetic classifications. As a result, certain diabetic cases in the production dataset were not well-represented by the training rules.

4. **Overfitting During Rule Construction**
   While the iterative rule construction process effectively achieved high training accuracy, it also led to overfitting. The KBS became finely tuned to the patterns in the training dataset but lacked the flexibility to generalize effectively to new cases. This is a well-documented limitation in rule-based systems, where increasing the number of highly specific rules can reduce adaptability to unseen data.

5. **Challenges of Socio-Demographic Factors**
   Unlike geographic attributes, socio-demographic factors are inherently dynamic and influenced by temporal and regional changes. Attributes such as median household income or education levels may shift significantly over time or vary between regions. This variability can impact the stability of predictive models, making it necessary to refine rule construction approaches to improve generalizability.

These factors highlight the complexities of applying socio-demographic data to predictive modelling. However, further refinements to the KBS, explored in the next section, demonstrate improvements in generalizability, particularly as accuracy on the training dataset is extended to 100%.

5.3.8 Implications and Recommendations

These findings highlight critical areas for improvement in the socio-demographic KBS, particularly in addressing overfitting and enhancing generalisability:

- **Expanding Training Data:**
  The observed performance gap between the training and production datasets suggests that expanding the dataset to include a more diverse and representative sample could enhance generalization. Integrating national or regional socio-demographic data alongside the local dataset could help capture a broader range of socio-economic patterns, reducing overfitting and improving prediction accuracy for unseen cases.

- **Dynamic Rule Refinement:**
  While the RDR framework enables incremental rule updates, introducing a mechanism for adaptive learning, where the system continually incorporates emerging socio-demographic trends could improve its responsiveness to changing social conditions. This would allow the KBS to dynamically adjust its rules based on evolving health disparities.

- **Complementary Attributes:**
  Enhancing the predictive power of the system may require integrating additional contextual factors, such as healthcare access, lifestyle indicators, and environmental influences. These attributes, when combined with socio-demographic determinants, may improve both sensitivity and specificity, particularly in cases where aggregated socio-economic data alone is insufficient for accurate classification.

The socio-demographic KBS has demonstrated strong potential, achieving high accuracy and specificity on the training dataset. However, the performance gap on the production dataset underscores the inherent challenges of working with socio-demographic data. Section 5.4 explores whether achieving 100% accuracy on the training dataset mitigates these challenges and improves real-world applicability. Addressing these challenges through expanded datasets, dynamic rule refinement, and complementary attributes will be critical for optimising the system's robustness and broad applicability.

5.4: Tailored Rule Development (100% accuracy on the training dataset)

As the socio-demographic KBS was further developed to achieve 100% accuracy on the training dataset, the complexity of rule construction increased significantly. While early rules primarily addressed broad patterns, later rules required more refined attribute combinations to handle

edge cases that the KBS initially failed to classify correctly. This involved not only filtering data but also examining interactions between multiple socio-demographic attributes, particularly in borderline cases.

Overall, achieving 100% accuracy required 133 rules, compared to 90 rules for 90.2% accuracy. The additional 43 rules were primarily designed to refine classifications for rare or borderline cases, ensuring complete accuracy across the training dataset.

5.4.1 Iterative Refinement and Final Accuracy

As demonstrated in Figure 4.1, the geographic KBS reached 80% accuracy after approximately 45 rules, with additional rules yielding only incremental improvements. The socio-demographic KBS, however, required 90 rules to reach 90.2% accuracy on the training dataset. Extending the system to 100% accuracy demanded a total of 133 rules, highlighting the greater complexity and variability of socio-demographic data compared to its geographic counterpart.

This increase in rule count underscores the need for a more nuanced approach to account for the aggregated nature of socio-demographic attributes such as income ranges, household sizes, and education levels. Unlike location-based attributes, which are often binary or discrete, socio-demographic factors overlap and interact in intricate ways, requiring more refined classifications. As accuracy improved, each additional rule played a crucial role in addressing rare or borderline cases, ensuring that previously misclassified instances were correctly categorised.

The final stages of rule construction focused on edge cases, where rules often involved multiple attributes simultaneously. The iterative refinement process remained consistent, involving misclassification analysis, targeted rule adjustments, and the integration of new socio-demographic conditions to further improve classification performance.

Appendix C summarises the full set of 133 rules and the incremental accuracy gains as rules were added, highlighting how socio-demographic attributes contributed to decisions. Rules 40–90 refined classifications using income, education and household size, while Rules 91– 133 resolved edge cases and overlapping patterns, bringing the training set to 100% accuracy. Beyond Rule 90, the remaining 43 rules were designed primarily to resolve edge cases, addressing overlapping socio-demographic patterns and subtle distinctions within the dataset. These refinements ensured that previously misclassified cases were correctly categorised, thereby meeting the 100% accuracy benchmark.

Figure 5.4 illustrates the accuracy progression throughout the rule development process. The initial stages of rule construction resulted in sharp accuracy gains, followed by diminishing returns as the system approached 133 rules. This pattern aligns with typical iterative knowledge base development, where early rules establish broad classifications, and later refinements fine-tune decision boundaries.

Extending rule development beyond 90 rules (90.2% accuracy) to 133 rules (100% accuracy) was necessitated by the increased complexity of socio-demographic data. Addressing these challenges required incremental rule refinement, demonstrating the adaptability and effectiveness of the RDR methodology in achieving a highly precise classification system.

5.4.2 Iterative Refinement and Final Accuracy

The iterative refinement process for the socio-demographic KBS followed a similar methodology to the geographic KBS, as described in Chapter 4. However, key differences emerged, particularly in how quickly the cumulative accuracy plateaued during rule construction.

Unlike the geographic KBS, where steady accuracy improvements continued until approximately 45 rules, the socio-demographic KBS experienced a sharp accuracy gain early on, plateauing around 6 rules (Figure 5.4). This suggests that socio-demographic attributes provided broader coverage per rule, allowing the system to capture meaningful patterns more efficiently in the early stages.

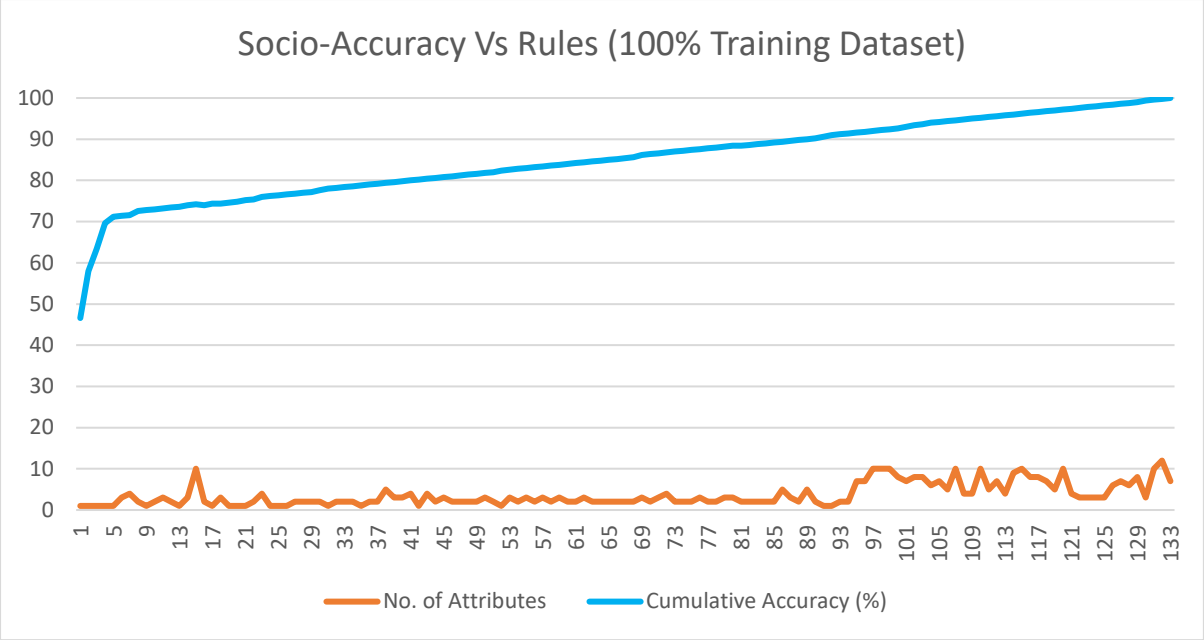

Figure 5. 4 The Excel rule builder mismatch cases output, showing cases that do not conform to any rules in the KBS.

Despite this early plateau, achieving 100% accuracy required 133 rules in total, reflecting the complexity and overlapping nature of socio-demographic attributes such as income ranges, education levels, and family size. These additional refinements were necessary to address nuanced cases and resolve edge classifications, ensuring that all cases in the training dataset were correctly categorised.

The accuracy progression observed in Figure 5.4 reveals a rapid rise in cumulative accuracy during the early rule iterations, with substantial gains achieved by Rule 10. The curve begins to plateau after a dozen rules or so, indicating that socio-demographic attributes quickly establish a strong classification foundation. However, subsequent rule additions contributed only incremental improvements, reflecting the challenges of refining exception cases and addressing overlapping socio-demographic patterns. This behaviour highlights key differences between geographic and socio-demographic data within the KBS framework. The geographic KBS reached 90.2% accuracy with 79 rules, whereas the socio-demographic KBS required 90 rules to achieve the same accuracy (section 4.4; section 5.3) . In contrast, socio-demographic attributes, being aggregated or continuous variables (e.g., income ranges, education levels), provided broader initial coverage but demanded more fine-tuned refinements to capture nuanced distinctions within the dataset.

The graph also illustrates the number of attributes per rule, represented by the orange line in Figure 5.4. Most rules incorporated between one and three attributes, but later stages exhibited notable spikes in complexity. These spikes reflect the effort required to refine edge-case exception, ensuring that patterns were correctly recognised across diverse socio-demographic profiles. This aligns with earlier findings that socio-demographic data necessitates greater granularity due to its inherent variability.

Ultimately, the socio-demographic KBS achieved 100% accuracy on the training dataset, an outcome driven by iterative refinements guided by real-time performance metrics. The transition from a rapid accuracy climb to a plateau underscores the importance of efficient rule development in knowledge-based systems, particularly when handling complex datasets. The earlier plateau observed in the socio-demographic KBS reinforces the initial effectiveness of socio-demographic attributes in capturing broad classification patterns but also highlights the increased effort required for fine-tuning edge cases as full accuracy is approached. These findings provide valuable insights into the distinct advantages and challenges posed by socio-demographic data in predictive model construction. Indeed, the generalisation achieved with the socio-economic attributes has immediate impact on the efficiency of the knowledge acquisition process. The added effort in the preprocessing to acquire these attributes is outweighed by the efficiency of the KA and more importantly with the generalisation of the KB (leading to higher accuracy on unseen data).

5.4.3 Real-Time Feedback and Performance Metrics

The iterative development of the socio-demographic KBS was supported by a robust real-time feedback system, which provided immediate insights into the performance of each rule iteration. These dynamic metrics were instrumental in guiding the refinement process, ensuring the system remained both adaptable and precise. By leveraging real-time data, the researcher could make informed decisions regarding rule refinement, addition, or termination based on observed diminishing returns.

The feedback system dynamically displayed several key performance metrics critical to the development process:

- Accuracy Rate:
  This metric measured the system's ability to classify cases correctly as diabetic or not diabetic. Accuracy was recalculated after each rule iteration, offering a clear and

consistent measure of progress. As the system was refined to reach 100% accuracy on the training dataset, this metric played a pivotal role in highlighting the remaining misclassified cases and evaluating the impact of each new rule.

- **Execution Speed:**
  The time required to apply the rules across the entire dataset was monitored to ensure the system remained efficient despite the growing number of rules. Even with 133 rules, the execution speed remained within acceptable limits, demonstrating that the KBS framework remained scalable and practical when applied to the socio-demographic dataset.

- **Dataset Coverage:**

  This metric identified the proportion of cases addressed by the rules at each stage of development. As additional rules were introduced, dataset coverage expanded, focusing particularly on cases that remained unclassified. These uncovered cases often informed the creation of new rules, ensuring that the final KBS comprehensively addressed the variability within the dataset.

As additional rules were introduced, dataset coverage expanded, focusing particularly on cases that remained unclassified:

- **Accuracy trends** were displayed graphically, allowing the researcher to track improvements over 133 rule iterations and observe diminishing returns in later stages.
- **Uncovered cases** were highlighted in pink, as shown in Figure 5.2, drawing attention to requiring intervention. This enabled a targeted refinement approach, ensuring rules were adapted effectively.

The real-time feedback loop provided invaluable support throughout the process. Misclassified cases were quickly identified, prompting refinements to existing rules or the addition of new ones tailored to the specific attributes of those cases. The system also highlighted the diminishing returns associated with later rules, allowing the researcher to prioritize rules that had the greatest impact on improving the model's generalizability.

Ultimately the iterative refinement process enabled the socio-demographic KBS to achieve 100% accuracy on the training dataset. While this milestone demonstrated the robustness of the approach, the insights provided by real-time feedback emphasised the system's broader applicability in handling complex socio-demographic determinants. By the conclusion of the

development process, the socio-demographic KBS not only matched the geographic KBS in its predictive capabilities but surpassed it in versatility. Its ability to generalize across diverse demographic contexts while maintaining computational efficiency underscores the value of socio-demographic data in predictive modelling.

5.4.4: Performance for Socio-Demographic KBS

The socio-demographic KBS exhibited varying performance levels when evaluated on the training and production datasets. On the training dataset, the system achieved a perfect accuracy of 100%, demonstrating the successful integration and refinement of 133 rules to classify every case correctly. However, when tested on the production dataset (unseen data), accuracy declined to 76.7%.

A closer examination of the metrics reveals that while specificity remained relatively robust at 100% for the training dataset and 88% for the production dataset, sensitivity experienced a significant decline from 100% on the training dataset to only 22% on the production dataset, as illustrated in Table 5.2.

	Accuracy	Specificity	Sensitivity
Training Dataset	100%	100%	100%
Production Dataset	76.7%	88%	22%

Table 5. 3 Results obtained for the training and production datasets using the socio-determinant enhanced dataset experiment. (100% accuracy on the training dataset)

These results underscore the complexities of generalising a socio-demographic KBS to unseen data. While the system achieved 100% accuracy on the training dataset, its lower performance on the production dataset (76.7% accuracy) highlights the challenges of applying socio-demographic patterns in broader, more diverse contexts. Rather than a conventional case of overfitting, where a model memorises training data and loses adaptability, the reduced performance may be attributed to the inherently regional and population-specific nature of socio-demographic data.

Socio-demographic attributes, such as median income, education levels, and household size, are typically aggregated and can vary significantly between regions. These disparities may not always be adequately captured in the production dataset, leading to mismatches between the learned patterns and new, unseen data. For instance, individuals in the production dataset may come

from suburbs with socio-demographic profiles absent from the training dataset, impacting the system's ability to classify cases with high sensitivity.

Additionally, the dataset's class imbalance (30% diabetic, 70% non-diabetic) likely biased the KBS towards optimising specificity, making it particularly effective at identifying non-diabetic cases (88% specificity) but less reliable in detecting true positive diabetic cases (22% sensitivity). The system's high specificity demonstrates that it correctly rules out non-diabetic cases but struggles to recognize positive diabetes cases, a limitation that can be addressed through data augmentation or weighted rule construction techniques. Rather than reflecting a fundamental limitation of socio-demographic predictors, these findings emphasize the importance of incorporating diverse and representative datasets to enhance the generalizability. Expanding the dataset to include a wider range of socio-demographic profiles, addressing class imbalance, and refining rule selection could improve sensitivity on unseen cases while retaining the strong specificity observed in both training and production datasets.

When compared to the 90.2% accuracy experiment, accuracy model further reduced sensitivity while showing a slight increase in specificity on unseen data. At 90.2% accuracy, the production dataset sensitivity was 39.3%, while specificity was 83.3%. In contrast, at 100% accuracy, sensitivity declined to 22%, while specificity improved to 88%. This pattern suggests that perfect accuracy on the training dataset strengthened precision but reduced the model's ability to generalize, reinforcing the trade-off between optimization on known data and adaptability to new cases.

These findings demonstrate the KBS's strong capacity to capture and encode socio-demographic patterns effectively but also highlight areas for improvement in generalizability. Future refinements, including balancing the dataset, incorporating additional socio-demographic features, and applying techniques to mitigate sensitivity reduction, could further enhance the system's ability to identify diabetic cases in unseen data. Nonetheless, the high specificity across both datasets remains a key strength, reinforcing the robustness of socio-demographic factors in predicting type 2 diabetes risk.

5.5 Evaluating the Socio-Demographic KBS on the Production Dataset

Building on the incremental knowledge acquisition approach outlined in previous sections, this section evaluates the performance of the socio-demographic KBS on the production (unseen) dataset. This step is crucial in validating the system's ability to generalize beyond the training dataset and accurately predict type 2 diabetes risk, based on socio-demographic attributes.

The geographic KBS, detailed in Chapter 4, demonstrated promising results when applied to the training dataset. However, its performance was somewhat constrained when tested on unseen data from the production dataset, primarily due to the inherent limitations of geographic attributes in providing a universally transferable predictive model. In contrast, the socio-demographic KBS was designed to overcome these limitations by leveraging broader, more adaptable social determinants such as household income, education levels, and family size. These attributes are not bound to specific locations and instead reflect systemic influences that affect diabetes risk across different populations.

5.5.1 Experimental Setup

The socio-demographic KBS was evaluated using the same structured approach detailed in Chapter 4. The rules developed incrementally from the training dataset were applied to the production dataset, which consisted of previously unseen cases. The primary goal was to assess the model's predictive accuracy, identify any misclassifications, and determine whether the rule set required additional refinements.

The evaluation focused on the following key metrics:

- **Accuracy:** the percentage of correctly classified cases.
- **False Positive/Negative Rates:** to assess model reliability and sensitivity.
- **Rule Adaptability:** how well the existing rules accommodate new cases without significant modification.

Results from these evaluations are summarised in the table in Appendix E, showing the accuracy rate for both the training and production (unseen) datasets as each rule is added. A graphical representation of these accuracy trends is provided in Figure 5.5.

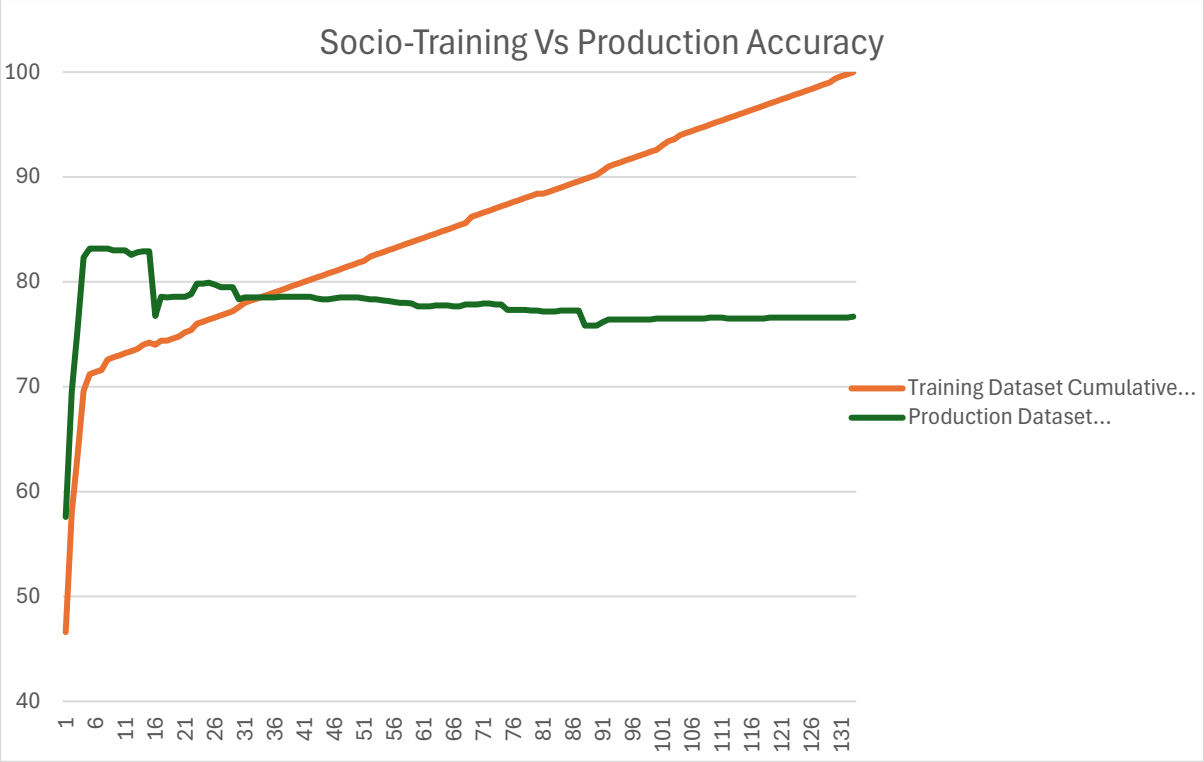

Figure 5. 5 Number of rules training Vs production (unseen) accuracy as each rule is added, including number of attributes used for each rule (100% accuracy on the training dataset)

5.5.2 Observations and Initial Findings

The results indicated that the socio-demographic KBS achieved higher accuracy than its geographic counterpart when tested on the production dataset. This comparison is discussed in detail in Chapter 6. The system successfully classified a larger proportion of cases correctly, demonstrating a superior ability to generalize across different populations. This can be attributed to the fact that socio-demographic attributes inherently encapsulate broader health determinants that are less region dependent than geographic indicators.

Additionally, the socio-demographic KBS required fewer modifications when applied to unseen data. While the geographic KBS faced challenges in adapting to new locations, since geographic attributes such as postcodes and street names are inherently region specific, the socio-demographic KBS exhibited greater transferability. This reinforces the hypothesis that social determinants offer a more robust foundation for predictive modelling in public health applications.

5.5.3 Preliminary Comparison with the Geographic KBS

While a detailed comparative discussion is presented in Chapter 6, it is important to note some key preliminary observations:

- **Higher Generalizability:** The socio-demographic KBS exhibited better adaptability when applied to the production dataset, whereas the geographic KBS required more rule modifications to maintain accuracy.
- **More Stable Performance:** The socio-demographic KBS produced fewer misclassifications, suggesting that attributes such as income, education, and household size provide stronger predictive signals than location-specific indicators.
- **Scalability:** Since socio-demographic factors are applicable across regions, this KBS model has the potential to be expanded and deployed in various geographical areas with minimal adjustments.

These initial findings highlight the advantages of incorporating socio-demographic data in predictive healthcare models. While geographic attributes can provide localised insights, they are inherently limited in scope and applicability. By contrast, socio-demographic attributes allow for more flexible and universally relevant rule development, making them a valuable tool for broader diabetes risk prediction initiatives.

5.5.4 Implications for Further Refinement

Given the encouraging results of the socio-demographic KBS, future iterations of the system will focus on further optimising rule efficiency and exploring additional social determinants that may enhance predictive accuracy. The integration of dynamic socio-economic trends, such as employment fluctuations and policy changes, could further improve the system's ability to anticipate shifts in diabetes risk within different populations.

