**Utilizing AI Expert Systems for Diabetes Diagnosis in Resource-Constrained Western Cape, Rural South Africa.**

**Researcher:**

**Craig Dun**

**Supervisor:**

**Dr Wilhelm Rothman**

**A thesis submitted in partial fulfilment of the PGDIP in Data Analysis requirements of IIE varsity college.**

**November 2023**

# Declaration

I hereby declare that the work presented in this thesis has not been submitted for any other degree or professional qualification, and that it is the result of my own independent work.

Craig Dun

Full Name Goes Here (Candidate)

2023/11/01

Date

# Abstract

Diabetes is a chronic disease that affects millions of people worldwide and poses significant challenges to healthcare systems. In resource-constrained areas such as the rural Western Cape in South Africa, the burden of diabetes diagnosis and management is exacerbated due to limited access to specialised healthcare professionals and diagnostic tools (Mayosi. et al 2012). This study explores the potential utilisation of artificial intelligence (AI) systems in improving diabetes diagnosis. AI expert systems are computer-based programs that mimic human decision-making processes by utilising algorithms and vast amounts of medical knowledge. These systems have demonstrated promising results in various medical fields, including diagnosing complex diseases like diabetes (Vayena and Blasimme (2018). By leveraging AI technology, it may be possible to bridge the gap between available resources and the increasing demand for accurate and timely diabetes diagnosis in resource-constrained settings.

A key component of this research involves developing machine learning models from a secondary dataset (Alex, 2021). These machine learning models will be trained using general human body information, devoid of specific regional data, allowing their application in diverse areas, including the rural Western Cape. By crafting machine learning models independent of contextual factors, their diagnostic accuracy will be solely determined by universally applicable human body data.

The utilisation of AI expert systems for diabetes diagnosis in the Western Cape, rural South Africa, offers a promising approach to bridge the healthcare gap in resource-constrained regions. By empowering healthcare providers with innovative tools, this initiative contributes to early diagnosis, improved patient outcomes, and ultimately, a reduction in the burden of diabetes on these communities (Vayena and Blasimme (2018).

# Acknowledgements

I am writing to express my heartfelt gratitude to Dr. Wilhelm Rothman and Dr. Deborah Ajumobi for your invaluable contributions and support throughout the research project. Your guidance, expertise, and unwavering assistance have played a pivotal role in the successful completion of this research endeavour.

I would also like to acknowledge the pivotal role played by my family and friends, who supported my project in various ways. This research would not have been possible without the combined efforts of all those involved. Your contributions have made a lasting impact on the quality and significance of my work.

Once again, thank you for all of your support.

# Table of contents

Contents

[● Declaration 1](#_Toc150282030)

[● Abstract 2](#_Toc150282031)

[● Acknowledgements 3](#_Toc150282032)

[● Table of contents 4](#_Toc150282033)

[● List of figures 6](#_Toc150282034)

[● List of tables 7](#_Toc150282035)

[Chapter 1: Introduction 1](#_Toc150282036)

[1.1 Background and aims. 1](#_Toc150282037)

[1.2 Problem statement 2](#_Toc150282038)

[1.3 Research objectives 3](#_Toc150282039)

[1.4 Research questions 3](#_Toc150282040)

[1.5 Relevance of research 3](#_Toc150282041)

[1.6 Thesis structure 4](#_Toc150282042)

[Chapter 2: Literature review 5](#_Toc150282043)

[2.1 Introduction 5](#_Toc150282044)

[2.2 Expert systems 5](#_Toc150282045)

[2.2.1 Diabetes in South Africa 7](#_Toc150282046)

[2.3 Conclusion 8](#_Toc150282047)

[Chapter 3: Methodology 9](#_Toc150282048)

[3.1 Introduction 9](#_Toc150282049)

[3.2 Research Design and Rationale 9](#_Toc150282050)

[3.3 Data Source 10](#_Toc150282051)

[3.4 Data Selection and sampling 11](#_Toc150282052)

[3.5 Data Splitting 12](#_Toc150282053)

[3.6 Machine Learning Models 12](#_Toc150282054)

[3.7 Ethics that were considered during the research. 13](#_Toc150282055)

[3.8 Conclusion 13](#_Toc150282056)

[Chapter 4: Empirical analysis and results 14](#_Toc150282057)

[4.1 Introduction 14](#_Toc150282058)

[4.2 Machine Learning Model Performance Analysis 14](#_Toc150282059)

[4.2.1 Performance Metrics 14](#_Toc150282060)

[4.2.2 Models evaluation 15](#_Toc150282061)

[4.2.3 Comparative Analysis 16](#_Toc150282062)

[4.3 Conclusion 17](#_Toc150282063)

[Chapter 5: Conclusion and recommendations 18](#_Toc150282064)

[5.1 Introduction 18](#_Toc150282065)

[5.2 Summary of Research 18](#_Toc150282066)

[5.3 Conclusion and Contributions 19](#_Toc150282067)

[5.4 Future research 19](#_Toc150282068)

[● References 21](#_Toc150282069)

[● Appendix A: Graphs 23](#_Toc150282070)

[● Appendix B: Tables 25](#_Toc150282071)

[● Appendix C: Code for the data analysis 26](#_Toc150282072)

# List of figures

*Figure 2 - Page 23*

*Figure 3 - Page 23*

*Figure 4 - Page 24*

*Figure 5 - Page 24*

# List of tables

*Table 1 - Page 25*

*Table 2 - Page 25*

*Table 3 - Page 25*

# Introduction

## Background and aims.

Diabetes, a persistent metabolic disorder, is characterised by elevated blood glucose levels and deficiencies in insulin production. It presents in two primary forms: type I, marked by insufficient pancreatic insulin production, and type II, wherein the pancreas produces reduced amounts of insulin. In South Africa, type II diabetes prevails as the most common variant, accounting for nearly 90% of diagnosed cases (Egan & Dinneen, 2019). The prevalence of diabetes in South Africa has become a pressing public health concern, particularly in under-resourced regions like rural areas, where access to well-equipped healthcare facilities and medical professionals is limited (Mayosi. et al 2012). This limitation results in delayed diagnosis for individuals who require diabetes screening, potentially leading to late detection. Timely and accurate diagnosis plays a vital role in preventing complications associated with delayed detection (Egan & Dinneen, 2019).