The next chapter deeply explores the comparative analysis between the geographic and socio-demographic KBS models, further unpacking the observed performance differences and investigating their broader implications for KBS development in healthcare.

5.6 Summary

This chapter outlined the incremental development of the socio-demographic KBS, expanding on the methodologies established in Chapter 4. Unlike the geographic KBS, which relied on location-based attributes such as street names and postcodes, the socio-demographic KBS incorporated broader contextual factors, including median household income, education levels, and mortgage costs. This approach enhanced the system's adaptability, enabling it to capture a more holistic representation of type 2 diabetes risk factors that extend beyond geographical constraints.

The iterative refinement process, using RDRs, facilitated structured rule development, ensuring incremental adaptation to the complexities of socio-demographic data. Initially, the system achieved 90.2% accuracy on the training dataset with 90 rules. However, further refinement was undertaken to push the training dataset to 100% accuracy, requiring a total of 133 rules. This increase in rule count reflects the greater complexity and variability of socio-demographic attributes, which often involve aggregated or overlapping data. The results reinforce that while achieving full accuracy is possible, ensuring generalizability to unseen data remains a challenge.

The chapter began by detailing the experimental platform used to construct the socio-demographic KBS, highlighting its dynamic rule development and iterative refinements. The integration of real-time performance metrics was particularly valuable, as it allowed continuous assessment of accuracy, execution speed, and dataset coverage. Several challenges unique to socio-demographic data were addressed, including the handling of aggregated values, missing information, and the standardization of socio-economic variables.

Despite these complexities, the KBS demonstrated strong adaptability, initially achieving 90.2% accuracy with 90 rules, a slightly higher count than the geographic KBS (79 rules). This increase reflected the greater variability within socio-demographic datasets. The tool's intuitive design ensured usability for non-technical experts, reinforcing its potential for broader application in public health and policymaking.

The transition from training to production datasets revealed a significant trade-off between specificity and sensitivity. At 90.2% accuracy, specificity was 83.3%, but sensitivity fell to 39.3%, highlighting challenges in detecting true diabetic cases within unseen data. Upon achieving 100% accuracy on the training dataset, specificity improved slightly to 88%, but sensitivity dropped further to 22%. This underscores how overfitting to the training dataset impacted generalization,

requiring future improvements to optimize sensitivity while maintaining specificity. The results emphasize that socio-demographic factors remain strong predictors of diabetes risk but must be carefully balanced to ensure adaptability across different populations.

The incremental refinement process demonstrated the strengths of the RDR approach, which allowed the system to accommodate nuanced socio-demographic variations. However, the increase from 90 to 133 rules highlights the balance between expanding rule complexity and maintaining model generalizability. While 100% accuracy on the training dataset was achieved, the performance gap on the production dataset reinforces the importance of dataset diversity and adaptive rule structures in improving real-world applicability. These findings validate the effectiveness of expert-driven rule-based models in capturing complex health determinants. The next step involves assessing how this KBS framework compares to alternative approaches, particularly ML methods. Chapter 6 analyses the performance of both the geographic and socio-demographic KBSs in comparison to ML models, identifying the strengths, limitations, and practical implications of each approach in predicting type 2 diabetes risk.

Chapter 6: Geographic & Socio-Demographic Production and ML Comparisons

The chapter is divided into five sections, each focusing on a specific aspect of the experimental results. Section 6.1 provides an overview of the chapter and outlines the main goals and structure of the experiments conducted. Section 6.2 evaluates the performance of the socio-demographic KBS when applied to the production (unseen) dataset, comparing its results against the training dataset. This analysis highlights the model's generalizability and its advantages over the geographic KBS. Section 6.3 establishes a structured comparison between RDRs and ML in KBS development, examining their fundamental differences in methodology, adaptability, and scalability. Section 6.4 extends the theoretical discussion by presenting real-world results from the application of Weka's J48 decision tree algorithm to the same datasets used in RDR development. For the ML baseline WEKA's J48 decision tree was used (Witten et al. 2023; Witten et al. 2025). It highlights the empirical performance of ML-based KBS construction in comparison to RDR-based incremental knowledge acquisition. Section 6.5 synthesises the key takeaways from this comparative analysis, reinforcing the advantages of RDRs in developing adaptable, interpretable, and expert-driven decision support systems. The discussion also sets the stage for the next chapter by emphasising the broader implications of knowledge-driven AI in predictive healthcare.

6.1 Overview

While Chapters 4 and 5 detailed the incremental development and performance of both the geographic KBS and the socio-demographic KBS, these experiments raise an important question: how well does the socio-demographic KBS generalize when applied to previously unseen data?

Unlike geographic attributes, which are inherently region-specific, socio-demographic attributes provide a broader, more adaptable foundation for predictive modelling. To assess this, Section 6.2 provides an evaluation of the performance of the socio-demographic KBS on the production dataset, comparing its results against the training dataset. This analysis highlights the model's generalizability, identifying key advantages and potential areas for refinement (e.g., Hill-Briggs et al., 2021; WHO, 2024).

Following this evaluation, the chapter compares RDR and ML in two domains: (i) whether ML can learn spatial patterns from geographic data and generalise as effectively as an expert-driven RDR, and (ii) whether ML captures complex, non-linear socio-demographic interactions and generalises across populations without extensive retraining, or whether RDR's incremental, expert-driven approach is more adaptable.:

1. **Geographic RDR vs. ML:**
    - Can ML effectively learn spatial patterns from geographic data in a way that rivals the expert-driven rule construction in RDR?
    - How well does ML handle geographic attributes and does it generalize as effectively as an RDR model that incorporates domain expertise?
2. **Socio-Demographic RDR vs. ML:**
    - Does ML capture the complex, non-linear interactions within socio-demographic attributes as effectively as an RDR system?
    - Can an ML model generalize across different populations without requiring extensive retraining, or does the expert-driven nature of RDR provide a more adaptable approach?

Through this comparison, the chapter aims to highlight the unique advantages of RDR, particularly in scenarios where data availability is limited, interpretability is crucial, and continuous updates are required. Unlike ML models, which remain static once trained, RDR-based KBS development is incremental, allowing subject-matter experts (SMEs) to inject new knowledge and refine rules dynamically.

The following sections consider data requirements and scalability, interpretability and transparency, and adaptability and update mechanisms. By analysing these factors, this chapter establishes the case for RDR as a superior alternative to ML for developing KBSs in healthcare, particularly in environments where expert knowledge plays a crucial role in decision making.

6.2 Performance Comparison: Socio-Demographic KBS on Training vs Production Data

This analysis operationalises the socio-ecological model of diabetes risk (Hill, Nielsen & Fox, 2013) by mapping each rule/metric to its corresponding social determinant.

This section evaluates the performance of the socio-demographic KBS when applied to both the training dataset and the production (unseen) dataset. The goal is to assess how well the rules developed during training generalize to unseen cases, and to determine whether any refinements are necessary for optimising predictive accuracy.

6.2.1 Accuracy Trends Across Datasets

The socio-demographic KBS was designed to incrementally refine its rule base using training data before being tested on previously unseen cases from the production dataset. As expected, the performance on the training dataset showed a steady increase in accuracy as rules were added. However, the critical test of the system's robustness lies in its application to new data.

A comparison of accuracy rates for both datasets is provided in the table in Appendix F, illustrating how accuracy evolved as more rules were incorporated. Additionally, Figure 6.1 presents a visual representation of these trends, highlighting differences between the geographic and the socio-demographic production datasets.

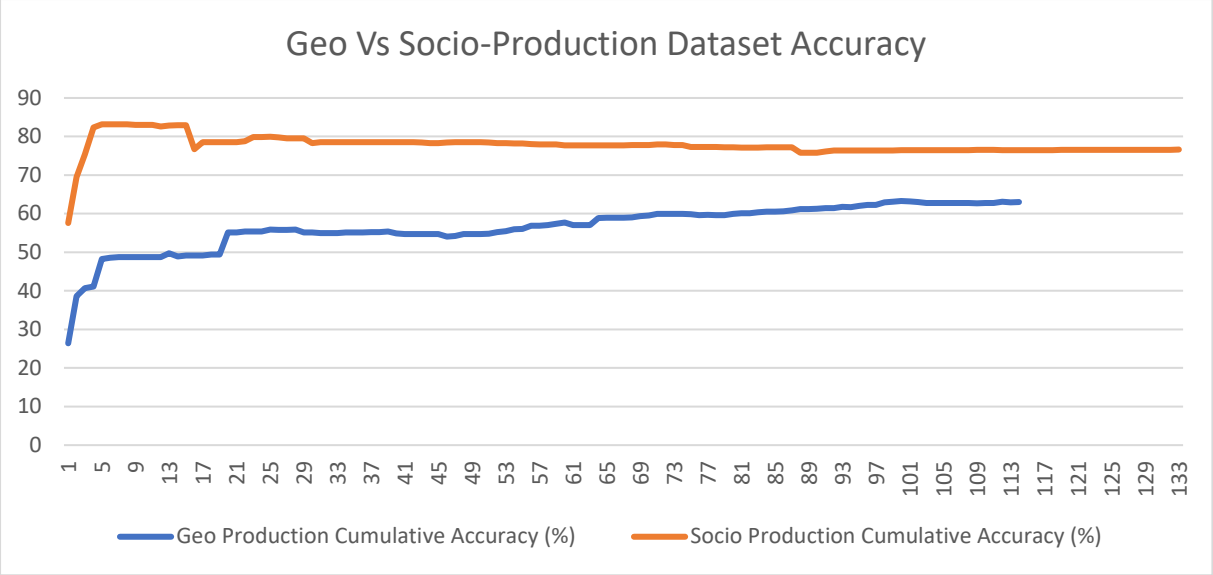

Figure 6. 1 Number of rules training Vs production (unseen) accuracy as each rule is added, including number of attributes used for each rule (100% accuracy on the training dataset)

Key observations:

- The training dataset reached high accuracy, demonstrating that the rules effectively captured patterns within the known data.
- The production dataset achieved comparable accuracy, reinforcing the generalizability of socio-demographic attributes.
- Minor performance deviations were observed in the production dataset, but these were significantly lower compared to the geographic KBS results in Chapter 4.
- As shown in Figure 6.1, the average accuracy on the socio-demographic production dataset was significantly higher than that on the geographic dataset, with the socio-demographic KBS averaging around 80%, compared to approximately 60% for the geographic KBS.

6.2.2 Key Findings and Refinements

The results indicate that the socio-demographic KBS outperforms the geographic KBS in terms of transferability and predictive accuracy when applied to unseen data. Unlike geographic attributes, which often require location specific adaptations, socio-demographic factors offer a more universal predictive foundation, leading to minimal performance drop-offs when transitioning from training to production datasets.

Further refinements may still be needed to optimize rule efficiency. Future work could focus on:

- fine tuning rules with lower confidence scores to improve generalizability.
- exploring additional socio-demographic attributes to enhance predictive power.
- examining error cases more closely to identify trends among misclassified instances.

6.2.3 Broader Implications for Knowledge-Based Systems

These findings reinforce the broader value of socio-demographic data in predictive modelling. Since such attributes are not constrained by geography, they provide an adaptable framework for diabetes risk assessment that can be applied across diverse populations with minimal recalibration. This adaptability is crucial for developing scalable, real-world decision support systems in public health.

The next section incorporates these insights into a broader comparative discussion, further analysing how the geographic and socio-demographic KBS models differ in practical application, rule adaptability, and long-term viability.

6.3 Comparative Framework: Ripple-Down Rules (RDR) vs. Machine Learning (ML)

This analysis operationalises the socio-ecological model of diabetes risk (Hill, Nielsen & Fox, 2013) by mapping each rule/metric to its corresponding social determinant.

This section establishes a structured comparative framework for evaluating RDR and ML in the development of KBS. While both approaches aim to extract meaningful patterns from data, their underlying methodologies, adaptability, and practical applications differ significantly.

6.3.1 Methodological Differences

At their core, RDR and ML employ fundamentally different paradigms in constructing predictive models:

- Ripple-Down Rules:
    - operate through incremental rule-building, where experts iteratively refine the system by adding human understandable rules.
    - require no predefined dataset size and thrive in data sparse environments, making it particularly effective in domains where data is scarce or expensive to collect.
    - are dynamic and continuously evolving, allowing the system to incorporate new knowledge without full retraining.
- Machine Learning:
    - relies on large scale datasets, using statistical learning techniques to identify patterns and relationships between input variables.
    - once trained, ML models are static, meaning new knowledge cannot be integrated without retraining the entire model.
    - often operates as a black-box system, where interpretability is sacrificed in favour of higher predictive performance.

Table 6.1 summarises the variations in features between using RDRs and ML to develop a KBS.

Feature	Ripple-Down Rules	Machine Learning
Development Approach	Incremental, expert-driven rule creation	Requires large-scale training on datasets
Adaptability	Continuously updated without retraining. As demonstrated in Section 6.2, the socio-demographic KBS exhibited superior generalization to unseen data, reinforcing its adaptability over geographic KBS models	Static model; requires full retraining for updates
Interpretability	Fully transparent, rule-based system	Often a black box, making decision rationale unclear
Data Requirements	Minimal; thrives in low-data environments	High; dependent on large, labelled datasets
Computational Complexity	Low; efficient with incremental updates	High; training ML models is resource-intensive

Table 6. 1 Comparing the various features of RDRs and ML in the development of a KBS.

The fundamental contrast between RDR's expert-driven, knowledge enhancing framework and ML's data-dependent, static model development serves as the foundation for a deeper comparative analysis in geographic and socio-demographic KBS contexts.

1. Human Interpretability & Trust
   - RDR provides transparent, interpretable decision logic, which is crucial for clinician trust in decision support systems.
   - ML models, particularly neural networks and deep learning models, often produce predictions without explainability, raising concerns about clinical accountability and regulatory compliance.
2. Data Availability & Domain Expertise
   - RDR is ideal for environments where data is scarce, as expert knowledge compensates for missing data.
   - ML requires extensive datasets, which may not always be available, particularly in specialised medical domains.
3. Scalability & Generalizability

- ML models often struggle with domain transferability, a model trained on one population may perform poorly on another due to dataset bias.
- RDR, with its incremental adaptability, ensures that new cases are integrated without losing previously learned knowledge, making it better suited for evolving medical insights.

4. Computational & Resource Constraints
   - Training ML models requires significant computational resources and expertise in hyperparameter tuning, feature engineering, and optimization.
   - RDR, on the other hand, requires minimal computation, making it an efficient alternative in real-world healthcare settings where processing power may be limited.

6.4 Comparison of Machine Learning and Ripple-Down Rules for Knowledge Base Development

This analysis operationalises the socio-ecological model of diabetes risk (Hill, Nielsen & Fox, 2013) by mapping each rule/metric to its corresponding social determinant.

The construction of a KBS requires a robust approach to pattern recognition, rule derivation, and adaptability. Traditional ML methods have been widely used for predictive modelling in healthcare, including type 2 diabetes risk assessment. However, RDR offer a fundamentally different methodology for KBS development, particularly in environments where domain expertise, interpretability, and incremental rule refinement are crucial.

This section provides a structured comparison between ML-based KBS construction and RDR-based incremental knowledge acquisition. The focus is on their respective advantages and limitations in handling medical and socio-demographic data, emphasising the adaptability of RDRs in dynamically evolving domains. In addition to this theoretical comparison, an empirical evaluation is presented, where Weka's J48 decision tree algorithm is applied to the same datasets used for RDR development (Witten et al. 2025; WEKA Team n.d.). This real-world comparison further illustrates the practical challenges and trade-offs associated with ML-based approaches, reinforcing the advantages of RDRs in expert-driven KBSs.

6.4.1 Data Requirements and Model Training

One of the most significant differences between ML-based and RDR-based KBS development is their approach to data requirements and model training.

- **Machine Learning Dependency on Large Datasets:**
  ML models require vast amounts of data to establish predictive accuracy. In many medical and socio-demographic contexts, such large datasets are not always available or may be incomplete due to privacy concerns, ethical considerations, and data collection inconsistencies (Obermeyer et al., 2019). Without substantial amounts of high-quality training data, ML models risk producing unreliable or biased predictions.
- **RDR's Ability to Function with Limited Data:**
  Unlike ML, RDR does not require extensive training datasets. Instead, it allows for incremental rule development, where a subject matter expert (SME) can iteratively refine the system based on misclassifications and observed cases. This adaptability makes RDR particularly suitable for domains where data is scarce or where expert-driven knowledge can supplement computational learning.

6.4.2 Interpretability and Explainability

Another fundamental distinction between ML and RDR is in their explainability, a critical factor in medical decision making.

- **The Black Box Nature of ML:**
  Many ML models, especially deep learning and neural networks, are often criticised for being black boxes (Lipton, 2018). While they can achieve high predictive accuracy, their internal decision-making processes are difficult to interpret. This lack of transparency can be problematic in healthcare settings, where decision justification is crucial.
- **RDR's Transparency and Justification:**
  RDR naturally provides explainable reasoning by constructing a rule hierarchy that mirrors expert thinking. Each rule is developed based on observable misclassifications, ensuring that decisions remain interpretable. This feature allows users to trace the reasoning behind each classification, a vital aspect of medical diagnostics and decision support.

6.4.3 Adaptability and Incremental Learning

Both ML and RDR can adapt to new information, but their mechanisms for doing so, differ significantly.

- **ML Requires Retraining from Scratch:**
  Once an ML model is trained, any new knowledge or data updates require retraining the entire model from scratch. This retraining process can be time consuming, computationally expensive, and impractical in fast-evolving medical and socio-demographic research.
- **RDR's Incremental Knowledge Acquisition:**
  In contrast, RDR allows for real-time updates without the need for full retraining. If a new socio-demographic factor is discovered to influence type 2 diabetes, an SME can immediately add a new rule without modifying the existing rule set. This ensures that the system remains dynamic and continuously improves over time.

6.4.4 Handling Evolving Medical Knowledge

Medical and socio-demographic knowledge is not static as new risk factors, environmental influences, and social determinants emerge over time. The ability to integrate these changes is crucial.

- **ML Struggles with Rapid Medical Advances:**
  ML models are trained on historical data, meaning they can quickly become outdated if new findings emerge (Ghassemi et al., 2020). Since retraining is cumbersome, ML-based KBSs are inherently static rather than evolving.
- **RDR's Ability to Stay Up to Date:**
  Because RDR operates incrementally, new findings can be immediately integrated as new rules. This feature is particularly advantageous for chronic disease management, where evolving social and environmental factors continually reshape the risk landscape for conditions like type 2 diabetes.

6.4.5 Empirical Comparison – J48 vs. RDR

To further illustrate the differences between ML-based and RDR-based KBS development, this section presents an empirical comparison using Weka's J48 decision tree algorithm. The same datasets used in the RDR experiments were processed through J48 to evaluate its performance on both the geographically specific dataset and the socio-demographic dataset.

Before running the J48 classification, preprocessing steps were necessary to align the datasets with ML requirements. Key adjustments included:

- removing the Diagnostic DM (years) attribute, as J48 would use it to generate trivial classification rules without considering other influential factors.
- standardising categorical values, particularly within the target attribute, ensuring that uppercase and lowercase values (e.g., Yes vs. yes) were treated uniformly.

Once the datasets were optimised, the J48 decision tree was trained and evaluated, providing insights into its classification behaviour. The following sections analyse J48's performance and compare it against the RDR-developed KBS.

6.4.5.1 Results from J48 on the Geographically Specific Dataset

The first dataset analysed through J48 was the geographically specific dataset, which included location-based features such as street names, suburb names, and postcodes.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Weka_2-weka.filters.unsupervised.attribute.Remove-R1-3-weka.filters.unsupervised.attribute.Remove-R23-24
Instances:    1200
Attributes:   28
              Times_Attended
              Withdrawn_Screening
              Gender
              Patient_Age
              Patient_Street
              Patient_Town
              Patient_Postcode
              GP_Street
              GP_Town
              GP_Postcode
              CVD_Status
              Diagnostic_CVD_years
              HT_Status
              Diagnostic_HT_years
              Alcohol
              Family_History_DM
              Family_History_CVD
              PHQ9
```

```
            Diet
            Exercise_Duration_hrsweek
            Exercise_Intensity
            Last_visit_to_GP
            Frequency_yr_diabetes_educator
            Waist_Circumference
            Heightm
            Weightkg
            BMI
            Target
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Patient_Town = Wodonga
|  Patient_Age <= 73
|  |  CVD_Status = No
|  |  |  HT_Status = No
|  |  |  |  Family_History_DM = No
|  |  |  |  |  Family_History_CVD = No
|  |  |  |  |  |  Patient_Street <= 5843
|  |  |  |  |  |  |  Patient_Age <= 60: No (22.0/3.0)
|  |  |  |  |  |  |  Patient_Age > 60: Yes (13.0/3.0)
|  |  |  |  |  |  Patient_Street > 5843: No (40.0/1.0)
|  |  |  |  |  Family_History_CVD = Yes: No (11.0)
|  |  |  |  Family_History_DM = Yes: No (55.0/4.0)
|  |  |  HT_Status = Yes
|  |  |  |  Diagnostic_HT_years <= 0.5
|  |  |  |  |  Family_History_CVD = No: Yes (6.0/1.0)
|  |  |  |  |  Family_History_CVD = Yes: No (2.0)
|  |  |  |  Diagnostic_HT_years > 0.5: No (6.0)
|  |  CVD_Status = Yes: Yes (3.0/1.0)
|  Patient_Age > 73
|  |  Gender = M: Yes (9.0/1.0)
|  |  Gender = F
|  |  |  CVD_Status = No
|  |  |  |  Times_Attended <= 1: No (4.0)
|  |  |  |  Times_Attended > 1
|  |  |  |  |  GP_Street <= 4528: Yes (7.0)
|  |  |  |  |  GP_Street > 4528: No (3.0/1.0)
|  |  |  CVD_Status = Yes: No (4.0)
Patient_Town = Wooragee: No (14.0/1.0)
Patient_Town = Sydney Olymic Park: No (1.0)
Patient_Town = Albury
|  Family_History_DM = No
|  |  Patient_Age <= 83: No (281.0/15.0)
|  |  Patient_Age > 83
|  |  |  GP_Postcode <= 2640: No (11.0/1.0)
|  |  |  GP_Postcode > 2640: Yes (7.0/1.0)
|  Family_History_DM = Yes
|  |  Weightkg <= 81.5
|  |  |  Patient_Street <= 1990: Yes (5.0)
|  |  |  Patient_Street > 1990: No (75.0/2.0)
|  |  Weightkg > 81.5
|  |  |  Heightm <= 1.8
|  |  |  |  Diagnostic_HT_years <= 0.5
|  |  |  |  |  Last_visit_to_GP <= 0
|  |  |  |  |  |  Waist_Circumference <= 118
|  |  |  |  |  |  |  Patient_Street <= 6563: Yes (12.0)
|  |  |  |  |  |  |  Patient_Street > 6563
|  |  |  |  |  |  |  |  BMI <= 29.1: Yes (7.0)
|  |  |  |  |  |  |  |  BMI > 29.1: No (5.0)
|  |  |  |  |  |  Waist_Circumference > 118: No (6.0)
|  |  |  |  |  Last_visit_to_GP > 0: Yes (8.0)
```

```
|  |  |  |  Diagnostic_HT_years > 0.5: No (5.0/1.0)
|  |  |  Heightm > 1.8: No (9.0)
Patient_Town = Lavington
|  Diagnostic_CVD_years <= 3
|  |  Waist_Circumference <= 119: No (184.0/15.0)
|  |  Waist_Circumference > 119
|  |  |  Waist_Circumference <= 127: No (4.0)
|  |  |  Waist_Circumference > 127: Yes (3.0)
|  Diagnostic_CVD_years > 3
|  |  Waist_Circumference <= 90: No (5.0)
|  |  Waist_Circumference > 90: Yes (8.0/1.0)
Patient_Town = Thurgoona
|  Patient_Street <= 2491: Yes (7.0)
|  Patient_Street > 2491
|  |  Patient_Age <= 68: No (47.0)
|  |  Patient_Age > 68
|  |  |  Patient_Street <= 9808: Yes (5.0)
|  |  |  Patient_Street > 9808: No (3.0)
Patient_Town = Bicton: No (1.0)
Patient_Town = Jindera
|  Heightm <= 1.7: No (14.0/1.0)
|  Heightm > 1.7: Yes (2.0)
Patient_Town = Barnawatha: No (6.0)
Patient_Town = Beechworth: No (9.0)
Patient_Town = Wangaratta
|  GP_Street <= 6955: No (9.0)
|  GP_Street > 6955: Yes (3.0)
Patient_Town = Culcairn
|  Waist_Circumference <= 99: No (6.0)
|  Waist_Circumference > 99: Yes (3.0)
Patient_Town = Baranduda: No (11.0)
Patient_Town = Unknown: No (1.0)
Patient_Town = Yackandandah: No (19.0/1.0)
Patient_Town = Wagga Wagga: No (1.0)
Patient_Town = Glenrowan: No (1.0)
Patient_Town = Howlong: No (23.0)
Patient_Town = Yarrawonga: No (3.0/1.0)
Patient_Town = Indigo Valley: No (11.0)
Patient_Town = Lake Rowan: No (1.0)
Patient_Town = Glenroy: Yes (5.0/1.0)
Patient_Town = Oaklands: No (2.0)
Patient_Town = Myrtleford: No (1.0)
Patient_Town = Albury East: No (2.0)
Patient_Town = Bellbridge: No (5.0)
Patient_Town = Chiltern: No (1.0)
Patient_Town = Daysdale: No (11.0)
Patient_Town = Tallangatta
|  GP_Street <= 5933
|  |  Family_History_DM = No
|  |  |  Times_Attended <= 3: No (5.0/1.0)
|  |  |  Times_Attended > 3: Yes (3.0)
|  |  Family_History_DM = Yes: Yes (2.0)
|  GP_Street > 5933: Yes (10.0)
Patient_Town = Point Lonsdale: No (2.0)
Patient_Town = Walla Walla: No (12.0)
Patient_Town = Rosewhite: No (1.0)
Patient_Town = Tangambalanga: No (4.0)
Patient_Town = Moorwatha: No (1.0)
Patient_Town = Wirlinga: No (6.0)
Patient_Town = Barnawartha: No (10.0/1.0)
Patient_Town = Shepparton: No (6.0/1.0)
Patient_Town = North Albury
|  CVD_Status = No: No (4.0)
|  CVD_Status = Yes: Yes (2.0)
Patient_Town = Woomaragama: No (2.0)
Patient_Town = West Wodonga: Yes (5.0/1.0)
Patient_Town = Morven: No (1.0)
```

```
Patient_Town = Kiewa: Yes (2.0/1.0)
Patient_Town = Wodonga West: No (4.0)
Patient_Town = Table Top: No (2.0)
Patient_Town = Mount Beauty: No (15.0)
Patient_Town = Kergunyah: No (1.0)
Patient_Town = Rutherglen: No (5.0)
Patient_Town = East Albury: No (6.0)
Patient_Town = Mitta Mitta: No (1.0)
Patient_Town = Woomargama: No (2.0)
Patient_Town = Bullioh
|   Waist_Circumference <= 0: Yes (2.0)
|   Waist_Circumference > 0: No (4.0)
Patient_Town = Dederang: No (4.0)
Patient_Town = South Albury: No (4.0)
Patient_Town = Huon: No (1.0)
Patient_Town = Eskdale: Yes (4.0)
Patient_Town = Holbrook: Yes (2.0)
Patient_Town = Norris Park , Lavington: Yes (4.0)
Patient_Town = Norris Park: Yes (4.0)
Patient_Town = Sterling, ACT: Yes (3.0)
Patient_Town = Vacy: No (2.0)
Patient_Town = Benalla: No (1.0)
Patient_Town = Hume Weir: No (1.0)
Patient_Town = Gerogery: No (1.0)
Patient_Town = Bright: No (1.0)

Number of Leaves  :         103

Size of the tree:    143


Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances        1139               94.9167 %
Incorrectly Classified Instances        61                5.0833 %
Kappa statistic                          0.7969
Mean absolute error                      0.0902
Root mean squared error                  0.2123
Relative absolute error                 33.085 %
Root relative squared error             57.5595 %
Total Number of Instances             1200

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.744    0.011    0.929      0.744    0.826      0.804    0.931     0.844     Yes
                 0.989    0.256    0.952      0.989    0.970      0.804    0.931     0.980     No
Weighted Avg.    0.949    0.217    0.948      0.949    0.947      0.804    0.931     0.958

=== Confusion Matrix ===

   a   b   <-- classified as
 145  50 |   a = Yes
  11 994 |   b = No
```

Figure 6. 2 . J48 output using the training dataset from the geographically specific dataset. This figure shows the statistics obtained by J48.

The decision tree constructed by J48 relied heavily on location-based attributes, forming branches based on patient suburb, postcode, and GP location. While this provided reasonable accuracy within the training dataset, it introduced significant limitations in generalizability, as geographic attributes tend to be specific to the dataset's regional scope. These patterns are visualised in figure 6.2, which provides a concise overview of the results.

To assess the generalization capability of J48, the trained model was applied to the unseen production dataset.

```
J48 pruned tree
------------------

Last_visit_to_diabetes_educator <= 0
|   Frequency_yr_GP <= 1: No (1119.0/136.0)
|   Frequency_yr_GP > 1
|   |   CVD_Status = No: Yes (22.0/5.0)
|   |   CVD_Status = Yes
|   |   |   Diagnostic_HT_years <= 3: No (11.0/1.0)
|   |   |   Diagnostic_HT_years > 3: Yes (2.0)
Last_visit_to_diabetes_educator > 0: Yes (46.0/7.0)

Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.07 seconds

=== Summary ===

Correctly Classified Instances        1051               87.5833 %
Incorrectly Classified Instances       149               12.4167 %
Kappa statistic                          0.3849
Mean absolute error                      0.217
Root mean squared error                  0.3294
Relative absolute error                 79.6004 %
Root relative squared error             89.2811 %
Total Number of Instances             1200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.297    0.012    0.829      0.297    0.438      0.449    0.644     0.368     Yes
                 0.988    0.703    0.879      0.988    0.930      0.449    0.644     0.879     No
Weighted Avg.    0.876    0.590    0.871      0.876    0.850      0.449    0.644     0.796

=== Confusion Matrix ===

   a   b   <-- classified as
  58 137 |   a = Yes
  12 993 |   b = No
```

Figure 6. 3 J48 output using the production dataset from the geographically specific dataset. This figure shows the statistics obtained by J48.

Figure 6.3 shows the J48 results from applying the model to the unseen production dataset for the geographically specific dataset. The production dataset results revealed a notable drop in

sensitivity, indicating that J48 struggled to classify unseen cases effectively. This behaviour suggests that while ML-based decision trees can fit training data well, their reliance on specific attributes (like location) makes them less adaptable when applied to broader datasets.

6.4.5.2 Results from J48 on the Socio-Demographic Dataset

The second dataset processed through J48 replaced location-based attributes with socio-demographic attributes, including median income, education level, and average household size. Figure 6.4 show the output results on the socio-demographic training dataset.

```
J48 pruned tree
------------------

Last_visit_to_diabetes_educator <= 0: No (471.0/119.0)
Last_visit_to_diabetes_educator > 0
|   Heightm <= 1.54: No (2.0)
|   Heightm > 1.54: Yes (27.0/2.0)

Number of Leaves  :     3

Size of the tree :      5


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         370               74      %
Incorrectly Classified Instances       130               26      %
Kappa statistic                          0.1568
Mean absolute error                      0.3777
Root mean squared error                  0.441
Relative absolute error                 91.9914 %
Root relative squared error             97.3866 %
Total Number of Instances              500

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.986    0.868    0.737      0.986    0.844      0.250    0.519     0.715     No
                 0.132    0.014    0.792      0.132    0.226      0.250    0.519     0.375     Yes
Weighted Avg.    0.740    0.622    0.753      0.740    0.666      0.250    0.519     0.617

=== Confusion Matrix ===

   a   b   <-- classified as
 351   5 |   a = No
 125  19 |   b = Yes
```

Figure 6. 4 . J48 output using the training dataset from the socio-determinant enhanced dataset. This figure shows the statistics obtained by J48.

The decision tree generated by J48 adapted to socio-demographic variables but continued to exhibit similar challenges in relation to the geographically specific dataset. Aggregated data (e.g.,

median income and education levels) led to overlapping classifications, making it difficult for J48 to consistently differentiate between diabetic and non-diabetic cases.

When applied to the production dataset, the same generalization issues emerged, with accuracy dropping significantly, as shown in Figure 6.5. This reinforces a key limitation of ML-driven decision trees as they rely on static rules derived from the training dataset without the flexibility to refine or adjust based on expert insights.

```
J48 pruned tree
------------------

Last_visit_to_GP <= 0
|   Last_visit_to_diabetes_educator <= 1: No (1118.0/136.0)
|   Last_visit_to_diabetes_educator > 1
|   |   CVD_Status = No
|   |   |   Family_History_CVD = No
|   |   |   |   Avg. Household Size <= 2.3: No (3.0/1.0)
|   |   |   |   Avg. Household Size > 2.3: Yes (14.0)
|   |   |   Family_History_CVD = Yes
|   |   |   |   Gender = M: No (2.0)
|   |   |   |   Gender = F: Yes (3.0/1.0)
|   |   CVD_Status = Yes
|   |   |   Diagnostic_HT_years <= 3: No (11.0/1.0)
|   |   |   Diagnostic_HT_years > 3: Yes (2.0)
Last_visit_to_GP > 0: Yes (46.0/7.0)

Number of Leaves  :     8

Size of the tree :      15


Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1040               86.7389 %
Incorrectly Classified Instances       159               13.2611 %
Kappa statistic                          0.3436
Mean absolute error                      0.223
Root mean squared error                  0.3401
Relative absolute error                 81.7397 %
Root relative squared error             92.1526 %
Total Number of Instances             1199

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.272    0.017    0.757      0.272    0.400      0.401    0.587     0.342     Yes
                 0.983    0.728    0.874      0.983    0.925      0.401    0.587     0.854     No
Weighted Avg.    0.867    0.613    0.855      0.867    0.840      0.401    0.587     0.771