In rural South Africa, healthcare providers often depend on outdated diagnostic approaches, primarily relying on clinical assessments made by healthcare professionals. These traditional techniques are susceptible to misdiagnosis and suboptimal treatment, heightening the risk of complications for individuals with diabetes (Mendenhall & Norris, 2015). In certain regions, healthcare professionals are assigned to multiple healthcare centres, necessitating patients to wait until the diabetes specialist is available at their specific location when seeking a diabetes specialist's expertise or diagnostic tests. This lack of dependable and easily accessible diagnostic resources exacerbates the challenge of late-stage diabetes diagnosis. Furthermore, in some instances, diabetes specialists, also referred to as diabetologists or endocrinologists, may not be accessible in these areas, leading to general practitioners or nurses conducting diabetes evaluations. General practitioners and nurses typically lack the specialised knowledge and experience required for accurate diabetes diagnosis, potentially resulting in diagnostic inaccuracies and substandard care (Mendenhall & Norris, 2015).

Expert systems represent specialised artificial intelligence (AI) agents endowed with the ability to emulate human decision-making processes by utilising a knowledge base and predefined reasoning rules (Zeki et al. 2012). These systems heavily depend on decision-making algorithms and incorporate domain-specific knowledge, often referred to as ontologies, to underpin their decision-making capabilities (Wahl et al. 2018). In our research, we aim to encapsulate diabetes knowledge and expertise through the formulation of rules, heuristics, and guidelines.

The application of expert systems as an alternative for diagnosing diabetes has exhibited promising outcomes in diverse healthcare settings. Multiple studies, such as the one conducted by Wahl et al. in 2018, have highlighted the substantial potential of expert systems in enhancing diagnostic precision, reducing human errors, and ultimately improving patient outcomes in the realm of diabetes healthcare management. Despite the accomplishments of expert systems in various research domains, their utilisation in diabetes diagnosis within rural South Africa remains largely uncharted territory.

## Problem statement

A critical need for improved diabetes diagnosis exists in the rural reaches of the Western Cape, South Africa, where access to medical resources is limited. This study conducts a thorough investigation of whether expert systems can be employed as a solution to this problem. Expert systems that harness the power of artificial intelligence hold the potential of reproducing the diagnostic acumen of human specialists, with the goal of overcoming the barriers that prevent correct diabetes diagnosis in these neglected rural areas. Against this context, the purpose of this research is to determine the viability and effectiveness of integrating expert systems to perform diabetes diagnosis. By assessing the potential impact of expert systems, this study aims to pave the way for improved diagnostic precision, timely interventions, and, ultimately, improved health outcomes for people living in rural parts of South Africa's Western Cape province. The findings fit with the deep importance of tackling healthcare disparities and improving the well-being of underrepresented communities in this attempt.

## Research objectives

* To assess the performance of machine learning models in predicting diabetes development in individuals and conduct a systematic analysis by comparing them using a variety of algorithms and methodologies.
* To determine the most effective predictive machine learning model as the foundation for later integration into an expert system framework.

## Research questions

* What machine learning models (MLMs) are currently available for diabetes diagnosis prediction, and how do their predictive accuracy and precision compare?
* Is a predictive model adequate to serve as a framework for diabetes diagnosis?

## Relevance of research

Utilising AI expert systems for diabetes diagnosis in resource-constrained Western Cape, rural South Africa carries significant relevance. It addresses the need for improved healthcare access in regions where medical facilities and specialised healthcare professionals are lacking. By offering early diabetes detection, it mitigates the risk of complications associated with late diagnoses. Moreover, it minimises the chances of misdiagnosis, ensuring that patients receive appropriate care.

This approach optimises the use of healthcare resources by automating routine diagnosis, allowing medical professionals to focus on complex cases. It contributes to enhancing patient outcomes through timely and accurate diagnoses. Furthermore, it fosters innovation and research, potentially leading to the development of specialised systems for addressing various healthcare challenges in similar underserved contexts.

In conclusion, the implementation of AI expert systems for diabetes diagnosis in resource-constrained rural South Africa has the potential to make healthcare more accessible, improve diagnostic accuracy, and enhance the overall well-being of individuals in these underserved areas.

## Thesis structure

**Chapter 1: Introduction**

In this chapter, we will present a summary of the research, describing the challenges of diabetes diagnosis in the resource-constrained Western Cape, Rural South Africa. We will outline the study objectives and scope, emphasising the importance of leveraging AI expert systems to address these difficulties. The chapter will function as the thesis's introduction and will establish the groundwork for the future chapters.

**Chapter 2: Literature Review**

This chapter will investigate the existing diabetes literature, as well as the usage of AI expert systems for medical diagnosis. We will look at the healthcare difficulties that are unique to resource-constrained areas, as well as the theoretical basis for AI in healthcare. In addition, we will outline key studies and their conclusions in order to build a solid theoretical foundation for our research.

**Chapter 3: Methodology**

The third chapter will describe the research technique and methodologies used in our study. We will go into the data collecting methods, AI model selection, and how we adapted it to the situation. Ethical considerations and the research implementation plan will also be discussed, ensuring that our technique is transparent and rigorous.