=== Confusion Matrix ===

   a   b   <-- classified as
  53 142 |   a = Yes
  17 987 |   b = No
```

Figure 6. 5 J48 output using the production dataset from the socio-determinant enhanced dataset. This figure shows the statistics obtained by J48.

The empirical comparison between J48 and RDR highlights several critical advantages of RDRs in KBS development:

- **Incremental Adaptation**: Unlike J48, which builds a static decision tree, RDRs continuously evolve as new cases are introduced, allowing for real-time adjustments and refinements.
- **Expert-Driven Rule Construction**: RDRs allow domain experts to inject their knowledge and experience into the system, ensuring that rules reflect real-world patterns beyond just statistical correlations in the training data.
- **Better Handling of Edge Cases**: While J48 struggled with edge cases and outliers, RDRs offered more nuanced rule construction, improving classification accuracy over time.
- **Scalability Beyond a Single Dataset**: J48's reliance on fixed rules means that expanding to new datasets requires retraining from scratch. RDRs, on the other hand, allow incremental modifications without the need for complete retraining.
- **Generalisation Across Regions**: The geographic KBS created using J48 demonstrated severe limitations in generalizability due to its reliance on location-based attributes. The RDR-based KBS, in contrast, was designed to extend beyond a single region, particularly in the socio-demographic model, which is adaptable to different populations and newly discovered relationships between socio-demographic factors and diabetes risk.
- **Performance**: The empirical RDR results presented in Section 6.2 illustrate that the socio-demographic KBS consistently outperforms the geographic KBS, particularly in its ability to maintain predictive accuracy across production datasets.

The J48 decision tree experiments confirmed that ML-based models require extensive data and retraining to maintain performance across unseen datasets. In contrast, RDR-based KBS construction demonstrated greater flexibility, allowing for incremental updates, expert driven refinements, and the better handling of socio-demographic complexity. These findings support the broader argument for RDR superiority in dynamically evolving domains, where expert knowledge and incremental refinements play a crucial role in maintaining an accurate and adaptable decision support system.

6.5 Summary

This chapter provided a comprehensive comparative analysis between RDRs and ML in the development of a KBS for type 2 diabetes prediction. The analysis emphasised both theoretical and practical differences, demonstrating why RDRs outperform ML in knowledge-based system construction, particularly in data-scarce environments.

To provide a clearer overview, Table 6.2 summarises the key differences between ML-based and RDR-based KBS development.

Feature	Machine Learning (ML)	Ripple-Down Rules (RDRs)
Data Requirement	Requires large datasets for training	Can function with minimal data
Interpretability	Often a black box	Fully explainable and interpretable
Adaptability	Requires full retraining for updates	Allows incremental rule addition
Handling New Medical Knowledge	Struggles with rapidly evolving research	Easily integrates new knowledge
Computational Cost	High; requires extensive computational resources	Low; rules are added manually based on expert input
Best Use Case	Works well for pattern recognition in large datasets	Best suited for expert-driven knowledge bases

Table 6. 2 Comparison of the various development features of RDRs and ML in the development of a KBS.

The key comparative findings are summarised in Table 6.2, which highlights the fundamental strengths and limitations of both methodologies. The analysis reinforced that while ML thrives on large datasets, it struggles in situations where data is limited, evolving, or requires human interpretability. Conversely, RDRs provide a transparent, expert-driven, and incrementally adaptable framework, making them particularly well suited for socio-demographic modelling and real-world decision support.

Additionally, our empirical results using WEKA's J48 validated these theoretical distinctions. J48 required extensive data preprocessing, suffered from a lack of interpretability, and demonstrated

limitations in handling evolving knowledge. In contrast, RDR-based KBS construction allowed expert-driven rule adjustments, supported incremental learning, and adapted dynamically to new insights, an essential feature in public health applications where new socio-demographic determinants may emerge over time.

The key takeaway points and future considerations from this chapter are as follows:

- ML approaches require large, clean datasets, lack transparency, and must be retrained from scratch to incorporate new knowledge.
- RDRs thrive in expert-driven environments, require minimal data, and allow real-time incremental rule addition without full retraining.
- The adaptability of RDR based KBSs makes them better suited for healthcare decision making, particularly in socio-demographic research, where domain expertise and evolving knowledge are crucial.
- Future research should explore hybrid approaches, combining RDRs with ML-based feature selection to optimize both pattern recognition (ML) and expert-driven knowledge integration (RDRs).
- The comparison between the geographic KBS and the socio-demographic KBS in the production dataset demonstrated that socio-demographic attributes provide a more generalizable and transferable model. The socio-demographic KBS consistently achieved higher accuracy (80%) than the geographic KBS (60%), reinforcing the limitations of geographic-based predictions beyond specific regions.

The ML landscape in healthcare continues to evolve rapidly, with new deep learning models and transfer learning approaches achieving improved performance on diverse datasets (Fregoso-Aparicio et al. 2021; Smith et al. 2023). Recent systematic reviews emphasize both the potential and limitations of ML, particularly regarding bias, data quality (Compton & Kang 2021), and real-world generalizability (Ghassemi et al., 2020; An et al., 2023).

Ultimately, this study highlights RDRs as the superior methodology for constructing knowledge-based systems in socio-demographic and healthcare applications.

This chapter's findings further validate the conclusions drawn in section 6.2, emphasising socio-demographic KBS's greater scalability, transferability, and resilience when applied to previously unseen data. The next chapter further expands on these findings, examining the broader implications for predictive modelling in chronic disease prevention and discussing how

knowledge-driven AI can be integrated into modern healthcare decision support systems, as well as in various other systems, not healthcare related.

Chapter 7: Summary, Discussion and Future Work

The research conducted in this thesis explored the potential of socio-demographic and geographic factors in predicting type 2 diabetes and the comparative effectiveness of Ripple-Down Rules (RDR) and machine learning (ML) methodologies in constructing a knowledge-based system (KBS). This study contributes to the broader field of public health by examining how socio-demographic factors influence diabetes prediction and management. It further explores the role of decision support systems in enhancing healthcare interventions. This chapter concludes the thesis by synthesising key insights gained from the research and critically evaluating the implications of the developed KBS in healthcare applications. It highlights the advantages of RDR-based systems in expert-driven environments and data-scarce contexts, while also addressing the broader challenges of integrating socio-demographic factors into predictive models. Furthermore, this chapter situates the research findings within the larger discourse on decision support technologies, healthcare policy, and future directions in intelligent systems for chronic disease management.

The chapter is divided into eight sections, each addressing key elements of the research and its implications for diabetes management. Section 7.1 provides a summary of the findings of the thesis. Section 7.2 explores the role of socio-demographic data in refining predictive models for chronic disease management. It considers the challenges of data aggregation, generalization, and real-world applicability in healthcare decision making. Section 7.3 provides a critical comparison of the RDR and ML methodologies used in this research, analysing their respective advantages and limitations in KBS development. This section discusses the importance of incremental rule development, expert involvement, and adaptability in decision support systems. Section 7.4 addresses the issues of data scarcity and scalability. Section 7.5 discusses the broader implications for policy in public health and clinical decision support tools. This section connects the study's findings to existing healthcare frameworks, emphasising how adaptive systems can enhance disease prediction and intervention strategies. Section 7.6 extends this discussion beyond healthcare, such as finance, insurance, and risk management. It highlights the scalability and adaptability of incremental rule-based learning in domains where expert knowledge is critical for decision making. Section 7.7 examines future research directions of this research exploring how it can also be applied to other industries. Section 7.8 concludes with final remarks reflecting on the study's contributions to KBS development and its potential implications for healthcare technology and policy.

7.1 Summary of Findings

The findings of this research highlight the importance of socio-demographic factors in type 2 diabetes prediction and the effectiveness of KBSs in capturing these relationships.

Key insights from the research include:

- **The Influence of Socio-Demographic Factors on Type 2 Diabetes:**
  Socioeconomic determinants such as income levels, education, household size, and access to healthcare significantly impact diabetes risk. These findings align with previous studies (e.g., Hill et al., 2013; Schwerdtle, 2016; Frier et al., 2020; Cooper, Mowbray & Johnson 2024) that emphasize the social determinants of health in chronic disease management.
- **Advantages of RDRs in KBS Development:**
  Unlike ML-based approaches, RDRs allow incremental rule development, integrate expert knowledge, and function effectively in data-limited environments. This adaptability ensures that new medical findings can be incorporated without requiring complete system retraining.
- **Comparative Performance of RDRs and ML:**
  Experimental results showed that RDR-based KBSs outperformed ML (J48 decision trees) in socio-demographic data interpretation. RDRs provided higher interpretability and adaptability, whereas ML models struggled with data scarcity and generalizability.

The following sections explore these findings in detail, starting with an in-depth analysis of how socio-demographic factors relate to type 2 diabetes.

7.2. The Role of Socio-Demographic Factors in Diabetes Prediction

Rather than proving causality, these results indicate that the RDR-based KBS detects stable associations between socio-demographic factors and risk outcomes in this dataset. The findings reinforce existing studies (Braveman et al., 2011; Cutler & Lleras-Muney, 2006; Cooper, Mowbray & Johnson 2024) that link socio-economic status to health outcomes. The socio-demographic KBS developed in this study provides a scalable framework for incorporating these determinants into predictive healthcare models.

The relationship between socio-demographic factors and Type 2 diabetes (T2D) is well established, with research highlighting how social determinants such as income, education, employment, and healthcare access influence diabetes prevalence and management (Hill, Nielsen & Fox, 2013; Sauliune & Kalediene, 2015; Schwerdtle, 2016; WHO 2024). Consistent with section 2.2, this thesis does not re-establish those relationships; rather, it operationalises them as computable, auditable rules within an RDR-based KBS. The findings of this study align with broader public health literature, which consistently underscores the significance of these factors in shaping health outcomes.

The results of the socio-demographic KBS indicate that income-related attributes play a crucial role in refining diabetes risk prediction. This was noted in Hill et al. (2013) who emphasised that socioeconomic disparities significantly impact diabetes prevalence, with lower-income populations exhibiting higher rates of diabetes due to limited access to healthcare, healthy food, and diabetes education. The findings suggest that economic constraints not only influence diabetes onset but also affect disease progression, as individuals from lower income backgrounds may experience delayed diagnoses and poorer management outcomes due to financial and logistical barriers to care. Indeed, the argument of the thesis that socio-demographic KBS models can bridge healthcare gaps by identifying at-risk populations and enabling targeted interventions is also supported by Schwerdtle (2016) who explores the impact of healthcare accessibility on diabetes risk and management. The study highlights how regional disparities in healthcare infrastructure contribute to differences in diabetes outcomes, particularly in rural and underserved areas. The work here concurs that individuals living in regions with limited healthcare facilities are at a higher risk of undiagnosed and poorly managed diabetes, reinforcing the importance of considering geographic accessibility alongside socio-economic determinants.

Additionally, Bull et al. (2017) discusses the critical role of physical activity and sedentary behaviour in diabetes risk, stressing that lower-income communities often have reduced access to recreational spaces and safe environments for exercise. Their findings suggest that socioeconomic constraints extend beyond financial limitations to include environmental and infrastructural barriers, which can discourage healthy behaviours and increase diabetes risk. The socio-demographic KBS incorporated exercise-related variables, further demonstrating how lifestyle factors interact with economic status to shape disease prevalence.

Frier et al. (2020) further highlight the intersection between financial stress and mental health in diabetes management. The study indicates that economic hardship is strongly associated with increased psychological distress, which in turn contributes to poor diabetes self-care, higher

HbA1c levels, and increased complications. These findings suggest that psychosocial stressors should be integrated into predictive models, reinforcing the argument for comprehensive socio-demographic KBS models that account for the broader determinants of health beyond purely medical or lifestyle attributes.

Beyond individual-level socio-demographic factors, systemic issues such as government policies, food security, and education systems also contribute to diabetes disparities. Research indicates that countries with strong social safety nets and accessible healthcare services tend to have lower diabetes-related complications (Smith et al., 2021). Similarly, public health initiatives targeting education and awareness have been shown to mitigate some of the socio-economic barriers to diabetes prevention and management (Jones et al., 2018).

The integration of socio-demographic factors into diabetes prediction models, such as the one developed in this research, offers a pathway toward more equitable healthcare interventions. By identifying high-risk groups based on both individual attributes (e.g., income, education) and structural factors (e.g., healthcare accessibility, policy frameworks), KBS models provide data-driven insights to inform public health policies and targeted intervention programs. It is important to recognise the multifaceted role of socio-demographic determinants in diabetes risk and management and reinforces the importance of integrating these factors into predictive models. The results from this study align with broader public health findings, demonstrating the validity and potential of socio-demographic KBS models in identifying at-risk populations and improving health outcomes through data-informed decision support systems.

7.3. Comparative Performance of RDRs and ML

A comparative analysis of RDR-based and ML-based KBS development showed distinct advantages and limitations for each approach:

- **Interpretability:** RDRs are fully explainable, whereas ML models often operate as black boxes.
- **Adaptability:** RDRs allow incremental updates without requiring model retraining, unlike ML models, which need large scale retraining for even minor adjustments.
- **Data Requirements:** ML relies on vast amounts of labelled data, whereas RDRs function effectively even with limited data availability.

- **Real-Time Knowledge Integration:** RDRs enable domain experts to inject new knowledge dynamically, making them ideal for fields like healthcare, where new research findings frequently emerge.

These findings align with previous studies (Compton & Kang, 2021; Eyerich et al., 2019), emphasising the utility of RDRs in knowledge-based system development.

The comparative analysis between RDRs and ML in KBS development underscores key distinctions in interpretability, adaptability, and data requirements, all of which have significant implications for healthcare decision support. Chapter 6 detailed how both approaches performed when applied to geographic and socio-demographic datasets, revealing the strengths and limitations of each method in capturing diabetes risk factors.

One of the most critical distinctions between RDRs and ML is their respective dependency on data availability. ML models, such as decision trees (J48 in Weka) and deep learning algorithms, require large, well-structured datasets to train effectively. Hill et al. (2013) note that many ML models struggle in medical research due to the limited availability of labelled healthcare data, making them less suitable for conditions where patient data is often incomplete or sparse. In contrast, RDRs is agnostic to data scarcity as they do not rely on large-scale training datasets but rather on incremental rule development guided by subject matter experts (SMEs). Schwerdtle (2016) argues that expert-driven systems play a crucial role in domains where real-world data is incomplete or difficult to generalize, reinforcing the importance of incremental knowledge acquisition in developing effective healthcare KBS models.

This difference was reflected in the research findings, while the ML-based J48 model required extensive preprocessing and a substantial dataset to function effectively, the RDR-based KBS achieved comparable accuracy with a fraction of the data. The ability of RDRs to function effectively in limited-data environments makes them particularly valuable for public health applications in low-resource settings, where access to extensive patient records may be constrained.

7.4 Addressing Data Scarcity and Scalability

One of the most significant challenges in predictive modelling for type 2 diabetes is the availability and quality of data. While ML approaches require large, well-structured datasets for accurate training, RDRs excel in data-scarce environments by allowing incremental knowledge acquisition

from experts. This section examines the impact of data scarcity on the development of a KBS and explores strategies to ensure its scalability and adaptability across different populations and healthcare contexts.

Key insights from the research include:

- **Challenges and Advantages of RDRs in Data-Limited Environments:**
  Unlike structured clinical data, socio-demographic datasets often suffer from missing values, aggregation issues, and regional inconsistencies. These limitations create challenges in ensuring robust predictive performance, particularly when applied to unseen populations
- **Why RDRs in Data Limited Environments:**
  Unlike ML models, which struggle when data is sparse, RDRs allow experts to iteratively refine rules based on their domain knowledge. This approach makes RDR based KBS models particularly effective in rural or low-resource settings, where comprehensive datasets may not be available.
- **Scalability of the Socio-Demographic KBS:**
  The socio-demographic KBS developed in Chapter 5 demonstrated greater adaptability than its geographic counterpart, as it relies on demographic trends rather than location-specific attributes. This enables broader applicability across different populations and geographic regions.
- **Future Directions for Expanding KBS Usability:**
  Incorporating dynamic rule updates and hybrid approaches where RDR-based models integrate external public health data sources could enhance both scalability and predictive power. The ability to incrementally update an RDR system ensures that it remains aligned with emerging healthcare insights.

The following sections further explore these topics, namely data scarcity challenges, the scalability of RDR-based KBS models in, and recommendations for future research in Section 7.3.

7.4.1 Challenges in Socio-Demographic Data Availability

One of the primary obstacles encountered in this research was obtaining high-quality, integrated socio-demographic and clinical datasets suitable for developing a KBS for type 2 diabetes

prediction. While large-scale medical datasets exist, they often lack socio-demographic indicators, and conversely, socio-demographic datasets are rarely linked to clinical health outcomes. This lack of integration posed a significant barrier to constructing a predictive model that incorporated social determinants of health.

Initially, multiple official sources were explored in an attempt to obtain comprehensive datasets. These included:

- **The Australian Bureau of Statistics (ABS):** Provided extensive socio-demographic data but lacked corresponding health indicators.
- **Diabetes Australia and Diabetes NSW:** Offered statistical insights into diabetes prevalence but without granular socio-economic linkage.
- **NSW Health:** While maintaining extensive clinical datasets, these were primarily focused on medical records and did not systematically incorporate socio-demographic variables.

Despite an exhaustive two-year search, no single dataset provided both health-related and socio-economic data in a structured format. This challenge underscored the difficulty of creating a truly comprehensive predictive system for T2D that reflects real-world social and medical intersections.

The breakthrough came from a diabetes complications research initiative at Charles Sturt University, which had compiled data from the Albury/Wodonga region, a border community between New South Wales and Victoria, Australia. This dataset contained medical attributes such as BMI, age, blood pressure, and family history of diabetes, alongside basic geographic information like patient and GP locations.

However, socio-demographic factors were not included directly. To address this, additional publicly available datasets from local government sources (Albury/Wodonga councils) and the ABS were integrated. Key attributes such as median household income, education levels, and average family size were extracted and mapped to corresponding geographic identifiers within the dataset.

These challenges highlighted several key limitations that influenced the development of the KBS:

1. **Limited Representativeness:**
    - The dataset covered a specific geographic region, raising concerns about its generalizability to other populations.

- Socio-demographic variables were regional aggregates, meaning that individual-level socio-economic factors were not captured directly.

2. **Data Preprocessing Complexities:**
    - Significant data cleaning was required, including handling missing values, standardising formats, and normalising socio-demographic attributes.
    - Some attributes had to be estimated or inferred, such as using median income levels for an area rather than individual income data.
3. **Challenges in Balancing Medical and Socio-Demographic Attributes:**
    - Socio-demographic factors often involve continuous variables (e.g., median income) rather than discrete classifications (e.g., postcode).
    - Unlike medical indicators, which have established clinical thresholds, socio-demographic variables required contextual interpretation to be useful in rule construction.

These limitations reinforced the importance of using RDRs for this research. Since RDRs allow for incremental adaptation, they were able to accommodate the progressive refinement of rules as more nuanced socio-demographic relationships were uncovered. This flexibility was particularly valuable given the uncertainties and inconsistencies in the available data.

7.4.2 Scalability of the RDR Approach

Despite these challenges, the research findings suggest that the RDR methodology offers a scalable solution that can be applied to different populations with minimal modifications. The flexibility of RDRs allows new rules to be incorporated based on regional variations, making them a viable tool for expanding diabetes risk assessment beyond the original dataset.

The scalability of RDRs is one of its most defining strengths, allowing for incremental expansion, continuous adaptation, and real-time refinement as new information becomes available (Compton and Jansen, 1990). Unlike machine learning (ML) models, which often require complete retraining when incorporating new data, RDRs facilitate a dynamic knowledge building process that seamlessly integrates new findings without disrupting the system's existing rule base.

While this research has focused on type 2 diabetes as a test case, the RDR approach is not inherently limited to diabetes prediction. Rather, type 2 diabetes was chosen as a vehicle to

demonstrate the superiority of RDRs in incrementally developing a KBS. This method can be extended to other medical conditions, including but not limited to:

- **Coronary Care:** incorporating evolving risk factors such as hypertension, lipid profiles, and lifestyle attributes to refine cardiovascular disease prediction.
- **Neurological Disorders:** incrementally building rules to predict and classify conditions like Alzheimer's disease, Parkinson's, or epilepsy as new biomarkers and risk factors emerge.
- **Oncology & Cancer Screening:** dynamically integrating new genetic markers, treatment responses, and patient history to improve early cancer detection and prognosis prediction (Sauer et al., 2020, Eyerich et al., 2019).

The adaptability of RDRs makes them particularly valuable in fast evolving medical fields, where research frequently uncovers new risk factors, treatment pathways, and diagnostic methodologies. Unlike static models, RDRs integrate new insights seamlessly, ensuring that decision support systems remain aligned with the latest clinical knowledge.

Another critical scalability advantage of RDRs lies in their ability to integrate human expertise. Unlike ML models, which derive their conclusions purely from data, RDRs allow subject matter experts (SMEs), such as clinicians, researchers, and medical practitioners, to directly inject their knowledge and experience into the system (Compton and Kang, 2021).

This capability is especially beneficial in medical contexts where certain causal relationships may not yet be statistically validated in large datasets but are well understood by experts. For instance, a cardiologist may recognize an emerging but not yet widely documented link between a specific medication and increased heart attack risk. With RDRs, such expert insights can be encoded as provisional, auditable rules tied to cornerstone cases and clearly flagged for review; they trigger monitoring or data collection rather than automated decisions until validated on held-out data or external evidence.

This human-in-the-loop approach enhances the interpretability and clinical trustworthiness of RDR-based systems. By combining empirical data with expert-driven rules, the system achieves higher accuracy, transparency, and adaptability, which is particularly valuable in domains where medical decisions must be explainable, and evidence based.

While this research has primarily focused on medical decision making, the principles underlying RDR-based knowledge acquisition are not restricted to healthcare. The incremental, expert-

driven approach of RDRs can be applied in various industries that require decision support systems capable of adapting to new conditions.

1. **Insurance Industry:**
    - **Risk Assessment & Premium Calculation:** RDRs can help dynamically adjust insurance premiums based on evolving risk factors (Džeroski et al., 2001).
    - Example: Based on vehicle type, driver age, accident history, and regional crime rates, RDRs can refine auto insurance premium calculations.
    - **Home & Life Insurance:** The approach could integrate climate risks, lifestyle changes, and emerging medical trends to adjust insurance pricing in real time.
2. **Financial & Investment Sectors:**
    - **Foreign Exchange (Forex) Prediction:** By incorporating macroeconomic indicators, market trends, and geopolitical events, an RDR-based system could provide adaptive currency forecasting models.
    - **Stock Market & Investment Analysis:** An RDR model could continuously integrate new market signals, allowing traders to make adaptive investment decisions.
    - **Loan & Credit Scoring:** RDRs can enhance predictive models used by banks to assess creditworthiness, incorporating new financial behaviours, policy changes, and economic conditions.
3. **Legal & Regulatory Compliance:**
    - RDRs can be used in regulatory compliance systems, dynamically adjusting legal interpretations based on new case precedents, government policies, and evolving regulatory frameworks.
4. **Cybersecurity & Threat Detection:**
    - **Adaptive Threat Intelligence:** Cybersecurity systems using RDRs can integrate newly discovered vulnerabilities, hacker strategies, and emerging attack vectors to refine real-time threat mitigation rules.

These examples are illustrative of potential transfer; they fall outside the evaluated scope of this thesis and would require domain-specific validation in future work.

7.5 Implications for Policy and Practice

The integration of socio-demographic factors into type 2 diabetes risk prediction has profound implications for healthcare policy and clinical practice. This research highlights the importance of moving beyond clinical indicators to incorporate broader social determinants of health, which play a pivotal role in shaping health outcomes (Cooper, Mowbray & Johnson 2024). Policies that acknowledge and integrate these factors into preventive strategies and healthcare planning could significantly enhance public health interventions (WHO, 2024; Hill-Briggs et al., 2021; Braveman & Gottlieb, 2014; Marmot, 2005). A more detailed discussion on this is provided in Section 7.5.1.

The ability to incrementally refine a KBS using RDRs introduces a new paradigm in decision support tools, particularly for public health agencies and policymakers. Unlike static machine learning models, RDR-based systems allow decision makers to incorporate evolving socio-demographic insights, leading to more adaptive and evidence-based policy interventions (Compton & Jansen, 1990; Richards et al., 2017).

Beyond public health, the scalability of this approach suggests its applicability in health insurance risk assessments, resource allocation for chronic disease management, and targeted community health programs. By embedding RDR-driven decision support tools into existing policy frameworks, governments and health organizations could more effectively identify high-risk populations, optimize resource distribution, and improve access to preventive care (Hill et al., 2013; Frier et al., 2020).

7.5.1 Integration of Socio-Demographic Factors in Public Health Policy

Given the strong relationship between socio-demographic factors and diabetes risk, public health strategies should incorporate these determinants into prevention and intervention frameworks. Policymakers can leverage socio-demographic KBS models to design targeted educational programs, improve healthcare accessibility in high-risk areas, and allocate resources more effectively.

7.5.2 Enhancing Clinical Decision Support Systems

The integration of socio-demographic insights into clinical decision support systems (CDSS) represents a significant advancement in personalised medicine. By combining patient-specific medical and lifestyle data with broader socio-economic variables, healthcare providers can generate more accurate risk assessments and intervention strategies tailored to individual patients. Unlike public health policies, which address population-wide strategies, CDSS enhancements focus on improving decision making at the patient level.

7.6 Broader Applications Beyond Healthcare

While this research has focused on type 2 diabetes prediction, the incremental development approach using RDRs is highly adaptable and extends beyond healthcare applications. The ability to incorporate expert knowledge, refine decision rules over time, and dynamically adjust to new data makes RDR-based knowledge systems particularly valuable in finance, insurance, and risk assessment industries.

7.6.1 Application in the Insurance Industry

The insurance sector relies heavily on risk assessment models to determine premiums, eligibility, and policy adjustments. Traditional actuarial models often depend on static risk profiles, which can quickly become outdated due to shifting market conditions. RDRs offer a dynamic alternative, allowing insurance companies to continuously refine underwriting policies based on real-world claims data, emerging risk trends, and regulatory changes (Frees et al., 2016; Charpentier, 2020). Illustrative insurance use-cases (health, auto, life) are summarised in section 7.6.1; we do not duplicate them here.

7.6.2 Application in Finance and Investment

The financial industry, particularly in areas such as risk modelling, credit scoring, and fraud detection, relies heavily on predictive analytics. ML models dominate this space, but their static nature can create blind spots when dealing with rapidly changing economic conditions. RDRs offer a more adaptable alternative, enabling financial institutions to incorporate new economic indicators, regulatory changes, and evolving consumer behaviours (Altman et al., 2017; Goodell, 2020).

For instance:

- **Credit risk assessment:** An RDR-driven KBS could incrementally refine credit scoring models based on emerging financial behaviours, rather than relying solely on historical credit reports.
- **Fraud detection:** Traditional fraud models often fail to detect new fraud tactics until extensive data retraining occurs. RDRs can help rapidly adjust detection rules based on real-time fraud patterns.
- **Stock market predictions:** Financial analysts could combine expert intuition with RDR-driven insights, allowing the model to continuously refine investment strategies based on market conditions and global economic indicators.

These are illustrative research directions; in domains like finance, any use would require domain-specific data, validation, and compliance review (no causal claims are made).

7.6.3 Broader Decision Support Systems

Beyond healthcare, insurance, and finance, RDRs can enhance decision making in fields such as supply chain management, legal compliance, and cybersecurity (Chen et al., 2021). In these domains, where expert-driven insights are crucial, RDRs bridge the gap between machine learning models and human expertise, allowing organizations to dynamically refine knowledge bases as new insights emerge.

From financial risk assessment to cybersecurity, RDR-based KBS solutions offer a scalable and interpretable alternative to traditional AI systems.

7.7 Future Research Directions

I now return to healthcare-specific directions; cross-industry applications were noted above as illustration only. The methodologies developed in this research can be extended to other chronic diseases, such as cardiovascular diseases and hypertension, which are also influenced by socio-demographic determinants.

7.7.1. Expanding RDR-Based KBS Development to Other Industries' KBS models.

While this research has demonstrated the efficacy of RDRs in constructing a KBS for type 2 diabetes prediction, future research should explore how RDR-based models can be further optimised across industries requiring dynamic decision support. Areas such as finance, insurance, and risk management present unique challenges that require industry-specific adaptations of RDR methodologies.

In finance, decision making is heavily influenced by a combination of economic indicators, market trends, risk assessment models, and regulatory compliance rules. Traditional machine learning models used in financial forecasting often rely on vast datasets and statistical pattern recognition, but they lack interpretability and adaptability to sudden economic shifts.

- Future research could explore how RDR-based KBS models can be incrementally updated to incorporate new financial regulations, economic events, or investment strategies without requiring full retraining.
- Applications in banking could include risk assessment for loans, dynamic credit scoring based on evolving financial behaviour, or fraud detection where expert knowledge continuously refines the detection criteria.
- Investment and trading models could benefit from real time expert inputs, allowing traders to update decision-making rules as new market trends emerge.

The insurance industry relies on complex decision-making processes where premiums, coverage, and claims approvals are determined by a range of risk factors. RDR-based systems could provide a transparent, explainable, and adaptable solution for managing insurance policies.

- Future studies could investigate how RDR can be used to develop KBS models that dynamically adjust insurance pricing based on evolving risk factors such as climate change, economic shifts, or new actuarial research.
- Claims processing and fraud detection could benefit from RDR's incremental learning capability, where new fraud patterns or emerging risks can be integrated as soon as they are identified by domain experts.

Beyond finance and insurance, RDR-based KBS models could be applied in a range of fields where domain expertise and evolving knowledge are critical:

- **Legal and regulatory compliance**: Laws and policies frequently change, requiring legal firms and compliance officers to continuously update their knowledge. RDR-based systems could assist in maintaining up-to-date rule-based decision support for legal compliance, contract analysis, or case law research.
- **Cybersecurity and threat detection**: The constantly evolving nature of cyber threats requires an adaptable approach to security rule enforcement. Future research could explore how RDR could be used for incrementally updating security policies based on newly identified vulnerabilities and attack patterns.
- **Environmental and sustainability monitoring**: Predicting climate-related risks and environmental sustainability indicators requires an evolving set of decision rules. RDR-based systems could assist in monitoring and responding to changes in pollution levels, deforestation rates, or renewable energy efficiency metrics.

To ensure the successful adoption of RDR in these industries, future research should investigate:

1. **Scalability Challenges**: How RDR models can be optimised for big data environments while maintaining interpretability.
2. **Integration with Machine Learning:** Hybrid approaches where RDR and ML techniques complement each other for enhanced predictive accuracy.
3. **Domain-Specific Adaptations:** Customising RDR frameworks to accommodate industry-specific challenges, such as regulatory compliance in finance or real-time threat detection in cybersecurity.

The methodology and findings of this thesis have clear practical applications across multiple domains. In healthcare, the developed RDR-based KBS can inform risk stratification and targeted screening for chronic conditions beyond diabetes, such as cardiovascular disease or hypertension. In the financial and insurance sectors, the system's adaptability supports more dynamic risk assessment and personalized policy adjustment, while in the legal field it could facilitate compliance monitoring and regulatory decision support. For policymakers, these capabilities enable more precise resource allocation and the design of data-driven public health interventions. The demonstrated scalability and flexibility of the approach ensure that it remains relevant to evolving real-world challenges, strengthening the practical and societal contributions of this research.

7.7.2 Hybrid Models Combining RDRs and ML

While RDRs demonstrate clear advantages, future studies could explore hybrid approaches that combine RDRs' interpretability with ML's pattern recognition capabilities. Such hybrid models could leverage ML's predictive power while retaining the adaptability and explainability of RDRs.

Future research could further investigate how explainable AI techniques, such as model-agnostic interpretability methods, can be integrated with RDR-ML hybrid systems to enhance transparency while maintaining predictive accuracy.

The recent surge in research on hybrid and explainable systems further highlights the importance of combining expert-driven and data-driven methods. While state-of-the-art ML models continue to advance, the challenges of data scarcity, transparency, and adaptability remain critical in healthcare applications (Smith et al., 2023; Holzinger et al., 2019). The findings of this thesis are well-aligned with these developments, reinforcing the need for KBS that can both incorporate expert knowledge and adapt to new data in real time.

7.8 Final Remarks

This research has shown that expert-driven knowledge systems, such as those built using RDRs, offer a powerful alternative to traditional machine learning—particularly in settings where data is limited, fragmented, or evolving. By enabling real-time updates and embedding domain expertise, RDR-based systems promote transparency and adaptability in healthcare prediction. These qualities position RDRs not only as a methodological contribution, but as a practical solution for developing intelligent systems that serve complex real-world needs. Beyond healthcare, the adaptability and incremental nature of RDRs present exciting opportunities for broader industry applications. As demonstrated in this research, the same methodology used for diabetes risk prediction can be extended to other fields requiring dynamic, expert-driven decision support. The ability to evolve dynamically in response to new information makes RDRs an invaluable tool for real-time decision making.

The insights gained from this study contribute to both academic research and real-world applications. By leveraging the strengths of RDR-based KBS models, this research paves the way for more dynamic, interpretable, and inclusive decision support systems. These findings underscore the importance of continued exploration into knowledge-based AI applications,

driving advancements in chronic disease prevention, financial forecasting, and other expert-driven fields.

References

Agardh, E., Allebeck, P., Hallqvist, J., Moradi, T., & Sidorchuk, A. 2019. Type 2 diabetes incidence and socio-economic position: a systematic review and meta-analysis. *Diabetologia*, 54(5), 960-965.

Alrige, M., Banjar, H., Shuaib, T., Ahmed, A. & Gharbawi, R. 2023, 'Knowledge-based dietary intake recommendations of nutrients for pediatric patients with Maple Syrup Urine Disease', *Healthcare*, 11, p. 301.

American Diabetes Association (2018), 'Economic costs of diabetes in the US in 2017', *Diabetes Care*, 41(5), pp. 917-928.

An, Q., Rahman, S., Zhou, J. & Kang, J.J. 2023, 'A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges', *Sensors*, 23(9), p. 4178.