**Chapter 4: Empirical Analysis and Results**

Here, we will provide the findings of our research, including an evaluation of the AI expert system's performance. We will compare its results to those of existing diagnostic procedures and provide an explanation of the findings. This chapter will provide an in-depth analysis of the system's effectiveness in identifying diabetes in the study area.

**Chapter 5: Discussion and Recommendations**

We will examine the significance of our findings for healthcare in resource-constrained places such as Western Cape, Rural South Africa, in Chapter 5. We will also discuss any difficulties or constraints identified throughout the research and make suggestions for future research or practical applications of AI in diabetes diagnosis.

# Literature review

## Introduction

In the past few years, there has been a growing body of literature focusing on the utilisation of expert systems for diagnosing diabetes in different healthcare settings (Zeki et al., 2012; Rotchford & Rotchford, 2002; Bhavadharini et al., 2016). This section intends to provide a comprehensive review and synthesis of these studies, encompassing their research methodologies, results, and constraints. This section will also explore studies that have been done to try and understand the state of medical care in resource constrained settings in South Africa.

## Expert systems

Zeki et al.(2012) conducted a study where an expert system that was used to diagnose diabetes was developed using fuzzy-logic. The study utilised a retrospective cohort design and compares the diagnostic accuracy of expert systems to that of one healthcare professional. The results demonstrate that the expert systems achieve a higher level of accuracy in diagnosing diabetes compared to the traditional methods. The study also concluded that expert systems have broader and vast knowledge compared to one physician and they can potentially perform well in areas where the patients do not have access to physicians who may diagnose them. The authors conclude that with expert systems, diabetes diagnosis can be done early, and patients may be helped by expert systems. The study was done in the context of Iran which might be different from the South African context. The study also acknowledged limitations in terms of sample size and generalizability, highlighting the need for larger-scale studies in different healthcare settings. Hence, we propose undertaking this type of study in the South Africa context.

Bhavadharini et al. (2016) conducted a study that emphasises the potential of alternative methods for diabetes diagnosis in resource-constrained settings. The study specifically focuses on using capillary blood glucose for screening gestational diabetes mellitus in rural India. The findings of the study indicate that venous plasma glucose (VPG) remains the preferred diagnostic test for gestational diabetes mellitus (GDM) in such settings. However, the study identifies a significant limitation with the capillary blood glucose (CBG) test, namely its inadequate sensitivity and specificity. Additionally, the study acknowledged that the sensitivity of diagnosing GDM using the International Association of Diabetes and Pregnancy Study Groups (IADPSG) criteria was still below optimal levels. This study provided evidence that the capillary blood glucose (CBG) test was not suitable for accurately diagnosing diabetes. Diagnostic accuracy should not be compromised based on the availability of resources; it should consistently align with the benchmark tests established by the IADPSG. Therefore, the potential use of expert systems in resource-constrained environments should be investigated as a potential alternative to the capillary blood glucose test.

Kentala et al. (1998) conducted a comparative study to assess the diagnostic accuracy of a neurotologic expert system in comparison to human experts. The study determines the level of agreement between the diagnoses generated by the expert system and those made by human experts. The data used for both the expert system and the human experts was identical, ensuring a fair comparison. The study revealed that when the expert system and human experts had access to the same patient information, the expert system outperformed the human experts in terms of diagnostic accuracy. The expert system achieved an accuracy rate of 65%, whereas the human experts achieved 54%. However, when the expert system received incomplete patient information while the human experts had access to complete information, the expert system's accuracy decreased to 65%, while the human experts maintained an accuracy of 69%. These findings consistently demonstrated the potential of expert systems to enhance diagnostic accuracy and reduce misdiagnosis rates. The study emphasised the importance of expert systems in improving patient outcomes and highlighted the need for further research on their implementation and scalability in different healthcare settings. This indicates the importance of exploring the broader application of expert systems to enhance healthcare practices and decision-making.

### Diabetes in South Africa

Rotchford and Rotchford (2002) assessed diabetes care and complications in rural South Africa. The researchers employed a mixed-methods approach, combining qualitative interviews with healthcare providers and quantitative analysis of diagnostic outcomes. The study highlighted that diabetes had been recognized as a disease requiring special attention in South Africa, with an emphasis on implementing a community-based primary care approach. The study found that there existed a substantial gap that needed to be addressed in order to have achieved satisfactory standards of care for diabetic patients in rural communities. The study emphasised the importance of providing additional training and support for nursing staff within the healthcare system. It concluded that improving the accuracy of diabetes diagnosis would significantly enhance diabetes care by enabling early detection and initiation of appropriate treatment. Based on the study's conclusion that more training is needed for existing healthcare professionals, it becomes evident that exploring new methods of diabetes diagnosis, such as utilising expert systems to support medical care in resource-constrained settings, should be a subject of further research.

Mendenhall and Norris (2015) study found that there was a need for systematic changes in healthcare practices to ensure effective diabetes education and management in clinical settings. Disparities between public and private healthcare systems, encompassing variations in diabetes counselling, medication availability, quality of care, and patient waiting times, often result in patients resorting to self-care rather than attending clinical visits. Such a shift toward self-care poses risks to the overall clinical system. Notably, structural barriers within the public health system present challenges to medication adherence, further exacerbating the issue. Consequently, it is imperative to prioritise and strengthen national efforts in diabetes counselling and management. Exploring alternative methods, such as utilising expert systems, could potentially aid in diabetes counselling, particularly in scenarios where physicians are not readily accessible. With an increasing number of diabetes patients and a limited increase in diabetes experts, individuals residing in economically disadvantaged communities face greater difficulty accessing quality healthcare compared to those in more affluent neighbourhoods. Given that the research was conducted with a sample size of only 27 women, conducting a study with a larger sample of diabetes patients would be beneficial to enhance the robustness of the findings.