Anjana, R.M., Unnikrishnan, R., Deepa, M., Pradeepa, R., Tandon, N., Das, A.K., Joshi, S., Bajaj, S., Jabbar, P.K. & Das, H.K. 2023, 'Metabolic non-communicable disease health report of India: the ICMR-INDIAB national cross-sectional study (ICMR-INDIAB-17)', *The Lancet Diabetes & Endocrinology*, vol. 11, no. 7, pp. 474-89.

Anjana, R. M., Deepa, M., Pradeepa, R., Mahanta, J., Narain, K., Das, H. K., & Kaur, T. 2017. Prevalence of diabetes and prediabetes in 15 states of India: results from the ICMR-INDIAB population-based cross-sectional study. *The Lancet Diabetes & Endocrinology*, 5(8), 585-596.

Association, A.D. 2018, 'Economic costs of diabetes in the US in 2017', Diabetes Care, vol. 41, no. 5, pp. 917-28.

Atun, R., Davies, J. I., Gale, E. A. M., & Boulton, A. J. M. 2017. Diabetes in sub-Saharan Africa: From clinical care to health policy. *The Lancet Diabetes & Endocrinology*, 5(8), 622-624.

Australian Institute of Health and Welfare (AIHW) 2024, *Diabetes: Australian facts – Health system expenditure*, AIHW, Canberra, viewed 11 August 2025, https://www.aihw.gov.au/reports/diabetes/diabetes/contents/impact-of-diabetes/health-system-expenditure.

Australian Institute of Health and Welfare (2016) Health expenditure Australia 2014-15, Canberra, 57. Viewed 18 February 2019, https://www.aihw.gov.au/getmedia/a13427b8-d5de-495d-8b8f-4fd114f135d0/20279.pdf.aspx?inline=true%5C%3E

Basu, S., Flood, D., Geldsetzer, P., Theilmann, M., Marcus, M. E., Ebert, C., Mayige, M., Wong-Mcclure, R., Farzadfar, F., Saeedi Moghaddam, S., Agoudavi, K., Norov, B., Houehanou, C., Andall-Brereton, G., Gurung, M., Brian, G., Bovet, P., Martins, J., Atun,

R., Bärnighausen, T., Vollmer, S., Manne-Goehler, J. & Davies, J. 2021. Estimated effect of increased diagnosis, treatment, and control of diabetes and its associated cardiovascular risk factors among low-income and middle-income countries: a microsimulation model. *Lancet Glob Health,* 9**,** e1539-e1552.

Benke, C., Autenrieth, L.K., Asselmann, E. & Pané-Farré, C.A. 2020, 'Lockdown, quarantine measures, and social distancing: Associations with depression, anxiety and distress at the beginning of the COVID-19 pandemic among adults from Germany', *Psychiatry Research*, 293, p. 113462.

Beydoun, G. & Hoffmann, A. 2013, 'Dynamic evaluation of the development process of knowledge-based information systems', *Knowledge and Information Systems*, 35(1), pp. 233-247.

Bindoff, I., Curtain, C., Peterson, G., Westbury, J. & Ling, T. 2014, 'Problems detected by a ripple-down rules-based medication review decision support system: are they relevant?', *Knowledge Management and Acquisition for Smart Systems and Services: 13th Pacific Rim Knowledge Acquisition Workshop (PKAW 2014)*, Springer, pp. 59-68.

Bindoff, I.K., Peterson, G.M. & Curtain, C. 2014, 'Computer System to Support Medication Reviews: a good but not new concept', International Journal of Clinical Pharmacy, vol. 36, no. 2, p. 218.

Bull, F.C., Al-Ansari, S.S., Biddle, S., Borodulin, K., Buman, M.P., Cardon, G., Carty, C., Chaput, J.-P., Chastin, S. & Chou, R. 2020, 'World Health Organization 2020 Guidelines on Physical Activity and Sedentary Behaviour', *British Journal of Sports Medicine*, 54(24), pp. 1451-1462.

Compton, P. & Kang, B.H. 2021, *Ripple-Down Rules: The Alternative to Machine Learning*, CRC Press.

Compton, P., & Kang, B. 2014. Ripple-Down Rules: Acquisition and Maintenance of Knowledge for Clinical Decision Support. Journal of Clinical Decision Support Systems, 5(1), 23-34.

Compton, P., Kim, Y.S. & Kang, B.H. 2014, 'Linked Production Rules: controlling inference with knowledge', Knowledge Management and Acquisition for Smart Systems and Services, pp. 84-98.

Compton, P., RDR Example, School of Computer Science & Engineering, University of New South Wales, Sydney, viewed 20 March 2017, <http://www.cse.unsw.edu.au/~cs9416/06s1/lectures/rdr/rdr_trace.html>.

Compton, P. & Richards, D. 2000, 'Generalising Ripple-Down rules', in R. Dieng & O. Corby (eds), 12th International Conference on Knowledge Engineering and Knowledge Management, EKAW 2000, Conference Paper, vol. 1937, Springer Verlag, pp. 380-6, viewed 2 October 2017, <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84958770061&partnerID=40&md5=d52491e82a94ffe41f9041706aabc411>.

Cooper, Z.W., Mowbray, O. & Johnson, L. 2024, 'Social determinants of health and diabetes: using a nationally representative sample to determine which SDoH model best predicts diabetes risk', *Clinical Diabetes and Endocrinology*, vol. 10, art. 4, viewed 11 August 2025,

https://clindiabetesendo.biomedcentral.com/articles/10.1186/s40842-023-00162-5. BioMed Central

Cutler, D.M. & Lleras-Muney, A. 2010, 'Understanding Differences in Health Behaviours by education', *Journal of Health Economics*, 29(1), pp. 1-28.

Diabetes around the world in 2021 2024, International Diabetes Federation, Brussels, Belgium, viewed 16 August 2024, <https://idf.org/about-diabetes/diabetes-facts-figures/>.

Diabetes in Australia: 2023 Snapshot 2023, Diabetes Australia, Canberra, ACT, viewed 19 September 2023, <https://www.diabetesaustralia.com.au/wp-content/uploads/2023-Snapshot-Diabetes-in-Australia.pdf/>

Diabetes in Australia (2015), *Diabetes Australia*, Canberra, viewed 17/08/2017, <https://www.diabetesaustralia.com.au/about-diabetes/diabetes-in-australia/\>.

Džeroski, S., De Raedt, L. & Driessens, K. 2001, 'Relational Reinforcement Learning', *Machine Learning*, vol. 43, pp. 7-52.

Eyerich, K., Brown, S.J., White, B.E.P., Tanaka, R.J., Bissonette, R., Dhar, S., Bieber, T., Hijnen, D.J., Guttman-Yassky, E. & Irvine, A. 2019, 'Human and Computational Models of Atopic Dermatitis: A Review and Perspectives by an Expert Panel of the International Eczema Council', *Journal of Allergy and Clinical Immunology*, 143(1), pp. 36-45.

Financial Report 2020–2021 (2021), Diabetes Australia, Canberra, ACT, viewed 20 October 2021, <https://www.diabetesaustralia.com.au/wp-content/uploads/211018-Diabetes-Australia-Limited-FS-30-June-2021-FINAL.pdf>

Firima, E., Gonzalez, L., Ursprung, F., Robinson, E., Huber, J., Belus, J.M., Raeber, F., Gupta, R., Deen, G.F. & Amstutz, A. 2023, 'Community-based models of care for management of type 2 diabetes mellitus among non-pregnant adults in sub-Saharan Africa: A scoping review', *Plos one*, vol. 18, no. 11, p. e0278353.

Fleiss, J.L. (2011), *Design and Analysis of Clinical Experiments*, John Wiley & Sons.

Fregoso-Aparicio, L., Noguez, J., Montesinos, L. & García-García, J.A. 2021, *'Machine learning and Deep Learning Predictive Models for Type 2 Diabetes: a Systematic Review', Diabetology & Metabolic Syndrome*, vol. 13, no. 1, pp. 1-22.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y. & Ranganath, R. 2020. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings,* 2020**,** 191.

Greenes, R. A., Bates, D. W., Kawamoto, K., Middleton, B., Osheroff, J. & Shahar, Y. 2018. Clinical decision support models and frameworks: seeking to address research issues underlying implementation successes and failures. *Journal of biomedical informatics,* 78**,** 134-143.

Habehh, H. & Gohel, S. 2021. Machine learning in healthcare. *Current genomics,* 22**,** 291.

Hamedan, F., Orooji, A., Sanadgol, H. & Sheikhtaheri, A. 2020. Clinical decision support system to predict chronic kidney disease: A fuzzy expert system approach. *International journal of medical informatics,* 138**,** 104134.

Han, S. C., Mirowski, L. & Kang, B. H. 2015. Exploring a role for MCRDR in enhancing telehealth diagnostics. *Multimedia Tools and Applications,* 74**,** 8467-8481.

Hill-Briggs, F. & Fitzpatrick, S.L. 2023, 'Overview of social determinants of health in the development of diabetes', *Diabetes care*, vol. 46, no. 9, pp. 1590-8.

Hill, J., Nielsen, M. & Fox, M. H. 2013. Understanding the Social Factors That Contribute to Diabetes: A Means to Informing Health Care and Social Policies for the Chronically Ill. *The Permanente Journal,* 17**,** 67 -72.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K. & Müller, H. 2019, 'Causability and Explainability of Artificial Intelligence in Medicine', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312.

Hurtubise, K., Rivard, L., Héguy, L., Berbari, J. & Camden, C. 2016. Virtual knowledge brokering: describing the roles and strategies used by knowledge brokers in a pediatric physiotherapy virtual community of practice. *Journal of Continuing Education in the Health Professions,* 36**,** 186-194.

Hussain, M., Aboalsamh, H., & Khan, A. 2021. Addressing the Social Determinants of Health Through Health Systems Strengthening. *Public Health Reviews*, 42, 210-223.

Hussain, M., Satti, F. A., Ali, S. I., Hussain, J., Ali, T., Kim, H.-S., Yoon, K.-H., Chung, T. & Lee, S. 2021. Intelligent knowledge consolidation: from data to wisdom. *Knowledge-Based Systems,* 234**,** 107578.

Hyeon, J., Oh, K.-J., Kim, Y. J., Chung, H., Kang, B. H. & Choi, H.-J. Constructing an initial knowledge base for medical domain expert system using induct RDR. Big Data and Smart Computing (BigComp), 2016 International Conference on, 2016. IEEE, 408-410.

International Diabetes Federation (IDF) 2024, *Diabetes around the world in 2021* [Online]. Brussels , Belgium:. Available: https://idf.org/about-diabetes/diabetes-facts-figures/ Viewed 16 August 2024.

International Diabetes Federation (IDF) 2021, *IDF Diabetes Atlas*, 10th edn, IDF, Brussels, Available: https://diabetesatlas.org/media/uploads/sites/3/2025/02/IDF_Atlas_10th_Edition_2021.pdf. [Accessed 11 August 2025]

Kantor, M., Wright, A., Burton, M., Fraser, G., Krall, M., Maviglia, S., Mohammed-Rajput, N., Simonaitis, L., Sonnenberg, F. & Middleton, B. 2011, 'Comparison of Computer-based Clinical Decision Support Systems and Content for Diabetes Mellitus', Applied Clinical Informatics.

Khamisi, Y. N. A., Khan, M. K. & Munive-Hernandez, J. E. 2019. The design of a knowledge-based system for quality management in healthcare: case study. *International Journal of Advanced Operations Management,* 11**,** 257-274.

Kivuyo, S., Birungi, J., Okebe, J., Wang, D., Ramaiya, K., Ainan, S., Tumuhairwe, F., Ouma, S., Namakoola, I. & Garrib, A. 2023, 'Integrated management of HIV, diabetes, and hypertension in sub-Saharan Africa (INTE-AFRICA): a pragmatic cluster-randomised, controlled trial', *The Lancet*, vol. 402, no. 10409, pp. 1241-50.

Kunutsor, S. K. & Laukkanen, J. A. 2017. Gamma-glutamyltransferase and risk of chronic kidney disease: a prospective cohort study. *Clinica chimica acta,* 473**,** 39-44.

Lee, C., Cohort, R. & Magliano, D. 2013. The Cost of Diabetes in adults in Australia. *Diabetes Research and Clinical Practice***,** Vol. 99 Issue 3 pages 385 - 390.

Lee, C. M. Y., Colagiuri, R., Magliano, D. J., & Shaw, J. E. 2019. The Cost of Diabetes in Australia: The Impact of Diabetes Complications on the Healthcare System. *Medical Journal of Australia*, 210(4), 180-186.

Lee, C. M. Y., Goode, B., Nørtoft, E., Shaw, J. E., Magliano, D. J. & Colagiuri, S. 2018. The cost of diabetes and obesity in Australia. *Journal of medical economics,* 21**,** 1001-1005.

Lee, S. K., Khambhati, J., Varghese, T., Stahl, E. P., Kumar, S., Sandesara, P. B., Wenger, N. K. & Sperling, L. S. 2017. Comprehensive primary prevention of cardiovascular disease in women. *Clinical Cardiology,* 40**,** 832-838.

Lipton, Z. C. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue,* 16**,** 31-57.

Liu, X., Zhang, L. & Chen, W. 2023. Trends in economic burden of type 2 diabetes in China: Based on longitudinal claim data. *Frontiers in Public Health,* 11.

Macciotta, A., Sacerdote, C., Giachino, C., Di Girolamo, C., Franco, M., van der Schouw, Y.T., Zamora-Ros, R., Weiderpass, E., Domenighetti, C. & Elbaz, A. 2025, 'Examining causal relationships between educational attainment and type 2 diabetes using genetic analysis: findings from the EPIC-InterAct study through Mendelian randomisation', *J Epidemiol Community Health*, vol. 79, no. 5, pp. 373-9.

Marmot, M. 2005, 'Social Determinants of Health Inequalities', *The Lancet*, 365(9464), pp. 1099-1104.

McCombe, G., Murtagh, S., Lazarus, J.V., Van Hout, M.C., Bachmann, M., Jaffar, S., Garrib, A., Ramaiya, K., Sewankambo, N.K. & Mfinanga, S. 2021, 'Integrating diabetes, hypertension and HIV care in sub-Saharan Africa: a Delphi consensus study on international best practice', *BMC Health Services Research*, vol. 21, no. 1, p. 1235.

Morgan, A. J., Wright, J. & Reavley, N. J. 2021. Review of Australian initiatives to reduce stigma towards people with complex mental illness: what exists and what works? *International Journal of Mental Health Systems,* 15**,** 1-51.

Mutyambizi, C., Chola, L., & Manne-Goehler, J. 2018. The economic Impact of Diabetes in Sub-Saharan Africa: A Systematic Review. *The Lancet Diabetes & Endocrinology*, 6(12), 908-919.

Mutyambizi, C., Pavlova, M., Chola, L., Hongoro, C. & Groot, W. 2018. Cost of diabetes mellitus in Africa: a systematic review of existing literature. *Globalization and health,* 14**,** 1-13.

Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science,* 366**,** 447-453.

Omar, A., Beydoun, G., Win, K., Shukla, N. & Baker, G. 2019. Socio-Technical Perspective on Managing Type II Diabetes. *ACIS2019: 30th Annual Conference on Information Systems.* Perth, Australia, Decemeber 9-11, 2019.

Omar, A., Beydoun, G., Win, K. T. & Jelinek, H. 2022. The Incremental Development of a Diabetes 2 Knowledge Base System using Ripple Down Rules. *Pacific Asia Conference on Information Systems, July 5-9 2022.* Taipei/Sydney Virtual Conference.

Parker, E. D., Lin, J., Mahoney, T., Ume, N., Yang, G., Gabbay, R. A., Elsayed, N. A. & Bannuru, R. R. 2023. Economic Costs of Diabetes in the U.S. in 2022. *Diabetes Care,* 47**,** 26-43.

Peffers, K., Tuunanen, T., Rothenberger, M.A. & Chatterjee, S. 2007, 'A Design Science Research Methodology for Information Systems Research', *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45-77.

Pickett, K.E. & Wilkinson, R.G. 2015, 'Income Inequality and Health: A Causal Review', *Social Science & Medicine*, 128, pp. 316-326.

Pratama, B. O. R., Nugroho, A. C. & Chandrawati, T. B. Expert System for Initial Diagnosis of Covid-19 Using the Forward Chaining Method and Ripple Down Rules. 2023 7th International Conference on Information Technology (InCIT), 2023. IEEE, 141-146.

Psyllidis, A. Ontology-based data integration from heterogeneous urban systems: A knowledge representation framework for smart cities. Proceedings of the 14th International Conference on Computers in Urban Planning and Urban Management (CUPUM'14), 2015.

Ramaprasad, A., Win, K. T., Syn, T., Beydoun, G. & Dawson, L. 2016. Australia's National Health Programs: An Ontological Mapping.

Rani, P., Kumar, R., Ahmed, N. M. S. & Jain, A. 2021. A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments,* 7**,** 263-275.

Reid, R.A., Mavoa, S., Foster, S., Gilmartin-Thomas, J. & Rachele, J.N. 2024, 'Spatially consistent annual socio-economic indexes for areas (SEIFA) data for australian statistical areas, 1996–2021'.

Sauer, K. S., Jungmann, S. M. & Witthöft, M. 2020. Emotional and behavioral consequences of the COVID-19 pandemic: The role of health anxiety, intolerance of uncertainty, and distress (in) tolerance. *International journal of environmental research and public health,* 17**,** 7241.

Sauliune, S. & Kalediene, R. 2015. Health Profile of the Urban Community Members in Lithuania: Do Socio-Demographic Factors Matter? *International Journal of Epidemiology,* 44**,** i81-i81.

Schwerdtle, P. 2016. Prevalence, Distribution And Impact. *Australian Nursing and Midwifery Journal,* 23.

Shaw, M. 2004. Housing and public health. *Annu. Rev. Public Health,* 25**,** 397-418.

Sim, L.L.W., Ban, K.H.K., Tan, T.W., Sethi, S.K., & Loh, T.P. (2017). Development of a clinical decision support system for diabetes care: A pilot study. *PLOS ONE*, 12.

Smith, L.A., Oakden-Rayner, L., Bird, A., Zeng, M., To, M.-S., Mukherjee, S. & Palmer, L.J. 2023, 'Machine learning and deep learning predictive models for long-term prognosis in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis', *The Lancet Digital Health*, vol. 5, no. 12, pp. e872-e81.

Smith, C., Mcnaughton, D. & Meyer, S. 2016. Client perceptions of group education in the management of type 2 diabetes mellitus in South Australia. *Australian Journal of Primary Health*.

Steele, C.J., Schöttker, B., Marshall, A.H., Kouvonen, A., O'Doherty, M.G., Mons, U., Saum, K.-U., Boffetta, P., Trichopoulou, A. & Brenner, H. 2017, 'Education achievement and type 2 diabetes—what mediates the relationship in older adults? Data from the ESTHER study: a population-based cohort study', *BMJ open*, vol. 7, no. 4, p. e013569.

Stringhini, S., Carmeli, C., Jokela, M., Avendaño, M., Muennig, P., Guida, F., Ricceri, F., D'errico, A., Barros, H., Bochud, M., Chadeau-Hyam, M., Clavel-Chapelon, F., Costa, G., Delpierre, C., Fraga, S., Goldberg, M., Giles, G. G., Krogh, V., Kelly-Irving, M., Layte, R., Lasserre, A. M., Marmot, M. G., Preisig, M., Shipley, M. J., Vollenweider, P., Zins, M., Kawachi, I., Steptoe, A., Mackenbach, J. P., Vineis, P., Kivimäki, M., Alenius, H., Avendano, M., Bochud, M., Carmeli, C., Carra, L., Castagné, R., Chadeau-Hyam, M., Clavel-Chapelon, F., Costa, G., Courtin, E., Delpierre, C., D'errico, A., Dugué, P. A., Elliott, P., Fraga, S., Gares, V., Giles, G., Goldberg, M., Greco, D., Hodge, A., Irving, M. K., Karisola, P., Kivimäki, M., Krogh, V., Lang, T., Layte, R., Lepage, B., Mackenbach, J., Marmot, M., Mccrory, C., Milne, R., Muennig, P., Nusselder, W., Panico, S., Petrovic, D., Polidoro, S., Preisig, M., Raitakari, O., Ribeiro, A. I., Ribeiro, A. I., Ricceri, F., Robinson, O., Valverde, J. R., Sacerdote, C., Satolli, R., Severi, G., Shipley, M. J., Stringhini, S., Tumino, R., Vineis, P., Vollenweider, P. & Zins, M. 2017. Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1·7 million men and women. *The Lancet,* 389**,** 1229-1237.

Stringhini, S., Tabak, A.G., Akbaraly, T.N., Sabia, S., Shipley, M.J., Marmot, M.G., Brunner, E.J., Batty, G.D., Bovet, P. & Kivimäki, M. 2012, 'Contribution of modifiable risk factors to social inequalities in type 2 diabetes: prospective Whitehall II cohort study', *Bmj*, vol. 345.

Taylor, L. 2019, *Interview with A. Omar*, Director, Centre for Epidemiology & Evidence, NSW Health, Sydney, 15th October 2019

Van Hout, M.-C., Bachmann, M., Lazarus, J.V., Shayo, E.H., Bukenya, D., Picchio, C.A., Nyirenda, M., Mfinanga, S.G., Birungi, J. & Okebe, J. 2020, 'Strengthening integration of

chronic care in Africa: protocol for the qualitative process evaluation of integrated HIV, diabetes and hypertension care in a cluster randomised controlled trial in Tanzania and Uganda', *BMJ open*, vol. 10, no. 10, p. e039237.

Varghese, J.S., Anjana, R.M., Geldsetzer, P., Sudharsanan, N., Manne-Goehler, J., Thirumurthy, H., Bhattacharyya, S., Narayan, K.V., Mohan, V. & Tandon, N. 2023, 'National estimates of the adult diabetes care continuum in India, 2019-2021', *JAMA Internal Medicine*, vol. 183, no. 9, pp. 963-72.

Walsh, S., De Jong, E. E., Van Timmeren, J. E., Ibrahim, A., Compter, I., Peerlings, J., Sanduleanu, S., Refaee, T., Keek, S. & Larue, R. T. 2019. Decision support systems in oncology. *JCO clinical cancer informatics,* 3**,** 1-9.

WEKA Team n.d., 'Class J48', *WEKA 3.8/3.9 API Documentation*, viewed 11 August 2025, https://weka.sourceforge.io/doc.stable/weka/classifiers/trees/J48.html. Weka n.d., 'Class J48', *WEKA 3.8/3.9 API Documentation*, viewed 11 August 2025, https://weka.sourceforge.io/doc.stable/weka/classifiers/trees/J48.html. Weka

Williams, R., Karuranga, S., Malanda, B., Saeedi, P., Basit, A., Besançon, S., Bommer, C., Esteghamati, A., Ogurtsova, K. & Zhang, P. 2020. Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes research and clinical practice,* 162**,** 108072.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J. & Foulds, J. 2025, *Data Mining: Practical Machine Learning Tools and Techniques*, 5th edn, Morgan Kaufmann, Cambridge, MA. (See overview: https://ml.cms.waikato.ac.nz/weka/book.html.) Waikato Machine Learning

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J. & Foulds, J. 2023, *Data Mining: Practical Machine Learning Tools and Techniques*, 5th edn, Morgan Kaufmann, Cambridge, MA.

Wong, V. 2016, *Interview with A. Omar*, Head of Diabetes, Liverpool Hospital, Sydney, 16th June 2016

World Health Organization (WHO) (2025), *Commission on Social Determinants of health, 2005-2208*, World Health Organisation, Geneva, viewed 11 August 2025 (who.int). https://www.who.int/teams/social-determinants-of-health/equity-and-health/world-report-on-social-determinants-of-health-equity/commission-on-social-determinants-of-health

World Health Organization (WHO) (2024), *Guidance on global monitoring for diabetes prevention and control*, WHO, Geneva, viewed 11 August 2025, https://www.who.int/publications/m/item/guidance-on-global-monitoring-for-diabetes-prevention-and-control.

Xu, Y., Wang, L., He, J., Bi, Y., Li, M., Wang, T., Wang, L., Jiang, Y., Dai, M. & Lu, J. 2013, 'Prevalence and Control of Diabetes in Chinese Adults', Jama, vol. 310, no. 9, pp. 948-59.

Yan, H., Zheng, J., Jiang, Y., Peng, C. & Li, Q. 2003, 'Development of a Decision Support System for Heart Disease Diagnosis Using Multilayer Perceptron', *2003 IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 5, IEEE, pp. V-V.

Appendices

Appendix A

Rule construction table for geographic training KBS using the geographically specific dataset (100% accuracy)

Case No	Attribute(s) Used	Rule	Rule Conclusion	Cumulative Accuracy (%)	Order Added	RDR No	If True Go To	If False Go to
4	Pat. P Code	IF PPC = 2640	Not Diabetic	33	1	1	2	46
1151	Pat. St. Code	IF PSTC=3656	Diabetic	69.2	11	2	12	13
1169	Pat. Street & BMI	IF PS=9603 AND BMI>=30	Diabetic	70.4	16	3	17	24
1192	Gender & Pat. Street	IF Gender= F AND PS=1453	Diabetic	71.4	19	4	exit	20
1227	Pat. St	IF PS=6236	Diabetic	75.8	28	5	exit	29
1230	Pat. St	IF PS=2100	Diabetic	76	29	6	exit	30
1234	Pat. St	IF PS=3130	Diabetic	76.6	31	7	exit	32
1247	Pat. St & last visit to Dr	IF PS=9940 AND LVTDr>0	Diabetic	77.8	35	8	exit	Exit
1252	Pat. St & Gender	IF PS=6408 AND Gender=F	Diabetic	78	36	9	exit	37
840	Gender, Age & Pat. St	IF PS=6408 AND Gender=F AND Age>66	Not Diabetic	78.2	37	10	exit	38