## 

## Conclusion

The studies reviewed above provide compelling evidence of the potential of expert systems to enhance diagnostic accuracy and improve patient outcomes in the context of diabetes diagnosis. These systems offer the valuable advantage of ensuring consistent and evidence-based decision-making, which is especially beneficial in resource-constrained settings. However, it is essential to conduct additional research to validate these findings on a larger scale and to address various challenges, including generalizability, scalability, and the establishment of standardised evaluation metrics. It is recommended that future research should explore the application of expert systems in rural South Africa, as it would contribute to strengthening the findings and provide greater confidence in their effectiveness.

# Methodology

## Introduction

This research study used a secondary dataset namely, Diabetes Health Indicators Dataset (Alex, 2021), to develop various predictive machine learning models for diabetes prediction. The models were developed based on numerical data that was collected from real-world observations and measurements. This aspect of the research aligns well with the empirical focus of positivism research philosophy (Winahyu & Piaseu, 2023). The research leveraged quantitative data and machine learning techniques to predict the likelihood of diabetes in individuals. This methodology section provides a comprehensive account of our research design, the underlying reasoning, the methods employed for data collection and analysis, as well as the strategies that directed our research.

## Research Design and Rationale

Selecting an appropriate research philosophy is important in making sure the methodology of a research aligns well with the research objectives. In line with Winahyu and Piaseu (2023), the positivism philosophy consists of a fundamental belief in the existence of a discernible reality that can be realised through empirical observation and data analysis. The positivist stance was reached upon discovery of secondary dataset that could be used to realise the research objectives of this study.

The Diabetes Health Indicators Dataset served as an empirical source for our research design. This dataset consisted of numerical data collected from real-world observations and measurements pertaining to diabetes and its correlated factors. The utilisation of this dataset was harmonious with the positivism philosophy, as it enabled the utilisation of factual and measurable data to formulate predictions based on empirical evidence. This approach underscores the value of empirical data in bolstering the creation of predictive models designed to objectively estimate the likelihood of diabetes in individuals.

Furthermore, the research design we employed was rooted in a quantitative approach. Quantitative approach resonates with positivism research philosophy (Winahyu & Piaseu, 2023). The quantitative approach encompassed data analysis, characterised by rigorous statistical methods and machine learning techniques, and constituted the foundational framework of our research (Winahyu & Piaseu, 2023). This approach was instrumental in the conversion of the extensive pool of numerical data into predictive models capable of making objective, data-driven assessments regarding diabetes diagnoses.

## Data Source

The data source for this project is hosted on Kaggle dataset platform. The dataset is derived from the Behavioural Risk Factor Surveillance System (BRFSS), which is an annual health-related telephone survey conducted by the Centres for Disease Control and Prevention (CDC). The BRFSS survey is a comprehensive data collection effort that compiles responses from over 400,000 Americans. It covers a wide range of health-related topics, including risk behaviours, chronic health conditions, and preventative services utilisation. This survey has been an invaluable source of public health data since its inception in 1984(Alex, 2021).

For our project, we specifically focused on data from the year 2015, which is available in a CSV format on Kaggle. This original dataset contains responses from 441,455 individuals and encompasses a total of 330 features. These features include both direct survey questions asked of participants and calculated variables derived from individual responses.

The dataset is divided into three distinct files, each designed for a specific research purpose. The first dataset, "diabetes\_012\_health\_indicators\_BRFSS2015.csv," is a clean dataset containing responses from 253,680 participants in the CDC's BRFSS2015. It features a target variable, "Diabetes\_012," with three classes: 0 (indicating no diabetes or diabetes only during pregnancy), 1 (representing prediabetes), and 2 (indicating diabetes). However, it's important to note that this dataset exhibits class imbalance, and it contains 21 additional variables describing various health indicators (Alex, 2021).

The second dataset, "diabetes\_binary\_5050split\_health\_indicators\_BRFSS2015.csv," comprises responses from 70,692 participants. It maintains an equal 50-50 split between respondents with no diabetes and those with either prediabetes or diabetes. The target variable, "Diabetes\_binary," has two classes: 0 (representing no diabetes) and 1 (indicating prediabetes or diabetes). Like the first dataset, it includes 21 feature variables and is balanced in terms of class distribution.

The third dataset, "diabetes\_binary\_health\_indicators\_BRFSS2015.csv," also features responses from 253,680 participants in the CDC's BRFSS2015. Similar to the second dataset, the target variable, "Diabetes\_binary," has two classes: 0 (indicating no diabetes) and 1 (representing prediabetes or diabetes). Notably, this dataset is not balanced. Like the previous datasets, it comprises 21 feature variables describing various health indicators (Alex, 2021).

## Data Selection and sampling

In the data selection and sampling section, we combined all the three subsets of the dataset to create one huge comprehensive dataset. From this merged dataset, we then conducted random sampling to obtain a sample size of 200,000 sample size. This approach was crucial to ensure a probability-based sampling method that adhered to the principles of randomness, allowing us to draw unbiased conclusions (Acharya et al., 2013). The combined dataset was also balanced, that is equal records of class 0 and equal records for class 1. This can be seen in the Figure 3 in Appendix A.

The original dataset contained 21 features, but after thorough data preprocessing, it became evident that only 9 features were pertinent for effective diabetes diagnosis. These essential features, namely HighBP, HighChol, CholCheck, BM', Smoker HeartDiseaseorAttack, PhysActivity, HvyAlcoholConsump, Age were identified as the most relevant indicators for our research objectives (Alex, 2021). This is evident in Figure 2, and Figure 4 in Appendix A.

By combining the subsets and performing random sampling, we aimed to achieve a representative sample that would enable us to draw robust conclusions while focusing on the critical features essential for diabetes diagnosis. This process ensured that our analysis was based on a balanced and relevant dataset, promoting the accuracy and effectiveness of our research.

## Data Splitting

In the context of this research, we have separated the dataset into two unique subsets which consists of a training set and a testing set. The training set encompasses the bigger portion of the data and therefore serves the purpose of training the model. This allocation follows the widely adopted 80/20 split, with 80 percent of the data reserved for training and 20 percent for testing.

The testing set is specifically reserved for fine-tuning hyperparameters and is saved for the final evaluation of the model's performance. This approach to data division aligns with a conventional practice in the field of machine learning. ensuring that models are assessed on previously unseen data, thereby offering a robust assessment of their generalised capabilities.

## Machine Learning Models

Upon careful examination of our dataset, it became apparent that the target variable "Diabetes\_binary" consisted of two classes: 0, indicating the absence of diabetes, and 1, representing diabetes. This binary classification naturally led us to frame our analysis as a classification problem, where the goal is to categorise data points into predefined classes. Therefore, we approached this research as a supervised learning example, as we had labelled data with known outcomes as according to Cunningham, Cord, and Delany (2008), supervised learning is well fitted where we have labelled data.

To address this classification problem, we decided to employ supervised machine learning models. Our primary objective was to develop three distinct models, each with its unique characteristics, to enable a comparative analysis aimed at selecting the most suitable model for recommendation. The performance of these models would be evaluated based on accuracy, a common metric for classification tasks. The model with the highest accuracy among the three would be considered the best fit for our research.

The supervised learning models that we chose for this research were Logistic Regression, Decision Trees, and the Random Forest Classifier. These models were selected for their ability to handle classification tasks effectively (Cunningham et al., 2008), and their performance was to be rigorously assessed to determine which one is the most proficient in predicting diabetes.

## 3.7 Ethics that were considered during the research.

**Data Ownership and Usage Rights:** We have acknowledged that the dataset employed in this study was made available under the CC0 (Public Domain) licence. We have done proper attribution to the Centres for Disease Control and Prevention (CDC) as it is necessary to show respect for their efforts in collecting and sharing this data.

**Confidentiality and Privacy:** Given the sensitive medical information contained in the dataset, safeguarding data privacy and confidentiality is of utmost importance. The dataset was already diligently de-identified and anonymized before our usage to protect the identities of individuals. However, a rigorous approach was maintained to avoid any inadvertent disclosure of personal identifiable information, ensuring the preservation of participants' sensitive data.

## 3.8 Conclusion

In this section, we discussed the design and rationale of the positivism research philosophy. We provided a detailed explanation of the data analysis methods and the procedural steps that were followed. In the next section, our focus will shift towards the empirical findings and the thorough analysis of the results that were obtained following the implementation of the research study.

# Empirical analysis and results

## Introduction

In this section, we look at the results that were obtained after running the three supervised machine learning models: Logistic Regression, Decision Tree, and Random Forest Classifier. This section presents a thorough examination of the performance metrics and a comparative analysis encompassing these three models. We delve into an in-depth evaluation of the results generated by each model, shedding light on their individual strengths and weaknesses, thus enabling a comprehensive understanding of their respective performances. By scrutinising key performance metrics, we present a holistic view of the capabilities and limitations of these models, ultimately facilitating the selection of the best model for diabetes diagnosis.

## Machine Learning Model Performance Analysis

### Performance Metrics

In this research, accuracy, recall, and precision were the key metrics that were used to analyse the performance of the machine learning models. Accuracy quantifies the percentage of correct predictions, offering a broad sense of how well the model performs. However, accuracy can be misleading in diabetes prediction in scenarios where there are imbalanced datasets amongst the prediction classes. Imbalanced data may inflate accuracy when there are more non-diabetic cases, leading to ineffective models (Cunningham et al., 2008).

Recall, crucial in medical diagnoses like diabetes prediction, minimises false negatives by identifying true diabetic cases. It gauges the model's ability to correctly identify individuals with diabetes (true positives) among all individuals who indeed have the condition. A high recall minimises the risk of false negatives, ensuring that people with diabetes are not overlooked by the model, a critical aspect of a life-affecting task (Junker, Hoch, & Dengel, 1999)..

Precision measures the proportion of true positive cases among the predicted positive cases (Junker, Hoch, & Dengel, 1999). High precision is crucial in situations where the consequences of a false positive are significant. It ensures that when the model predicts an individual as diabetic, the likelihood of them truly having diabetes is substantial, thereby increasing confidence in the model's predictions.

Therefore, selecting the best diabetes prediction model involves balancing these metrics accordingly. While high accuracy is desirable, recall and precision weigh more in situations with significant consequences. The optimal model aligns with the practical implications and goals of diabetes prediction, ensuring effective and reliable outcomes.