1275	Pat. St	IF PS=9104	Diabetic	78.4	38	11	exit	39

2279	Age, Pat. St & Pat. P Code	IF Age<=61 AND PS=9104 AND PPC=2640	Not Diabetic	93.2	88	12	exit	89
1283	Pat. St	IF PS=9104	Diabetic	79	40	13	exit	41
675	Pat. St & GP St	IF PS=6440 AND GP St=2644	Not Diabetic	79.2	41	14	exit	42
1923	Age, Pat. St & Pat. P Code	IF Age>70 AND PS=6440 AND PPC=2640	Diabetic	88.2	70	15	exit	71
1307	Pat P Code, Age & BMI	IF PPC=2640 AND Age>73 AND BMI>28	Diabetic	80.2	46	16	47	68
952	Gender, Pat P Code, Age & PHQ9	IF Gender=F AND PPC=2640 AND Age<75 AND PHQ9=0	Not Diabetic	80.6	47	17	exit	48
1633	Pat. St	IF PS=6144	Not Diabetic	85.4	63	18	exit	64
1984	Fam. Hist. DM, Pat. St & Pat. P Code	IF FHDM=No AND PS=2120 AND PPC=2640	Not Diabetic	88.8	72	19	exit	73
2244	Gender, Pat. St & Pat. P Code	IF Gender=F AND PS=4342 AND PPC=2640	Not Diabetic	91.6	84	20	exit	85
2270	Gender, Pat. St & Pat. P Code	IF Gender=F AND PS=8592 AND PPC=2640	Not Diabetic	92.2	86	21	exit	exit
2288	Gender, Pat. St & Pat. P Code	IF Gender=F AND PS=6538 AND PPC=2640	Not Diabetic	93.4	89	22	exit	90
2601	Pat. St & Pat. P Code	IF PS=9940 AND PPC=2640	Not Diabetic	96.8	103	23	exit	104
1477	Pat. St	IF PS=8535	Diabetic	81.6	51	24	exit	52
1479	Pat. St & HT Stat.	IF PS=7006 AND HT=Yes	Diabetic	81.8	52	25	exit	53
1554	Pat. St	IF PS=6797	Diabetic	83	55	26	exit	56

1583	Pat. St & GP St	IF PS=8425 AND GPST=1453	Diabetic	85.2	62	27	exit	63
1716	Gender, Pat P St, Age & HT Stat	IF Gender=M AND PST=9004 AND Age=67 AND HTSTAT=Yes	Diabetic	87	65	28	exit	66
1830	Pat. St & Pat. P Code	IF PS=6052 AND PPC=2640	Diabetic	87.4	67	29	exit	exit
1868	Pat. St & Pat. P Code	IF PS=2609 AND PPC=2640	Diabetic	87.4	68	30	69	78
1881	Age, Pat. St & Pat. P Code	IF Age>70 AND PS=9808 AND PPC=2640	Diabetic	87.8	69	31	exit	70
2066	Pat. St & Pat. P Code	IF PS=6627 AND PPC=2640	Diabetic	89.8	77	32	exit	exit
1963	Pat. St, Pat. P Code & Alcohol	IF PS=6627 AND PPC=2640 AND Alch.=No	Not Diabetic	90	78	33	79	80
2068	Gender, Age, Pat P Code, CVD_stat, HT Stat, Alchol, Waist Circum. & BMI	If Gender=F AND Age>=69 AND PPC=2640 AND CVDSTAT=Yes AND HTSTAT=Yes AND Alch=Yes AND WST CIRM>=100 AND BMI>=26	Diabetic	90.2	79	34	exit	exit
2107	Pat. St & Pat. P Code	IF PS=5602 AND PPC=2640	Not Diabetic	90.6	81	35	exit	82
2145	Gender, Age, Pat. St & Pat. P Code	IF Gender=M AND Age=60 PS=9170 AND PPC=2640	Diabetic	90.8	82	36	exit	83
2150	Gender, Age, Pat. St & Pat. P Code	IF Gender=F AND Age>=72 PS=3616 AND PPC=2640	Diabetic	91.2	83	37	exit	84
2303	Age, Pat. St & Pat. P Code	IF Age>=65 AND PS=2598 AND PPC=2640	Diabetic	93.8	90	38	exit	91

2357	Gender, Pat. St & Pat. P Code	IF Gender=F AND PS=3877 AND PPC=2640	Diabetic	94.4	93	39	exit	exit
2379	Gender, Pat. St & Pat. P Code	IF Gender=F AND PS=7974 AND PPC=2640	Diabetic	94.6	94	40	exit	95
2440	Pat. St & Pat. P Code	IF PS=2120 AND PPC=2640	Diabetic	95.2	97	41	exit	98
2517	Gender, Age, Pat. St & Pat. P Code	IF Gender=F AND Age>=76 PS=6230 AND PPC=2640	Diabetic	96.4	101	42	exit	exit
2649	Age, Pat. St & Pat. P Code	IF Age=52 AND PPC=2640 AND HT_Stat= Yes AND BMI>29	Diabetic	97.2	104	43	exit	105
2664	HT_Stat, Age, GP P Code, Fam Hist DM & Diet	IF HT_Stat=Yes AND Age>=59 GPPC=2641 AND FamHist_DM=Yes AND Diet=No AND BMI>27	Diabetic	97.8	107	44	exit	exit
2700	Gender, Pat. St & Pat. P Code	IF Gender=M AND PS=7745 AND PPC=2640	Diabetic	99.8	113	45	exit	114
6	Pat. P Code	IF PPC = 2641	Not Diabetic	48	2	46	exit	3
1168	Pat. Town	IF PTwn=Norris Park, Lavington	Diabetic	70.2	15	47	exit	exit
1186	Pat. Town	IF PTwn=Norris Park	Diabetic	71	18	48	exit	19
1193	Withdrn. Scrn & Hist DM	IF PWTHSCN=Yes AND HISTDM=Yes	Diabetic	74.4	22	49	exit	23
968	Pat. Str.	IF PS=8201	Not Diabetic	74.6	23	50	exit	exit
1282	Pat. St & Cardio Vasc.Stat	IF PS=6939 AND CVD=Yes	Diabetic	78.8	39	51	exit	40
1286	Pat. St & GP St	IF PS=3219 AND GP St=4342	Diabetic	79.4	42	52	exit	43

1487	Pat. St	IF PS=3241	Diabetic	82.2	53	53	exit	54
1494	Pat. St	IF PS=6797	Diabetic	82.6	54	54	exit	55
1556	Pat. St	IF PS=1154	Diabetic	83.6	56	55	exit	57
1568	Pat. St	IF PS=5169	Diabetic	84	58	56	exit	59
1570	Pat. St	IF PS=5169	Diabetic	84.2	59	57	exit	60
1793	Gender, Age & Pat. Street	IF Gender=M AND PST=2491 AND	Diabetic	87.2	66	58	exit	67
1947	Pat. St & Pat. P Code	IF PS=1813 AND PPC=2641	Diabetic	88.6	71	59	exit	72
2012	Gender, Age, Pat. St & Pat. P Code	IF Gender=M AND Age=76 PS=6773 AND PPC=2641	Diabetic	89.2	74	60	exit	75
2017	Pat. St & Pat. P Code	IF PS=6773 AND PPC=2641	Diabetic	89.4	75	61	exit	76
2247	Gender, Alcohol, Pat. St & Pat. P Code	IF Gender=F AND Alchol=Yes PS=1453 AND PPC=2641	Diabetic	91.8	85	62	86	87
2317	Pat. St & Pat. P Code	IF PS=6758 AND PPC=2641	Diabetic	94	91	63	exit	92
2452	Gender, Pat. St, Pat. P Code & HT Stat	IF Gender=F AND PS=7099 AND PPC=2641 AND HTStat=Yes	Diabetic	95.8	98	64	exit	exit
2448	Gender, Pat. St & Pat. P Code	IF Gender=M AND PS=5378 AND PPC=2641	Diabetic	96	99	65	exit	100
2674	Gender, Pat. St & Pat. P Code	IF Gender=F AND PS=2878 AND PPC=2641	Diabetic	99	110	66	exit	exit

2731	Gender, Pat. St & Pat. P Code	IF Gender=F AND PS=5138 AND PPC=2641	Diabetic	100	114	67	exit	exit
16	Age	IF Age< = 55	Not Diabetic	54.2	3	68	exit	4
1134	Pat. St. Code	IF PSTC=5980	Diabetic	68.6	9	69	10	11
1139	Pat. St. Code	IF PSTC=3468	Diabetic	69	10	70	exit	exit
1155	Pat. St. Code	IF PSTC=7614	Diabetic	69.4	12	71	exit	exit
1157	Pat. Post Code	IF PPC=2644	Diabetic	69.6	13	72	14	16
1160	Pat. Post Code	IF PPC=3677	Diabetic	70	14	73	15	exit
1172	Pat. Street	IF PS=3960	Diabetic	70.6	17	74	exit	18
1472	Gender & Pat. Street	IF Gender=F AND PST=3867	Diabetic	80.8	48	75	exit	49
1473	Pat. St	IF PS=4792	Diabetic	81	49	76	50	51
2651	Gender, Pat. St & Pat. P Code	IF Gender=M AND PS=5196 AND HT_Stat=Yes	Diabetic	97.4	105	77	106	107
390	Pat. P Code	IF PPC = 2642	Not Diabetic	54.8	4	78	exit	5
1217	Age & History DM	IF Age>=63 AND HistDM=Yes	Diabetic	75	25	79	exit	26
416	Pat. P Code	IF PPC = 3690	Not Diabetic	64	5	80	exit	6
1732	PHQ9	IF PHQ9>0	Diabetic	74.2	21	81	exit	22

1216	Gender, Age & Pat. St	IF Gender = F AND Age>=60 AND Pat St= 5755	Diabetic	75.2	26	82	exit	27
1233	Pat. St	IF PS=3509	Diabetic	76.4	30	83	exit	31
1245	Pat. St	IF PS=3130	Diabetic	77.4	34	84	35	36
1303	Pat. St	IF PS=8014	Diabetic	79.8	44	85	exit	45
941	Pat. St & Age	IF PS=8014 AND Age<=75	Not Diabetic	80	45	86	exit	exit
1476	Pat. St	IF PS=3336	Diabetic	81.2	50	87	exit	exit
1562	Pat. St & Pat. P Code	IF PS=2028 AND PPC=3690	Diabetic	83.8	57	88	exit	58
1573	Pat. St	IF PS=7322	Diabetic	84.8	60	89	exit	61
1580	Pat. St	IF PS=7322	Diabetic	85	61	90	exit	62
1994	Pat. St & Pat. P Code	IF PS=2210 AND PPC=3690	Diabetic	89	73	91	exit	74
2078	Withdrawn Screening, Gender Age, & Pat. P Code	IF WthdrnScrn=Unknown AND Gender=F AND Age>=79 AND PPC=2640	Diabetic	89.6	76	92	exit	77
2336	Gender, Pat. St & Pat. P Code	IF Gender=F AND PS=5627 AND PPC=2640	Not Diabetic	94.2	92	93	93	94
2098	Gender, Pat. St & Pat. P Code	IF Gender=F AND PS=4399 AND PPC=3690	Diabetic	90.4	80	94	81	99
2277	Pat. St & Pat. P Code	IF PS=7620 AND PPC=3690	Diabetic	92.8	87	95	exit	88

2663	Gender, Pat. St & HT_Stat	IF Gender=F AND PS=5196 AND HT_Stat=Yes	Diabetic	97.6	106	96	exit	exit
2668	Pat. St & Pat. P Code	IF PS=8792 AND PPC=3690	Diabetic	98.6	109	97	110	111
2701	Gender, Pat. St & Pat. P Code	IF Gender=F AND PS=1304 AND PPC=3690	Diabetic	99.4	111	98	112	113
420	Pat. P Code	IF PPC = 3688	Not Diabetic	64.4	6	99	exit	7
486	Pat. P Code	IF PPC = 2660	Not Diabetic	64.8	7	100	exit	8
2510	Age, HT Stat & BMI	If Age>=70 AND HT_Stat=Yes AND BMI>23	Diabetic	96.2	100	101	101	102
503	HT_Status	IF HT_Status=No	Not Diabetic	68.4	8	102	exit	9
1197	GP Street	IF GP St=8872	Diabetic	74.8	24	103	exit	25
1225	Pat P Code	IF PPC= 2611	Diabetic	75.4	27	104	exit	28
1240	Pat P Code	IF PPC= 3700	Diabetic	76.8	32	105	33	34
1042	Gender & Pat. St	IF Gender=M AND PS=3867	Not Diabetic	77	33	106	exit	exit
1289	Pat. St & BMI	IF PS=7745 AND BMI>=24.5	Diabetic	79.6	43	107	exit	44
1423	PHQ9	IF PHQ9=0	Not Diabetic	74	20	108	exit	21
1662	Pat. P Code	IF PPC=3700	Diabetic	86.8	64	109	exit	65
2583	HT_Stat, Age, GP P Code, Alcohol & Diet	IF HT_Stat=Yes AND Age<=67 GPS=1453 AND Alchl=No AND Diet=No	Not Diabetic	96.6	102	110	103	108

2416	Pat. St & Pat. P Code	IF PS=5120 AND PPC=2659	Diabetic	94.8	95	111	exit	96
1778	Gender, Weight, Pat. St & Pat. P Code	IF Gender=F AND Wght<=90 PS=5120 AND PPC=2659	Diabetic	95	96	112	exit	97
2665	Age, Pat. St & Pat. P Code	IF Age<=61 AND PS=8344 PPC=2643	Diabetic	98.4	108	113	109	exit
2691	Gender, Pat. St & Pat. P Code	IF Gender=F AND PS=7745 AND PPC=3749	Diabetic	99.6	112	114	exit	exit

Appendix B

Rule construction table for socio-demographic KBS (90.2% accuracy) using the socio-demographic dataset.

Case No	No. of Attributes	Attribute(s) Used	Rule	Rule Conclusion	Cumulative Accuracy (%)	Rule No	RDR No	If True Go To	If False Go To
4	1	Fam Hist. DM	IF FHDM = No	Not Diabetic	46.6	1	1	2	44
1134	3	Gender, Pat Age & Med. Age	IF Gender = F AND Pat. Age = 41 AND Med. Age = 42	Diabetic	71.4	6	2	exit	3
1142	2	Med Age & CVD Status	IF Med Age = 37 AND CVD_Stat = Yes	Diabetic	72.6	8	3	4	6
1105	1	Pat. Age	IF Pat. Age = 84	Not Diabetic	72.8	9	4	exit	5
1511	2	Pat. Age & Med. Age	IF Pat. Age = 52 AND Med. Age = 37	Not Diabetic	84.2	61	5	exit	exit
1151	2	Withdrawn Screening & Pat. Age	IF Witdrn Scr = Death AND Pat. Age = 51	Diabetic	73	10	6	exit	7
1168	10	Times Attended, Gender, Pat. Age, Med. Age, Med. Fam Income, Med. Mortgage, Med. Rent, Avg. Houshold Size, Last Visit GP & BMI	IF Times Attended = 3 AND Gender = F AND Pat. Age = 54 AND Med. Age = 40 AND Med. Fam. Income = $ 5056 AND Med. Mortg. = $1300 AND Med.	Diabetic	74.2	15	7	exit	8

			Rent = $860 AND Avg. Household Size = 2.2 AND Last Visit GP = 1 AND BMI = 22.5						
1169	2	Gender & BMI	IF Gender = F AND BMI > 30	Diabetic	74	16	8	9	24
103	1	Times Attended	IF Times attended = 2	Not Diabetic	74.4	17	9	10	12
1216	1	Pat. Age	IF Pat. Age = 63	Diabetic	78.8	35	10	exit	11
1561	2	Pat. Age & Med. Age	IF Pat. Age = 60 AND Med. Age = 40	Diabetic	84.8	64	11	exit	exit
372	3	Withdrawn screening, Pat. Age & Highest Education (Male)	IF Withdrawn Screening = Unknown AND Pat. Age = 55 AND Highest Edu. (male) = Cert IV	Not Diabetic	74.4	18	12	exit	13
416	1	Withdrawn screening	IF Withdrawn Screening = Death	Not Diabetic	74.6	19	13	exit	14
503	1	Med. Rent	IF Med. Rent = 780	Not Diabetic	74.8	20	14	exit	15
612	1	PHQ9	IF PHQ9 = 7	Not Diabetic	75.2	21	15	exit	16
642	2	Pat. Age & Highest Edu. (Female)	IF Pat. Age AND Highest Edu. (Female) = Cert IV	Not Diabetic	75.4	22	16	exit	17

837	4	Med. Age, Med. Fam. Income, Med. Mortg. Payments & Med. Rent	IF Med. Age = 40 AND Med. Fam. Income = $5056 AND Med. Fam. Mortg = $1300 AND Med. Rent = $860	Not Diabetic	76	23	17	18	19
1568	2	Pat. Age & Times Attended	IF Pat. Age = 60 AND Times Attended = 3	Diabetic	85	65	18	exit	exit
885	1	BMI	IF BMI = 30.85	Not Diabetic	76.2	24	19	exit	20
963	1	Pat. Age	IF Pat. Age = 74	Not Diabetic	76.4	25	20	exit	21
1778	3	Pat. Age, Med. Age & Times Attended	IF Pat. Age = 83 AND Med. Age = 42 ANDTimes Attended = 1	Not Diabetic	87.6	76	21	exit	22
1891	2	Pat. Age & Med. Age	IF Pat. Age > 65 AND Med. Age = 51	Not Diabetic	88.8	83	22	exit	23
1961	3	Pat. Age, Times Attend & Diet	IF Pat. Age > 60 AND Times Attend = 1 AND Diet = Yes	Not Diabetic	89.6	87	23	exit	exit
1233	5	Pat. Age, Times Attended, Med. Age, Med. Fam. Income, Med. Mortg.	IF Pat. Age = 64 AND Times Attended = 3 AND Med. Age = 37 AND	Diabetic	79.4	38	24	exit	25

		Payments & Med. Rent	Med. Fam. Income = $5924 AND Med. Mortg. Payments = $1387 AND Med. Rent = $1387						
1234	3	Gender, Pat. Age, & BMI	IF Gender = F AND Pat. Age = 64 AND BMI = 27.2	Diabetic	79.6	39	25	exit	26
1245	4	Gender, Pat. Age, Med. Age & Exc. Int	IF Gender = F AND Pat. Age = 65 AND Med. Age = 37 AND Excs. Int = Moderate	Diabetic	80	41	26	27	28
689	1	Times Attended	IF Times Attended = 6	Not Diabetic	80.2	42	27	exit	exit
1257	4	Times Attended, Gender, Pat. Age, Med. Age & Med. Age	IF Times Attended = 1 AND Gender = F AND Pat. Age = 67 AND Med. Age =	Diabetic	80.4	43	28	exit	29
1289	2	Pat. Age, & Med. Age	IF Pat. Age = 73 & Med. Age = 48	Diabetic	81.2	47	29	exit	30
1305	3	Times Attended, Pat. Age, & Med. Age	IF Times Attended = 7 AND Pat. Age = 76 AND Med. Age = 40	Diabetic	81.8	50	30	exit	31
1329	3	Pat. Age. Gender & BMI	IF Pat. Age = 85 AND Gender =	Diabetic	82.6	53	31	exit	32

			F AND BMI > 28						
1473	3	Gender, Pat. Age & Med. Age	IF Gender = F AND Pat. Age = 49 AND Med. Age = 37	Diabetic	83	55	32	exit	33
1477	3	Pat. Age. Gender & BMI	IF Pat. Age = 57 AND Gender = F AND BMI > 27	Diabetic	83.4	57	33	exit	34
1487	3	Times Attended, Pat. Age, & Med. Age	IF Times Attended = 4 AND Pat. Age = 64 AND Med. Age = 40	Diabetic	83.8	59	34	exit	35
1580	2	Pat. Age & Med. Age	IF Pat. Age = 79 AND Med. Age = 37	Diabetic	85.4	67	35	exit	36
1583	2	Pat. Age & Gender	IF Pat. Age > 89 AND Gender = F	Diabetic	85.6	68	36	exit	37
1711	3	Pat. Age, Med. Age & HT Status	IF Pat. Age = 63 AND Med. Age = 40 AND HT_Status = Yes	Diabetic	86.6	71	37	exit	38
1727	2	Pat. Age & Med. Age	IF Pat. Age = 74AND Med. Age = 48	Diabetic	87.2	74	38	exit	39
1730	2	Pat. Age & Times Attended	IF Pat. Age = 77 AND Times Attended = 1	Diabetic	87.4	75	39	exit	40
1840	3	Pat. Age, Med. Age & Times Attended	IF Pat. Age > 75 AND Med. Age = 48 AND Times	Diabetic	88.2	79	40	exit	41

			Attended =>1						
1881	2	Pat. Age & Times Attended	IF Pat. Age = 77 AND Times Attended > 3	Not Diabetic	88.4	81	41	42	43
1354	2	Fam. Hist. & Exc. Duration	Fam. Hist. = Yes AND Exc. Duration = 2	Not Diabetic	88.6	82	42	exit	exit
1947	5	Pat. Age, Med. Age, Times Attended, Withdrawn Screening & Gender	IF Pat. Age = 61 AND Med. Age = 40 AND Times Attended = 1 AND Withdrawn Screening = Yes AND Gender = F	Diabetic	89.4	86	43	exit	exit
12	1	Gender	IF Gender = M	Not Diabetic	58	2	44	45	62
1139	4	Pat. Age, Med Age, Med. Fam Income & Med. Mortg Payments	IF Pat. Age = 47 Med Age = 37 AND Med Fam. Income = $5924 AND Med Mortg. Payments = $1387	Diabetic	71.6	7	45	exit	46
1160	3	Withdrawn Screening, Gender & Rent (Monthly)	IF Withdrawn Screening = Unknown AND Gender = M AND Rent = $880	Diabetic	74	14	46	exit	47

1186	2	Times Attended & Med. Age.	IF Times Attended = 4 AND Med. Age = 40	Diabetic	76.8	27	47	exit	48
1217	2	Pat. Age & Weight	IF Pat Age = 63 AND Weight = 90	Diabetic	79	36	48	exit	49
1125	2	Pat. Age & Med. Age	IF Pat Age = 64 AND Med. Age = 38	Diabetic	79.2	37	49	exit	50
1275	2	Times Attended, & Pat. Age	IF Times Attended = 1 AND Pat. Age = 70	Diabetic	80.6	44	50	exit	51
1286	2	Pat. Age, & Med. Age	IF Pat. Age = 73 & Med. Age = 40	Diabetic	81	46	51	exit	52
1295	2	Pat. Age, & Med. Age	IF Pat. Age = 73 & Med. Age = 48	Diabetic	81.4	48	52	exit	53
1303	2	Pat. Age, & Med. Age	IF Pat. Age = 76 & Med. Age = 37	Diabetic	81.6	49	53	exit	54
1307	2	Pat. Age, & Med. Age	IF Pat. Age = 77 AND Med. Age = 39	Diabetic	82	51	54	exit	55
1398	2	Times Attended AND Pat. Age	IF Times Attended = 5 AND Pat. Age = 74	Diabetic	82.8	54	55	exit	56
1479	2	Times Attended AND Pat. Age	IF Times Attended = 4 AND Pat. Age = 59	Diabetic	83.6	58	56	exit	57
1560	3	Withdrawn Screening, Pat. Age & HT_Status	IF Withdrawn Screening = not interesting AND Pat. Age = 58 AND	Diabetic	84.4	62	57	exit	58

			HT_Status = Yes						
1716	4	Pat. Age, CVD Status, HT Status & Exc. Int.	IF Pat. Age = 67 AND CVD Status = Yes AND HT_Status = Yes AND Exc. Int = Mod	Diabetic	86.8	72	58	exit	59
1717	2	Pat. Age & Times Attended	IF Pat. Age = 68 AND Times Attended = 5	Diabetic	87	73	59	exit	60
1793	2	Pat. Age & Med. Age	IF Pat. Age = 75 AND Med. Age = 40	Diabetic	87.8	77	60	exit	61
1830	2	Pat. Age & Exc. Int.	IF Pat. Age = 58 AND Exc. Int = Mod	Diabetic	88	78	61	exit	exit
35	1	Pat. Age	IF Pat Age<60	Not Diabetic	63.4	3	62	63	70
1155	3	Pat. Age & Exercise Inst.	IF Apt. Age =52 AND Med Age = 37 AND Exc. Int. = None	Diabetic	73.2	11	63	exit	64
1157	2	Gender & Withdrawn Screening	IF Gender = F AND Withdrawn Scr. = Yes	Diabetic	73.4	12	64	65	66
448	1	Med Rent (Monthly)	If Med. Rent = $1000	Not Diabetic	73.6	13	65	exit	exit
1172	1	Med. Age	IF Med. Age = 38	Diabetic	76.6	26	66	exit	67
1192	2	Pat. Age & Weight	If Pat. Age = 59 AND Weight = 65	Diabetic	77	28	67	exit	68
1476	2	Pat. Age & Med. Age	IF Pat. Age = 56 AND Med. Age = 37	Diabetic	83.2	56	68	exit	69

1868	3	Pat. Age, HT_Stat & Diet	IF Pat. Age > 46 AND HT_Stat = Yes AND Diet = Yes	Diabetic	88.4	80	69	exit	exit
558	1	CVD_Status	IF CVD_Status = No	Not Diabetic	69.6	4	70	71	88
1209	2	Med. Age CVD_Stat	IF Med. Age = 39 AND CVD_Stat = No	Diabetic	77.6	30	71	72	77
582	1	Pat. Age	IF Pat. Age = 62	Not Diabetic	78	31	72	exit	73
675	2	Pat. Age & Times Attended	IF Pat. Age = 65 AND Times Attended =2	Not Diabetic	78.2	32	73	exit	74
728	2	Pat. Age & Exc. Int	IF Pat. Age = 67 AND Exc. Int. = moderate	Not Diabetic	78.4	33	74	exit	75
810	2	Pat. Age & Times Attended	IF Pat. Age = 69 AND Times Attended =1	Not Diabetic	78.6	34	75	exit	76
1897	2	Pat. Age & Times Attended	IF Pat. Age < 71 AND Times Attended = 5	Not Diabetic	89	84	76	exit	exit
1244	3	Withdrawn Screening, Pat. Age, & Med. Age	IF Withdrawn Screening = Yes AND Pat. Age = 65 & Med. Age = 40	Diabetic	79.8	40	77	exit	78
1315	1	Pat. Age	IF Pat. Age = 80	Diabetic	82.4	52	78	exit	79
1494	2	Pat. Age & Med. Age	IF Pat. Age = 68 AND	Diabetic	84	60	79	exit	80

			Med. Age = 40						
1561	2	Pat. Age & Med. Age	IF Pat. Age = 60 AND Med. Age = 40	Diabetic	84.6	63	80	exit	81
1570	2	Pat. Age & Med. Age	IF Pat. Age = 67 AND Med. Age = 40	Diabetic	85.2	66	81	exit	82
1646	3	Pat. Age, Med. Age & HT Status	IF Pat. Age > 60 AND Med. Age = 40 AND HT_Status = Yes	Diabetic	86.2	69	82	83	84
1899	2	Pat. Age & Times Attended	IF Pat. Age = 71 AND Times Attended = 1	Not Diabetic	89.2	85	83	exit	exit
1662	2	Pat. Age & Med. Age	IF Pat. Age => 80 AND Med. Age = 48	Diabetic	86.4	70	84	exit	85
	2	Pat. Age & Times Attended	IF Pat. Age > 64 AND Times Attended = 1	Not Diabetic	89.8	88	85	86	exit
674	5	Times Attended, Withdrawn Screening, HT Stat., Fam. Hist. CVD & Diet	IF Pat. Times Attended = 1 AND Withdrawn Screening = Unknown AND HT Stat. = No AND Fam. Hist. CVD = No AND Diet = No	Not Diabetic	90	89	86	exit	87
947	2	Pat. Age & Med. Age	IF Pat. Age > 65 AND	Not	90.2	90	87	exit	exit

			Med. Age = 51						
968	1	Gender	IF Gender = M	Not Diabetic	71.2	5	88	89	exit
1197	2	Withdrawn Screening & Med. Age.	IF Withdrawn Screening = 4 AND Med. Age = 44	Diabetic	77.2	29	89	exit	90
1282	3	Withdrawn Screening, Pat. Age, & Exc. Int	IF Withdrawn Screening = Unknown AND Pat. Age = 72 & Exc. Int = Low	Diabetic	80.8	45	90	exit	exit

Appendix C

Rule construction table for the training dataset using the socio-demographic dataset (100% accuracy).

Case No.	No. of Attributes	Attribute(s) Used	Rule	Rule Conclusion	Cumulative Accuracy (%)	Order Added	RDR No	If True Go To	If False Go to
4	1	Fam Hist. DM	IF FHDM = No	Not Diabetic	46.6	1	1	2	67
1134	3	Gender, Pat Age & Med. Age	IF Gender = F AND Pat. Age = 41 AND Med. Age = 42	Diabetic	71.4	6	2	exit	3
1142	2	Med Age & CVD Status	IF Med Age = 37 AND CVD_Stat = Yes	Diabetic	72.6	8	3	4	6
1105	1	Pat. Age	IF Pat. Age = 84	Not Diabetic	72.8	9	4	exit	5
1511	2	Pat. Age & Med. Age	IF Pat. Age = 52 AND Med. Age = 37	Not Diabetic	84.2	61	5	exit	exit

1151	2	Withdrawn Screening & Pat. Age	IF Witdrn Scr = Death AND Pat. Age = 51	Diabetic	73	10	6	exit	7
1168	10	Times Attended, Gender, Pat. Age, Med. Age, Med. Fam Income, Med. Mortgage, Med. Rent, Avg. Houshold Size, Last Visit GP & BMI	IF Times Attended = 3 AND Gender = F AND Pat. Age = 54 AND Med. Age = 40 AND Med. Fam. Income = $ 5056 AND Med. Mortg. = $1300 AND Med. Rent = $860 AND Avg. Household Size = 2.2 AND Last Visit GP = 1 AND BMI = 22.5	Diabetic	74.2	15	7	exit	8
1169	2	Gender & BMI	IF Gender = F AND BMI > 30	Diabetic	74	16	8	9	31
103	1	Times Attended	IF Times attended = 2	Not Diabetic	74.4	17	9	10	15
1216	1	Pat. Age	IF Pat. Age = 63	Diabetic	78.8	35	10	exit	11
1561	2	Pat. Age & Med. Age	IF Pat. Age = 60 AND Med. Age = 40	Diabetic	84.8	64	11	exit	12

2317	10	Pat Age Med Age Med tot income Med Mortage Med Rent Avg Household size Edu lvl Female HT Status Fam Hist CVD Exc Dur Exc Int	IF Pat Age=80 AND Med Age=40 AND Med tot income=5056 AND Med Mortage=1300 AND Med Rent Avg=860 AND Household size=2.2 AND Edu lvl Female=Cert IV AND HT Status=Yes AND Fam Hist CVD=Yes AND Exc Dur=7 AND Exc Int=Moderate	Diabetic	94.6	107	12	exit	13
2649	3	Hgt Wght BMI	IF Hgt=1.78 AND Wght=115 AND BMI=36.3	Diabetic	97.6	122	13	exit	14
2656	3	Pat Age HT BMI	IF Pat Age=54 AND HT Stat=Yes AND BMI=30.1194	Diabetic	98.2	125	14	exit	exit
372	3	Withdrawn screening, Pat. Age & Highest Education (Male)	IF Withdrawn Screening = Unknown AND Pat. Age = 55 AND Highest Edu. (male) = Cert IV	Not Diabetic	74.4	18	15	exit	16
416	1	Withdrawn screening	IF Withdrawn Screening = Death	Not Diabetic	74.6	19	16	exit	17
503	1	Med. Rent	IF Med. Rent = 780	Not Diabetic	74.8	20	17	exit	18
612	1	PHQ9	IF PHQ9 = 7	Not Diabetic	75.2	21	18	exit	19

642	2	Pat. Age & Highest Edu. (Female)	IF Pat. Age AND Highest Edu. (Female) = Cert IV	Not Diabetic	75.4	22	19	exit	20
837	4	Med. Age, Med. Fam. Income, Med. Mortg. Payments & Med. Rent	IF Med. Age = 40 AND Med. Fam. Income = $5056 AND Med. Fam. Mortg = $1300 AND Med. Rent = $860	Not Diabetic	76	23	20	21	22
1568	2	Pat. Age & Times Attended	IF Pat. Age = 60 AND Times Attended = 3	Diabetic	85	65	21	exit	exit
885	1	BMI	IF BMI = 30.85	Not Diabetic	76.2	24	22	exit	23
963	1	Pat. Age	IF Pat. Age = 74	Not Diabetic	76.4	25	23	exit	24
1778	3	Pat. Age, Med. Age & Times Attended	IF Pat. Age = 83 AND Med. Age = 42 ANDTimes Attended = 1	Not Diabetic	87.6	76	24	exit	25
1891	2	Pat. Age & Med. Age	IF Pat. Age > 65 AND Med. Age = 51	Not Diabetic	88.8	83	25	exit	26
1961	3	Pat. Age, Times Attend & Diet	IF Pat. Age > 60 AND Times Attend = 1 AND Diet = Yes	Not Diabetic	89.6	87	26	exit	27
2104	7	Times attd Pat Age HT Stat Alcohol PHQ9 Exc Dur Last Visit GP & BMI	IF Times attd=3 AND Pat Age=72 AND HT Stat=Yes AND Alcohol=Yes AND PHQ9=1 AND Exc Dur=6 AND Last Visit GP=1 & BMI=36	Not Diabetic	93	101	27	exit	28

2372	5	Times attd Pat Age HT Status Fam Hist CVD Stat Fam Hist CVD	IF Times attd=10 AND Pat Age=68 AND HT Status=Yes AND Fam Hist CVD=Yes AND Stat=Yes	Not Diabetic	95.4	111	28	exit	29
2583	8	Med Age Med tot income Med Mortage Med Rent Avg Household size Hi Edu Lvl Female CVD Stat HT Status	IF Med Age=48 AND Med tot income=5548 AND Med Mortage=1083 AND Med Rent=760 AND Avg Household size=2.3 AND Hi Edu Lvl Female=Cert IV AND CVD Stat=Yes AND HT Status=Yes	Not Diabetic	96.4	116	29	exit	30