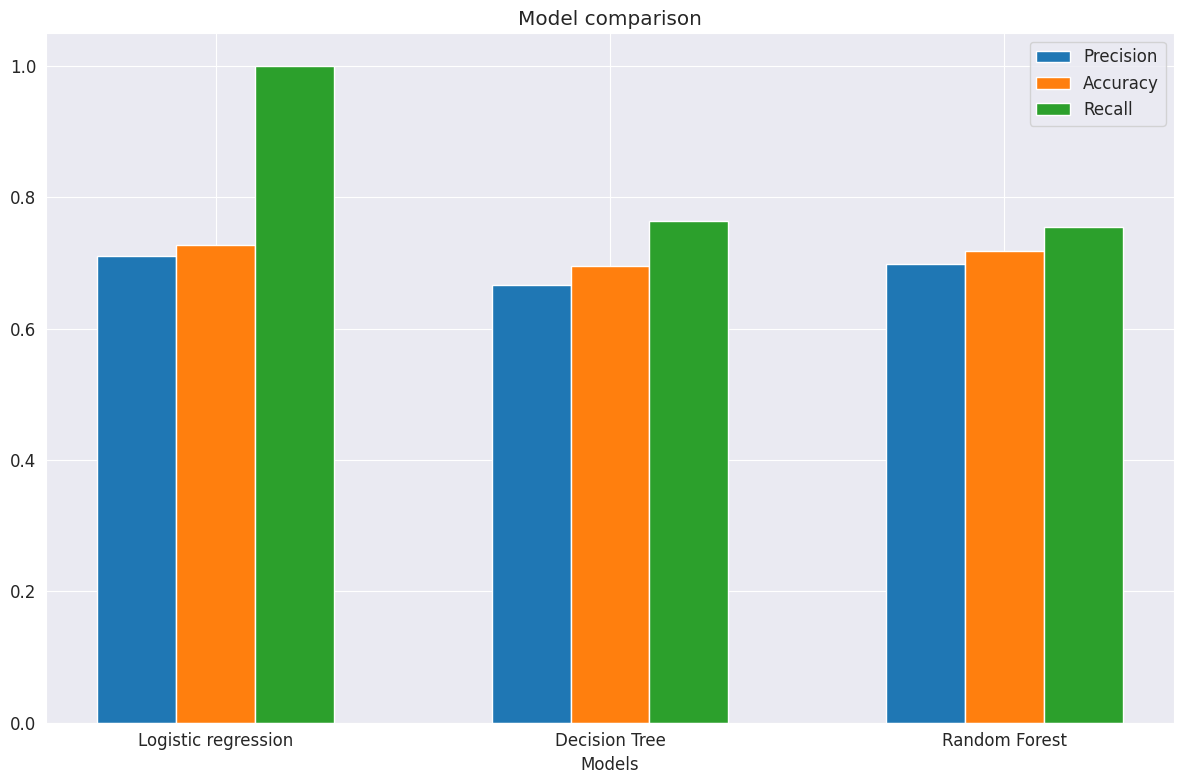
### Models evaluation

The most common issue that is encountered in supervised learning is overfitting where the accuracy of the training dataset is quite higher compared to that of the testing data. We decided to evaluate our models to check whether there was overfitting.

In our research study, as per Table 1 in Appendix B, we observed that logistic regression did not exhibit overfitting, primarily because the dataset was fairly and equally distributed across the two classes, allowing the model to generalise effectively as the decision boundary could be learned well during the training process. However, when it came to decision trees and random forest classifiers, we encountered overfitting issues. This overfitting was evident when the training accuracy significantly exceeded the testing accuracy, indicating that the models were capturing noise in the training data.

To address these overfitting problems, we employed hyperparameter tuning techniques to both decision tree and random forest classifier algorithms. Specifically, we adjusted hyperparameters like `n\_estimators`, `max\_depth`, `min\_samples\_leaf`, `max\_features`, and `min\_samples\_split`. By fine-tuning these parameters, for instance, setting `n\_estimators=50`, `max\_depth=2`, `min\_samples\_leaf=10`, `max\_features='sqrt'`, and `min\_samples\_split=20`, and according to Table 2 in Appendix B, we managed to resolve the overfitting issues. These adjustments struck a better balance between model complexity and generalisation, resulting in improved model performance and reliability in making predictions on new, unseen data.

### Comparative Analysis



*Figure 1: Graph showing the accuracy, precision and recall for each algorithm.*

In this comparative analysis we evaluate the accuracy, precision, and recall for the three supervised learning models. From Figure 1 above and Table 3 in Appendix B , Logistic Regression had the highest accuracy at 72.9%, benefiting from a well-balanced dataset. The Decision Tree followed closely with 69.6% accuracy, offering a balanced approach. The Random Forest Classifier, with 71.8% accuracy, strikes a reasonable balance between precision and recall. Logistic Regression achieves perfect recall, making it an ideal choice for diabetes diagnosis as it excels in both accuracy and recall. The Decision Tree strikes a balanced trade-off with a recall of 0.765. In contrast, the Random Forest Classifier maintains a good balance with a recall of 0.745. Since the study was looking at the balance of the three-performance metrics, the best model that can be recommended to be deployed in health care facilities to help with diabetes diagnosis. Therefore, the best machine learning model that was obtained from the research is logistic regression.

## Conclusion

From the results obtained and according to Figure 1, it is clear that the Logistic regression machine learning model was the best amongst all three due to its performance across all the three metrics that were used to determine the best model.

# Conclusion and recommendations

## Introduction

In the concluding section of this paper, we will revisit the research objectives and determine the extent to which they have been successfully achieved, while also reflecting on the key insights and findings derived from the research results.

## Summary of Research

In this research, our primary aim was to compare various machine learning models and systematically assess their performance using different performance metrics. We were further motivated to identify the most accurate predictive model, one that could lay the foundation for integration into an expert system framework for diabetes diagnosis.

Our comparative analysis, rooted in rigorous evaluation criteria of accuracy, precision, and recall, has yielded valuable insights. Among the three supervised learning models considered, Logistic Regression was the best model, having an accuracy rate of 72.9%. This accuracy was notably underpinned by the utilisation of a well-balanced dataset. Logistic Regression also had the best recall, making it the best choice for diabetes diagnosis as Logistics regression excels not only in overall accuracy but also in the ability to identify cases of diabetes effectively.

We can confidently assert that Logistic Regression is the most suitable machine learning model for deployment in healthcare facilities to enhance diabetes diagnosis. It offers a robust foundation for integration into expert systems, promising more accurate and data-driven healthcare solutions. By choosing Logistic Regression, we aim to empower healthcare professionals with a tool that combines exceptional accuracy and recall, ultimately improving patient care in the field of diabetes diagnosis.