2617	4	Times Attd CVD Stat HT Status Exc Int	IF Times Attd=6 AND CVD Stat=Yes HT Stat=Yes AND Exc Int=Moderate	Not Diabetic	97.4	121	30	exit	exit
1233	5	Pat. Age, Times Attended, Med. Age, Med. Fam. Income, Med. Mortg. Payments & Med. Rent	IF Pat. Age = 64 AND Times Attended = 3 AND Med. Age = 37 AND Med. Fam. Income = $5924 AND Med. Mortg. Payments = $1387 AND Med. Rent = $1387	Diabetic	79.4	38	31	exit	32
1234	3	Gender, Pat. Age, & BMI	IF Gender = F AND Pat. Age = 64 AND BMI = 27.2	Diabetic	79.6	39	32	exit	33

1245	4	Gender, Pat. Age, Med. Age & Exc. Int	IF Gender = F AND Pat. Age = 65 AND Med. Age = 37 AND Excs. Int = Moderate	Diabetic	80	41	33	34	35
689	1	Times Attended	IF Times Attended = 6	Not Diabetic	80.2	42	34	exit	exit
1257	4	Times Attended, Gender, Pat. Age, Med. Age & Med. Age	IF Times Attended = 1 AND Gender = F AND Pat. Age = 67 AND Med. Age =	Diabetic	80.4	43	35	exit	36
1289	2	Pat. Age, & Med. Age	IF Pat. Age = 73 & Med. Age = 48	Diabetic	81.2	47	36	exit	37
1305	3	Times Attended, Pat. Age, & Med. Age	IF Times Attended = 7 AND Pat. Age = 76 AND Med. Age = 40	Diabetic	81.8	50	37	exit	38
1329	3	Pat. Age. Gender & BMI	IF Pat. Age = 85 AND Gender = F AND BMI > 28	Diabetic	82.6	53	38	exit	39
1473	3	Gender, Pat. Age & Med. Age	IF Gender = F AND Pat. Age = 49 AND Med. Age = 37	Diabetic	83	55	39	exit	40
1477	3	Pat. Age. Gender & BMI	IF Pat. Age = 57 AND Gender = F AND BMI > 27	Diabetic	83.4	57	40	exit	41
1487	3	Times Attended, Pat. Age, & Med. Age	IF Times Attended = 4 AND Pat. Age = 64 AND Med. Age = 40	Diabetic	83.8	59	41	exit	42
1580	2	Pat. Age & Med. Age	IF Pat. Age = 79 AND Med. Age = 37	Diabetic	85.4	67	42	exit	43

1583	2	Pat. Age & Gender	IF Pat. Age > 89 AND Gender = F	Diabetic	85.6	68	43	exit	44
1711	3	Pat. Age, Med. Age & HT Status	IF Pat. Age = 63 AND Med. Age = 40 AND HT_Status = Yes	Diabetic	86.6	71	44	exit	45
1727	2	Pat. Age & Med. Age	IF Pat. Age = 74AND Med. Age = 48	Diabetic	87.2	74	45	exit	46
1730	2	Pat. Age & Times Attended	IF Pat. Age = 77 AND Times Attended = 1	Diabetic	87.4	75	46	exit	47
1840	3	Pat. Age, Med. Age & Times Attended	IF Pat. Age > 75 AND Med. Age = 48 AND Times Attended =>1	Diabetic	88.2	79	47	exit	48
1881	2	Pat. Age & Times Attended	IF Pat. Age = 77 AND Times Attended > 3	Not Diabetic	88.4	81	48	49	50
1354	2	Fam. Hist. & Exc. Duration	Fam. Hist. = Yes AND Exc. Duration = 2	Not Diabetic	88.6	82	49	exit	exit
1947	5	Pat. Age, Med. Age, Times Attended, Withdrawn Screening & Gender	IF Pat. Age = 61 AND Med. Age = 40 AND Times Attended = 1 AND Withdrawn Screening = Yes AND Gender = F	Diabetic	89.4	86	50	exit	51
2012	2	Exc. INt & Fam Hist DM	IF Exc Int >=5 AND Fam Hist DM=No	Not Diabetic	91.2	93	51	52	56

2012	7	Times Atd., Gender, HT Stat., Alcohol, Fam Hist CVD, Diet & Fam Hist DM	IF Times Att = 6 AND Gender=M AND CVD Stat=No AND HT Stat=Yes AND Alcohol=Yes AND Fam Hist CVD=No AND Fam Hist DM=No AND Diet=No	Diabetic	91.8	96	52	exit	53
2014	10	Pat Age, Med Age, Med Fam Income, Med Mortg, Med Rent, Avg Household Size, High Educ Levl (Female), Diet, Exc Int & Last Vis to GP	If Pat Age=77 AND Med Age=39 AND Med Fam Income=$5252 AND Med Mortg=$1517 AND Med Rent=$1000 AND Avg Household Size= 2.4 AND High Educ Levl (Female)= Bach AND Diet=Yes AND Exc Int=Low AND Last Vis to GP=2	Diabetic	92	97	53	exit	54

2068	10	Times Attd Gender Med Age All socio, CVD Stat	If Gender=F AND Med Age=39 AND Med Fam Income=$5252 AND Med Mortg=$1517 AND Med Rent=$1000 AND Avg Household Size= 2.4 AND High Educ Levl (Female)= Bach AND CVD Stat=Yes	Diabetic	92.4	99	54	exit	55
2440	9	Times attd Pat Age Med Age Med tot income Med Mortage Med Rent Avg Household size Hi Edu Female Alcohol	IF Times attd=3 AND Pat Age=86 AND Med Age=39 AND Med tot income=5252 AND Med Mortage=1517 AND Med Rent=1000 AND Avg Household size=2.4 AND Hi Edu Female=Bachelor AND Alcohol=No	Diabetic	96	114	55	exit	exit

2145	8	Pat Age CVD Stat HT Stat Alcohol Exc Int Last vst GP Last vst Diab edu Freq diab edu	IF Pat Age=60 AND CVD Stat=Yes AND HT Stat=Yes AND Alcohol=Yes AND Exc Int=moderate AND Last vst GP=2 AND Last vst Diab edu=2 AND Freq diab edu=2	Diabetic	93.6	103	56	exit	57
2308	5	Times Attd Withdrawn Scr. Pat Age Edu lvl Female HT Stat	IF Times Attd=1 AND Withdrawn Scr=Yes AND Pat Age=68 AND Edu lvl Female=Cert IV AND HT Stat=Yes	Diabetic	94.4	106	57	exit	58
2357	4	Times Attd Gender CVD Stat HT Stat	IF Times Attd=3 AND Gender=F AND CVD Stat=Yes AND HT Stat=Yes	Diabetic	95	109	58	exit	59

2448	10	Times Attd Pat Age Med Age Med tot income Med Mortage Med Rent Avg Household size Hi Edu Lvl Female CVD Stat HT Status	IF Times Attd=3 AND Pat Age=55 AND Med Age=40 AND Med tot income=5056 AND Med Mortage=1300 AND Med Rent=860 AND Avg Household size=2.2 AND Hi Edu Lvl Female=Cert IV AND CVD Stat=Yes AND HT Status=Yes	Diabetic	96.2	115	59	exit	60
2468	7	Med Age Med tot income Med Mortage Med Rent Avg Hi Edu Lvl Female CVD Stat HT Status	IF Med Age=42 AND Med tot income=5360 AND Med Mortage=1000 AND Med Rent=600 AND Hi Edu Lvl Female=Cert IV AND HT Status=Yes AND CVD Stat=Yes	Diabetic	96.8	118	60	exit	61
2650	3	Times Attd Withdraw Scrn Pat Age	IF Times Attd=1 AND Withdraw Scrn=Yes AND Pat Age=52	Diabetic	97.8	123	61	exit	62

2668	7	Times Attd Pat Age Med Age Med fam Income Med Mortg Hi Edu Lvl (Female) HT Stat	IF Times Attd=3 AND Pat Age=60 AND Med Age=37 AND Med fam Income=5924 AND Med Mortg=1387 AND Hi Edu Lvl (Female)=Cert III & IV AND HT Stat=Yes	Diabetic	98.6	127	62	exit	63
2691	8	Times Attd Gender Med Age Med fam Income Med Mortg Med Rent Household Size HT Stat	IF Times Attd= 4 AND Gender=F AND Med Age=48 AND Med fam Income=6808 AND Med Mortg=1560 AND Med Rent=1080 AND Household Size=2.5 AND HT Stat=Yes	Diabetic	99	129	63	exit	64
2700	3	Times Attd Pat Age HT Stat	IF Times Attd=4 AND Pat Age=68 AND HT Stat=Yes	Diabetic	99.4	130	64	exit	65

2708	10	Times Attd Gender Pat Age Med Age Med fam Income Med Mortg Hi Edu Lvl (Female) HT Stat PHQ9 BMI	IF Times Attd=1 AND Gender=F AND Pat Age=71 AND Med Age=37 AND Med fam Income=5924 AND Med Mortg=1387 AND Hi Edu Lvl (Female)=Cert III & IV AND HT Stat=Yes AND PHQ9=0 AND BMI=25.2	Diabetic	99.6	131	65	exit	66
2731	12	Times Attd Gender Pat Age Med Age Med fam Income Med Mortg Med Rent Household Size Hi Edu Lvl (Female) CVD Stat HT Stat Exc Duration	IF Times Attd=4 AND Gender=F AND Pat Age=79 AND Med Age=40 AND Med fam Income=5056 AND Med Mortg=1300 AND Med Rent=860 AND HouseHld Size=2.2 AND Hi Edu Lvl (Female)=Cert III & IV AND CVD Stat=Yes AND HT Stat=Yes AND Exc Dur=0	Diabetic	99.8	132	66	exit	exit
12	1	Gender	IF Gender = M	Not Diabetic	58	2	67	68	92

1139	4	Pat. Age, Med Age, Med. Fam Income & Med. Mortg Payments	IF Pat. Age = 47 Med Age = 37 AND Med Fam. Income = $5924 AND Med Mortg. Payments = $1387	Diabetic	71.6	7	68	exit	69
1160	3	Withdrawn Screening, Gender & Rent (Monthly)	IF Withdrawn Screening = Unknown AND Gender = M AND Rent = $880	Diabetic	74	14	69	exit	70
1186	2	Times Attended & Med. Age.	IF Times Attended = 4 AND Med. Age = 40	Diabetic	76.8	27	70	exit	71
1217	2	Pat. Age & Weight	IF Pat Age = 63 AND Weight = 90	Diabetic	79	36	71	exit	72
1125	2	Pat. Age & Med. Age	IF Pat Age = 64 AND Med. Age = 38	Diabetic	79.2	37	72	exit	73
1275	2	Times Attended, & Pat. Age	IF Times Attended = 1 AND Pat. Age = 70	Diabetic	80.6	44	73	exit	74
1286	2	Pat. Age, & Med. Age	IF Pat. Age = 73 & Med. Age = 40	Diabetic	81	46	74	exit	75
1295	2	Pat. Age, & Med. Age	IF Pat. Age = 73 & Med. Age = 48	Diabetic	81.4	48	75	exit	76
1303	2	Pat. Age, & Med. Age	IF Pat. Age = 76 & Med. Age = 37	Diabetic	81.6	49	76	exit	77
1307	2	Pat. Age, & Med. Age	IF Pat. Age = 77 AND Med. Age = 39	Diabetic	82	51	77	exit	78
1398	2	Times Attended AND Pat. Age	IF Times Attended = 5 AND Pat. Age = 74	Diabetic	82.8	54	78	exit	79

1479	2	Times Attended AND Pat. Age	IF Times Attended = 4 AND Pat. Age = 59	Diabetic	83.6	58	79	exit	80
1560	3	Withdrawn Screening, Pat. Age & HT_Status	IF Withdrawn Screening = not interesting AND Pat. Age = 58 AND HT_Status = Yes	Diabetic	84.4	62	80	exit	81
1716	4	Pat. Age, CVD Status, HT Status & Exc. Int.	IF Pat. Age = 67 AND CVD Status = Yes AND HT_Status = Yes AND Exc. Int = Mod	Diabetic	86.8	72	81	exit	82
1717	2	Pat. Age & Times Attended	IF Pat. Age = 68 AND Times Attended = 5	Diabetic	87	73	82	exit	83
1793	2	Pat. Age & Med. Age	IF Pat. Age = 75 AND Med. Age = 40	Diabetic	87.8	77	83	exit	84
1830	2	Pat. Age & Exc. Int.	IF Pat. Age = 58 AND Exc. Int = Mod	Diabetic	88	78	84	exit	85
2670	1	BMI	IF BMI > 40	Diabetic	91	92	85	exit	86
2303	2	Patient Age & BMI	IF PA>45 AND BMI=27.4	Diabetic	91.4	94	86	exit	87
2011	7	Times Att., Gender, HT Stat., Alcohol, Fam Hist CVD, Diet & Ex Int.	IF Times Att = 6 AND Gender=M AND HT Stat=Yes AND Alcohol=Yes AND Fam Hist CVD=No AND Diet=No AND Exc Int=Moderate	Diabetic	91.6	95	87	exit	88

2047		Times Attd Pat Age HT Stat. Alcohol Exc Dur Last Vist GP & Freq GP	IF Times Attd=1 AND Pat Age=83 AND HT Stat=Yes AND Alcohol=Yes AND Exc Dur=5 AND Last Vist GP=1 AND Freq GP=12	Diabetic	92.2	98	88	exit	89
2510	10	Med Age Med tot income Med Mortage Med Rent Avg Hi Edu Lvl Female CVD Stat HT Status Diet	IF Med Age=42 AND Med tot income=5176 AND Med Mortage=979 AND Med Rent=720 AND Avg Household Size=2.5 Hi Edu Lvl Female=Cert IV AND HT Status=Yes AND CVD Stat=Yes AND Diet=Yes	Diabetic	97.2	120	89	exit	90
2651	3	Times Attd Withdraw Scrn HT Stat	IF Times Attd=2 AND Withdraw Scrn=Yes AND HT Stat=Yes	Diabetic	98	124	90	exit	91

2738	7	Times Attd Med Age Med fam Income Med Mortg Med Rent Household Size Hi Edu Lvl (Female)	IF Times Attd=1 AND Med Age=48 AND Med fam Income=5548 AND Med Mortg=1083 AND Med Rent=760 AND HouseHld Size=2.3 AND Hi Edu Lvl (Female)=Cert III & IV	Diabetic	100	133	91	exit	exit
35	1	Pat. Age	IF Pat Age<60	Not Diabetic	63.4	3	92	93	103
1155	3	Pat. Age & Exercise Inst.	IF Apt. Age =52 AND Med Age = 37 AND Exc. Int. = None	Diabetic	73.2	11	93	exit	94
1157	2	Gender & Withdrawn Screening	IF Gender = F AND Withdrawn Scr. = Yes	Diabetic	73.4	12	94	95	96
448	1	Med Rent (Monthly)	If Med. Rent = $1000	Not Diabetic	73.6	13	95	exit	exit
1172	1	Med. Age	IF Med. Age = 38	Diabetic	76.6	26	96	exit	97
1192	2	Pat. Age & Weight	If Pat. Age = 59 AND Weight = 65	Diabetic	77	28	97	exit	98
1476	2	Pat. Age & Med. Age	IF Pat. Age = 56 AND Med. Age = 37	Diabetic	83.2	56	98	exit	99
1868	3	Pat. Age, HT_Stat & Diet	IF Pat. Age > 46 AND HT_Stat = Yes AND Diet = Yes	Diabetic	88.4	80	99	exit	100
2665	1	BMI	IF BMI > 40	Diabetic	90.6	91	100	exit	101

2452	8	Times Attd Med Age Med tot income Med Mortage Med Rent Avg Household size Hi Edu Lvl Female CVD Stat HT Status	IF Times Attd=3 AND Med Age=40 AND Med tot income=5056 AND Med Mortage=1300 AND Med Rent=860 AND Avg Household size=2.2 AND Hi Edu Lvl Female=Cert IV AND HT Status=Yes	Diabetic	96.6	117	101	exit	102
2663	6	Times Attd Pat Age Med Age Med fam Income Med Mortg Hi Edu Lvl (Female)	IF Times Attd=2 AND Pat Age=58 AND Med Age=37 AND Med fam Income=5924 AND Med Mortg=1387 AND Hi Edu Lvl (Female)=Cert III & IV	Diabetic	98.4	126	102	exit	exit
558	1	CVD_Status	IF CVD_Status = No	Not Diabetic	69.6	4	103	104	127
1209	2	Med. Age CVD_Stat	IF Med. Age = 39 AND CVD_Stat = No	Diabetic	77.6	30	104	105	113
582	1	Pat. Age	IF Pat. Age = 62	Not Diabetic	78	31	105	exit	106
675	2	Pat. Age & Times Attended	IF Pat. Age = 65 AND Times Attended =2	Not Diabetic	78.2	32	106	exit	107
728	2	Pat. Age & Exc. Int	IF Pat. Age = 67 AND Exc. Int. = moderate	Not Diabetic	78.4	33	107	exit	108
810	2	Pat. Age & Times Attended	IF Pat. Age = 69 AND Times Attended =1	Not Diabetic	78.6	34	108	exit	109

1897	2	Pat. Age & Times Attended	IF Pat. Age < 71 AND Times Attended = 5	Not Diabetic	89	84	109	exit	110
2199	6	Times attd HT Stat Alcohol Exc Dur Exc Int Last vist GP	IF Times attd=7 AND HT Stat=Yes AND Alcohol=Yes Exc Dur=1 Exc Int=low AND Last vist GP=1	Not Diabetic	94	104	110	exit	111
2334	4	Times Attd Pat Age HT Stat BMI	IF Times Attd=3 AND Pat Age=60 AND HT Stat=Yes AND BMI=25.8	Not Diabetic	94.8	108	111	exit	112
1461	5	Times Attd Pat Age HT Stat Freq Diab Edu BMI	IF Times Attd=2 AND Pat Age=77 AND HT Stat=Yes AND Freq Diab Edu=100 AND BMI=26.6	Not Diabetic	97	119	112	exit	exit
1244	3	Withdrawn Screening, Pat. Age, & Med. Age	IF Withdrawn Screening = Yes AND Pat. Age = 65 & Med. Age = 40	Diabetic	79.8	40	113	exit	114
1315	1	Pat. Age	IF Pat. Age = 80	Diabetic	82.4	52	114	exit	115
1494	2	Pat. Age & Med. Age	IF Pat. Age = 68 AND Med. Age = 40	Diabetic	84	60	115	exit	116
1561	2	Pat. Age & Med. Age	IF Pat. Age = 60 AND Med. Age = 40	Diabetic	84.6	63	116	exit	117
1570	2	Pat. Age & Med. Age	IF Pat. Age = 67 AND Med. Age = 40	Diabetic	85.2	66	117	exit	118

1646	3	Pat. Age, Med. Age & HT Status	IF Pat. Age > 60 AND Med. Age = 40 AND HT_Status = Yes	Diabetic	86.2	69	118	119	121
1899	2	Pat. Age & Times Attended	IF Pat. Age = 71 AND Times Attended = 1	Not Diabetic	89.2	85	119	exit	120
2411	4	Times attd Withdrawn Scrn Gender Pat Age	IF Times attd=3 AND Withdrawn Scrn=Unknown AND Gender=F AND Pat Age=75	Not Diabetic	95.8	113	120	exit	exit
1662	2	Pat. Age & Med. Age	IF Pat. Age => 80 AND Med. Age = 48	Diabetic	86.4	70	121	exit	122
	2	Pat. Age & Times Attended	IF Pat. Age > 64 AND Times Attended = 1	Not Diabetic	89.8	88	122	123	125
674	5	Times Attended, Withdrawn Screening, HT Stat., Fam. Hist. CVD & Diet	IF Pat. Times Attended = 1 AND Withdrawn Screening = Unknown AND HT Stat. = No AND Fam. Hist. CVD = No AND Diet = No	Not Diabetic	90	89	123	exit	124
947	2	Pat. Age & Med. Age	IF Pat. Age > 65 AND Med. Age = 51	Not Diabetic	90.2	90	124	exit	exit

2305	7	Highest Edu lvl (Male) HT Stat Exc Dur Exc Int Last vst GP Freq Gp Last visit to diab Edu	IF Highest Edu lvl (Male)=Bachelor AND HT Stat=Yes AND Exc Dur=3 AND Exc Int=High AND Last vst GP=2 AND Freq GP=4 AND Last visit to diab Edu=5	Diabetic	94.2	105	125	exit	126
2687	6	Med Age Med fam Income Med Mortg Med Rent Household Size HT Stat	IF Med Age=48 AND Med fam Income=5548 AND Med Mortg=1083 AND Med Rent=760 AND Household Size=2.3 AND HT Stat=Yes	Diabetic	98.8	128	126	exit	exit
968	1	Gender	IF Gender = M	Not Diabetic	71.2	5	127	128	exit
1197	2	Withdrawn Screening & Med. Age.	IF Withdrawn Screening = 4 AND Med. Age = 44	Diabetic	77.2	29	128	exit	129
1282	3	Withdrawn Screening, Pat. Age, & Exc. Int	IF Withdrawn Screening = Unknown AND Pat. Age = 72 & Exc. Int = Low	Diabetic	80.8	45	129	exit	130

2077	8	All Socio Fam Hist CVD	If Med Age=39 AND Med Fam Income=$5252 AND Med Mortg=$1517 AND Med Rent=$1000 AND Avg Household Size= 2.4 AND High Educ Levl (Female)= Bach AND CVD Stat=Yes	Diabetic	92.6	100	130	exit	131
2117	8	Time attd withdrawn Scrn Pat Age HT Stat Alcohol Fam Hist CVD Exc Int Freq vst GP	IF Time attd= 5 AND withdrawn Scrn=Unknown AND Pat Age=79 AND HT Stat=Yes AND Alcohol=Yes AND Fam Hist CVD=Yes AND Exc Int=moderate AND Freq vst GP=4	Diabetic	93.4	102	131	exit	132

2367	10	Times attd Withdrawn Scrn Pat Age Med Age Med tot income Med Mortage Med Rent Avg Household size HT Status Fam Hist CVD Diet	If Times attd=3 AND Withdrawn Scrn=Unknown AND Pat Age=73 AND Med Age=37 AND Med tot income=5924 AND Med Mortage= 1387 AND Med Rent=1000 AND Avg Household size=2.4 AND HT Status=Yes AND Fam Hist CVD=Yes AND Diet=Yes	Diabetic	95.2	110	132	exit	133
2379	7	Med Age Med Tot income Med Mortage Med Rent Avg Household size Hi Edu lvl (Female) HT Status	IF Med Age=39 AND Med Tot income=5252 AND Med Mortage=1517 AND Med Rent Avg=1000 AND Household size=2.4 AND Hi Edu lvl (Female)=Bachelor AND HT Status=Yes	Diabetic	95.6	112	133	exit	exit

Appendix D

On the geographic KBS, this table shows the accuracy rate on the training dataset against the accuracy rate on the production unseen dataset.

Rule No	Training Cumulative Accuracy (%)	Production Cumulative Accuracy (%)
1	33	26.4
2	48	38.63
3	54.2	40.66
4	54.8	41.14
5	64	48.27
6	64.4	48.55
7	64.8	48.75
8	68.4	48.75
9	68.6	48.75
10	69	48.75
11	69.2	48.75
12	69.4	48.75
13	69.6	49.75
14	70	48.94
15	70.2	49.13
16	70.4	49.13
17	70.6	49.13
18	71	49.42
19	71.4	49.42
20	74	55.11
21	74.2	55.11
22	74.4	55.39
23	74.6	55.39
24	74.8	55.39
25	75	55.88
26	75.2	55.78
27	75.4	55.78
28	75.8	55.88
29	76	55.11
30	76.4	55.11
31	76.6	55.01
32	76.8	55.01
33	77	55.01
34	77.4	55.11
35	77.8	55.11
36	78	55.1
37	78.2	55.2
38	78.4	55.2
39	78.8	55.39
40	79	54.91
41	79.2	54.72
42	79.4	54.72

43	79.6	54.72
44	79.8	54.72
45	80	54.72
46	80.2	54.05
47	80.6	54.24
48	80.8	54.72
49	81	54.72
50	81.2	54.72
51	81.6	54.82
52	81.8	55.2
53	82.2	55.49
54	82.6	55.97
55	83	56.07
56	83.6	56.84
57	83.8	56.84
58	84	57.03
59	84.2	57.42
60	84.8	57.71
61	85	57.03
62	85.2	57.03
63	85.4	57.03
64	86.8	58.89
65	87	58.96
66	87.2	58.96
67	87.4	58.96
68	87.4	59.06
69	87.8	59.34
70	88.2	59.54
71	88.6	59.92
72	88.8	59.92
73	89	59.92
74	89.2	59.92
75	89.4	59.83
76	89.6	59.63
77	89.8	59.73
78	90	59.63
79	90.2	59.63
80	90.4	59.92
81	90.6	60.12
82	90.8	60.12
83	91.2	60.4
84	91.6	60.5
85	91.8	60.5
86	92.2	60.6
87	92.8	60.89
88	93.2	61.18
89	93.4	61.18
90	93.8	61.27
91	94	61.46
92	94.2	61.46
93	94.4	61.75
94	94.6	61.66
95	94.8	62.04

96	95	62.24
97	95.2	62.24
98	95.8	62.91
99	96	63.1
100	96.2	63.29
101	96.4	63.2
102	96.6	63.01
103	96.8	62.81
104	97.2	62.81
105	97.4	62.81
106	97.6	62.81
107	97.8	62.81
108	98.4	62.81
109	98.6	62.72
110	99	62.81
111	99.4	62.81
112	99.6	63.1
113	99.8	62.91
114	100	63.01

Appendix E

On the socio KBS, this table shows the accuracy rate on the training dataset against the accuracy rate on the production unseen dataset.

Rule No	Id	Training Dataset Cumulative Accuracy (%)	Production Dataset Cumulative Accuracy (%)
1	4	46.6	57.58
2	12	58	69.5
3	35	63.4	75.42
4	558	69.6	82.33
5	968	71.2	83.17
6	1134	71.4	83.17
7	1139	71.6	83.17
8	1142	72.6	83.17
9	1105	72.8	83
10	1151	73	83
11	1155	73.2	83
12	1157	73.4	82.58
13	448	73.6	82.83
14	1160	74	82.92
15	1168	74.2	82.92
16	1169	74	76.75
17	103	74.4	78.58
18	372	74.4	78.5
19	416	74.6	78.58
20	503	74.8	78.58
21	612	75.2	78.58
22	642	75.4	78.83
23	837	76	79.83

24	885	76.2	79.83
25	963	76.4	79.92
26	1172	76.6	79.75
27	1186	76.8	79.5
28	1192	77	79.5
29	1197	77.2	79.5
30	1209	77.6	78.33
31	582	78	78.5
32	675	78.2	78.5
33	728	78.4	78.5
34	810	78.6	78.5
35	1216	78.8	78.5
36	1217	79	78.5
37	1125	79.2	78.58
38	1233	79.4	78.58
39	1234	79.6	78.58
40	1244	79.8	78.58
41	1245	80	78.58
42	689	80.2	78.58
43	1257	80.4	78.42
44	1275	80.6	78.33
45	1282	80.8	78.33
46	1286	81	78.42
47	1289	81.2	78.5
48	1295	81.4	78.5
49	1303	81.6	78.5
50	1305	81.8	78.5
51	1307	82	78.42
52	1315	82.4	78.33
53	1329	82.6	78.33
54	1398	82.8	78.25
55	1473	83	78.17
56	1476	83.2	78.08
57	1477	83.4	78
58	1479	83.6	78
59	1487	83.8	77.92
60	1494	84	77.67
61	1511	84.2	77.67
62	1560	84.4	77.67
63	1561	84.6	77.75
64	1561	84.8	77.75
65	1568	85	77.75
66	1570	85.2	77.67
67	1580	85.4	77.67
68	1583	85.6	77.83
69	1646	86.2	77.83
70	1662	86.4	77.83
71	1711	86.6	77.92
72	1716	86.8	77.92
73	1717	87	77.83
74	1727	87.2	77.83
75	1730	87.4	77.33
76	1778	87.6	77.33

77	1793	87.8	77.33
78	1830	88	77.33
79	1840	88.2	77.25
80	1868	88.4	77.25
81	1881	88.4	77.17
82	1354	88.6	77.17
83	1891	88.8	77.17
84	1897	89	77.25
85	1899	89.2	77.25
86	1947	89.4	77.25
87	1961	89.6	77.25
88	1978	89.8	75.83
89	674	90	75.83
90	947	90.2	75.83
91	2665	90.6	76.17
92	2670	91	76.42
93	2012	91.2	76.42
94	2303	91.4	76.42
95	2011	91.6	76.42
96	2012	91.8	76.42
97	2014	92	76.42
98	2047	92.2	76.42
99	2068	92.4	76.42
100	2077	92.6	76.5
101	2104	93	76.5
102	2117	93.4	76.5
103	2145	93.6	76.5
104	2199	94	76.5
105	2305	94.2	76.5
106	2308	94.4	76.5
107	2317	94.6	76.5
108	2334	94.8	76.5
109	2357	95	76.58
110	2367	95.2	76.58
111	2372	95.4	76.58
112	2379	95.6	76.5
113	2411	95.8	76.5
114	2440	96	76.5
115	2448	96.2	76.5
116	2583	96.4	76.5
117	2452	96.6	76.5
118	2468	96.8	76.5
119	1461	97	76.58
120	2510	97.2	76.58
121	2617	97.4	76.58
122	2649	97.6	76.58
123	2650	97.8	76.58
124	2651	98	76.58
125	2656	98.2	76.58
126	2663	98.4	76.58
127	2668	98.6	76.58
128	2687	98.8	76.58
129	2691	99	76.58

130	2700	99.4	76.58
131	2708	99.6	76.58
132	2731	99.8	76.58
133	2738	100	76.67

Appendix F

This table shows the accuracy rate on the geo Vs socio on the production unseen datasets.

Rule No	Geo Production Cumulative Accuracy (%)	Socio Production Cumulative Accuracy (%)
1	26.4	57.58
2	38.63	69.5
3	40.66	75.42
4	41.14	82.33
5	48.27	83.17
6	48.55	83.17
7	48.75	83.17
8	48.75	83.17
9	48.75	83
10	48.75	83
11	48.75	83
12	48.75	82.58
13	49.75	82.83
14	48.94	82.92
15	49.13	82.92
16	49.13	76.75
17	49.13	78.58
18	49.42	78.5
19	49.42	78.58
20	55.11	78.58
21	55.11	78.58
22	55.39	78.83
23	55.39	79.83
24	55.39	79.83
25	55.88	79.92
26	55.78	79.75
27	55.78	79.5
28	55.88	79.5
29	55.11	79.5
30	55.11	78.33

31	55.01	78.5
32	55.01	78.5
33	55.01	78.5
34	55.11	78.5
35	55.11	78.5
36	55.1	78.5
37	55.2	78.58
38	55.2	78.58
39	55.39	78.58
40	54.91	78.58
41	54.72	78.58
42	54.72	78.58
43	54.72	78.42
44	54.72	78.33
45	54.72	78.33
46	54.05	78.42
47	54.24	78.5
48	54.72	78.5
49	54.72	78.5
50	54.72	78.5
51	54.82	78.42
52	55.2	78.33
53	55.49	78.33
54	55.97	78.25
55	56.07	78.17
56	56.84	78.08
57	56.84	78
58	57.03	78
59	57.42	77.92
60	57.71	77.67
61	57.03	77.67
62	57.03	77.67
63	57.03	77.75
64	58.89	77.75
65	58.96	77.75
66	58.96	77.67
67	58.96	77.67
68	59.06	77.83
69	59.34	77.83
70	59.54	77.83
71	59.92	77.92

72	59.92	77.92
73	59.92	77.83
74	59.92	77.83
75	59.83	77.33
76	59.63	77.33
77	59.73	77.33
78	59.63	77.33
79	59.63	77.25
80	59.92	77.25
81	60.12	77.17
82	60.12	77.17
83	60.4	77.17
84	60.5	77.25
85	60.5	77.25
86	60.6	77.25
87	60.89	77.25
88	61.18	75.83
89	61.18	75.83
90	61.27	75.83
91	61.46	76.17
92	61.46	76.42
93	61.75	76.42
94	61.66	76.42
95	62.04	76.42
96	62.24	76.42
97	62.24	76.42
98	62.91	76.42
99	63.1	76.42
100	63.29	76.5
101	63.2	76.5
102	63.01	76.5
103	62.81	76.5
104	62.81	76.5
105	62.81	76.5
106	62.81	76.5
107	62.81	76.5
108	62.81	76.5
109	62.72	76.58
110	62.81	76.58
111	62.81	76.58
112	63.1	76.5

113	62.91	76.5
114	63.01	76.5
115		76.5
116		76.5
117		76.5
118		76.5
119		76.58
120		76.58
121		76.58
122		76.58
123		76.58
124		76.58
125		76.58
126		76.58
127		76.58
128		76.58
129		76.58
130		76.58
131		76.58
132		76.58
133		76.67

Glossary of Terms

DSS – Decision Support System

GIS - Geographic Information System

KBS – Knowledge-Based System

RDR – Ripple-Down Rules

ML – Machine Learning

SDoH – Social Determinants of Health

SME – Subject Matter Expert

Copyright Statement