## Conclusion and Contributions

The research undertaken aligns closely with the initially proposed contributions, which sought to harness AI-driven solutions, particularly decision tree models, to enhance the accuracy and efficiency of diabetes diagnosis in resource-constrained settings, with a specific focus on rural South Africa. The results of this study have significant implications for these proposed contributions and extend to a broader academic and scientific context.

In conclusion, the study's findings unequivocally point to Logistic Regression as the most accurate predictive model for diabetes diagnosis, achieving an impressive 72.9% accuracy and excelling in recall. This determination represents a crucial step towards realising the goal of improving diabetes diagnosis accuracy in resource-limited settings. By embracing Logistic Regression as the recommended model, we are advancing AI-driven healthcare solutions tailored for regions facing healthcare resource constraints.

Moreover, this research not only makes a noteworthy contribution to the field of diabetes diagnosis but also adds to the existing knowledge base on the application of AI in healthcare. It underlines the potential of AI in addressing healthcare challenges and highlights the critical role of machine learning models in enhancing patient care. Ultimately, the study's contribution lies in bridging the gap between limited resources and improved healthcare outcomes through the adoption of data-driven solutions.

## Future research

There are several opportunities that this research can be enhanced through.

Firstly, ensuring the ongoing effectiveness of machine learning models, such as Logistic Regression is continuously monitored and adapted are essential to maintain the models' accuracy and relevance amidst evolving healthcare landscapes. Future research should emphasise systematic evaluations and optimizations to guarantee sustained performance in diverse healthcare settings. This approach will enhance the model's dependability and foster trust among medical professionals and patients.

Secondly, future research must prioritise patient-centric healthcare by actively involving patients in the diagnostic process. This entails creating user-friendly interfaces, transparent explanations, and tools that empower patients to engage with their health data. Patient-centric solutions improve the overall patient experience, enhance adherence to treatment plans, and contribute to more accurate diagnoses.

Finally, expanding research to diverse geographic regions with varying healthcare infrastructures will shed light on the model's adaptability and effectiveness in different settings. This approach supports tailored implementations, addressing regional challenges and promoting equitable access to AI-driven healthcare solutions. It also contributes to a comprehensive understanding of AI's potential in improving global healthcare outcomes.

# References

Vayena, E. and Blasimme, A., 2018. Health research with big data: time for systemic oversight. The journal of law, medicine & ethics, 46(1), pp.119-129.

American Diabetes Association, 2010. Diagnosis and classification of diabetes mellitus. *Diabetes care*, *33*(Supplement\_1), pp.S62-S69.

Alex, T. (2021, August). Diabetes prediction dataset, Version 1. Retrieved October 20, 2023 from <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>.

Winahyu, K.M. and Piaseu, N., 2023. Philosophical and methodological perspective in developing nursing knowledge through research in diabetes. Journal of Holistic Nursing Science, 10(1), pp.58-64.

Egan, A.M. and Dinneen, S.F., 2019. What is diabetes?. Medicine, 47(1), pp.1-4.

Mendenhall, E. and Norris, S.A., 2015. Diabetes care among urban women in Soweto, South Africa: a qualitative study. BMC Public Health, 15(1), pp.1-7.

Zeki, T., Malakooti, M., Ataeipoor, Y. and Tabibi, S., 2012. An expert system for diabetes diagnosis. American Academic & Scholarly Research Journal, 4(5).

Mayosi, B. M., Flisher, A. J., Lalloo, U. G., Sitas, F., Tollman, S. M., & Bradshaw, D., 2012. The burden of non-communicable diseases in South Africa. The Lancet, 380(9859), 915-924.

Wahl, B., Cossy-Gantner, A., Germann, S. and Schwalbe, N.R., 2018. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings?. BMJ global health, 3(4), p.e000798.

International Diabetes Federation., 2019. IDF Diabetes Atlas, 9th edition, viewed 08 June 2023, from <https://www.diabetesatlas.org>.

Rotchford, A.P. and Rotchford, K.M., 2002. Diabetes in rural South Africa-an assessment of care and complications. South African Medical Journal, 92(7), pp.536-541.

Bhavadharini, B., Mahalakshmi, M.M., Maheswari, K., Kalaiyarasi, G., Anjana, R.M., Deepa, M., Ranjani, H., Priya, M., Uma, R., Usha, S. and Pastakia, S.D., 2016. Use of capillary blood glucose for screening for gestational diabetes mellitus in resource-constrained settings. Acta diabetologica, 53, pp.91-97.

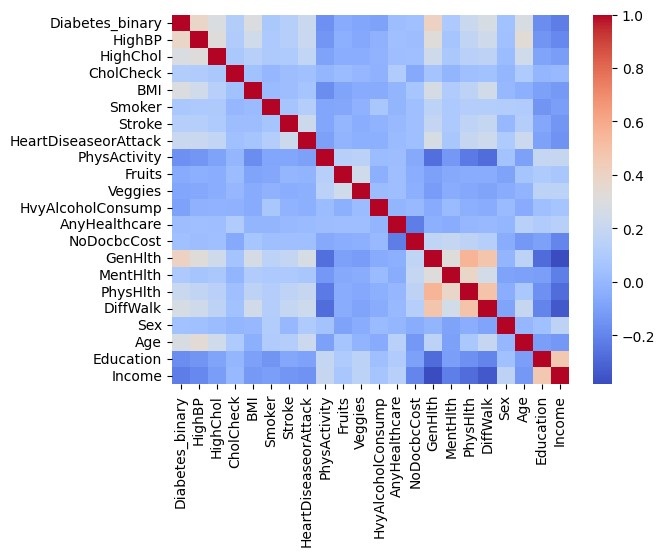
Kentala, E., Juhola, M., Auramo, Y. and Pyykkö, I., 1998. Comparison between diagnoses of human experts and a neurotologic expert system. Annals of Otology, Rhinology & Laryngology, 107(2), pp.135-140.

Acharya, A.S., Prakash, A., Saxena, P. and Nigam, A., 2013. Sampling: Why and how of it. Indian Journal of Medical Specialties, 4(2), pp.330-333.

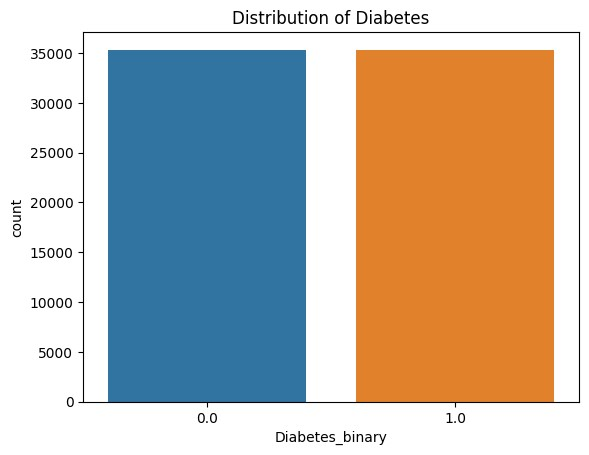
Cunningham, P., Cord, M. and Delany, S.J., 2008. Supervised learning. In Machine learning techniques for multimedia: case studies on organization and retrieval (pp. 21-49). Berlin, Heidelberg: Springer Berlin Heidelberg.

Junker, M., Hoch, R. and Dengel, A., 1999, September. On the evaluation of document analysis components by recall, precision, and accuracy. In Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No. PR00318) (pp. 713-716). IEEE.

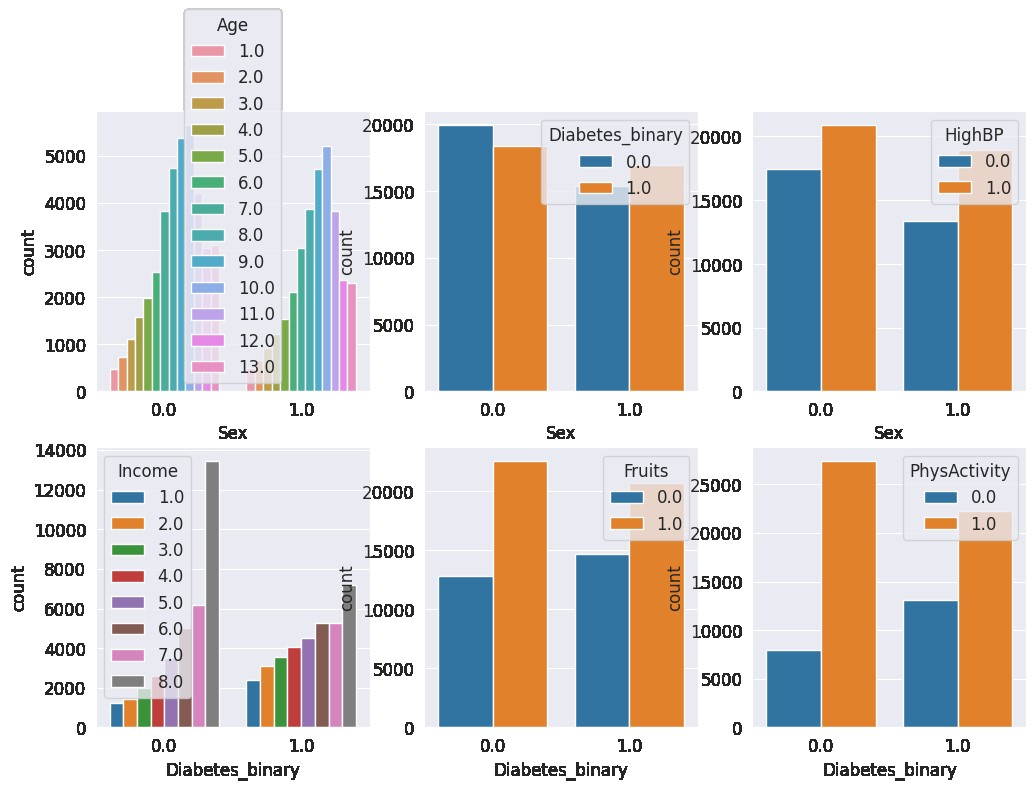
# Appendix A: Graphs



*Figure 2: Graph showing the co-variance of the various data variables that exists in the dataset.*



*Figure 3: Graph showing the distribution of Diabetes classes in the dataset.*



*Figure 4: Graphs showing various factors that affects the Diabetes*

# Appendix B: Tables

|  |  |  |
| --- | --- | --- |
| **Model** | **Training data accuracy** | **Testing data accuracy** |
| Logistic Regression | 0.727 | 0.729 |
| Decision Tree | 0.700 | 0.695 |
| Random Forest Classifier | 0.765 | 0.715 |

Table 1: Table showing accuracy of the training and testing dataset prior to hyper-parameter tuning

|  |  |  |
| --- | --- | --- |
| **Model** | **Training data accuracy** | **Testing data accuracy** |
| Logistic Regression | 0.727 | 0.729 |
| Decision Tree | 0.652 | 0.696 |
| Random Forest Classifier | 0.715 | 0.718 |

Table 2: Table showing accuracy of the training and testing dataset after hyper-parameter tuning

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** |
| Logistic Regression | 0.729 | 0.70 | 1.0 |
| Decision Tree | 0.696 | 0.698 | 0.765 |
| Random Forest classifier | 0.718 | 0.683 | 0.745 |

Table 3: Table showing accuracy, precision and recall results after hyper-parameter tuning

# Appendix C: Code for the data analysis

The code that was used for the data analysis can be found here: <https://github.com/R4pidAce/Reseach8412/blob/main/Research_Analysis_Craig.ipynb>