Comparative Analysis of Machine Learning Techniques for Predicting Diabetes

Conference Paper · October 2024

DOI: 10.46254/EU07.20240073

CITATIONS
0

READS
23

6 authors, including:

Isaac Murere
National University of Science and Technology, Bulawayo
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Sibusisiwe Dube
National University of Science and Technology, Bulawayo
42 PUBLICATIONS **118** CITATIONS
SEE PROFILE

Martin Muduva
Midlands State University
12 PUBLICATIONS **7** CITATIONS
SEE PROFILE

7th European Conference on Industrial Engineering and Operations Management
Augsburg, Germany, July 16-18, 2024
Publisher: IEOM Society International, USA DOI: 10.46254/EU07.20240073
Published: July 16, 2024

Comparative Analysis of Machine Learning Techniques for Predicting Diabetes

Isaac Murere, Belinda Ndlovu, and Sibusisiwe Dube
Department of Data Analytics and Informatics, Faculty of Applied Sciences
National University of Science and Technology, Bulawayo, Zimbabwe
P.O Box AC939, Ascot, Bulawayo, Zimbabwe
n02222456r@students.nust.ac.zw, belinda.ndlovu@nust.ac.zw, sibusisiwe.dube@nust.ac.zw

Martin Muduva
Department Information and Marketing Science Department
Midlands State University
Senga Road, Gweru, Zimbabwe
muduvam@staff.msu.ac.zw

Fungai Jacqueline Kiwa
Department of ICT and Electronics at Chinhoyi University of Technology
Chinhoyi University of Technology
Private Bag 7724, Chinhoyi
jkiwa@cut.ac.zw

Abstract

Diabetes, a chronic illness causing serious health problems, affects millions of people globally. With cases expected to rise, effective strategies for managing, detecting, and preventing the disease are essential. Artificial intelligence (AI) and machine learning (ML) have become powerful allies in this fight. These advancements aid in the automated detection of eye complications (retinopathy), supporting clinical decisions, identifying high-risk populations, and empowering patients to manage their health. The significant public health challenge of diabetes in Zimbabwe, impacting all demographics, highlights the need for better solutions. This research aims to develop a precise predictive model for diabetes using the CRISP-DM methodology. Machine learning techniques like random forest, Naive Bayes, XGBoost, decision trees, and support vector machines, were used to predict the presence of diabetes. The results revealed that the random forest approach outperformed other models, demonstrating a larger area under the curve (AUC).

Keywords

Machine Learning (ML), Artificial Intelligence (AI), Diabetes care, Random Forest, Naive Bayes, XGBoost, Decision Trees, Support Vector Machines.

1. Introduction

Diabetes is a chronic illness that has grown to be a significant global public health concern (Hanlon et al. 2020). Diabetes caused around 1.5 million deaths in 2019 and is predicted to cause 5 million deaths by 2030 (Wou et al. 2019). This translates to a significant burden on healthcare systems, with global costs reaching billions of dollars. The disease is particularly concerning in low- and middle-income countries (Tripathi et al. 2022). The World Health Organization (WHO) warns of a diabetes crisis, with millions already affected, especially in developing countries and these numbers are expected to climb even higher in the coming years (Mozaffarian 2020).

Diabetes has reached epidemic proportions globally, with over 400 million people currently affected and a projected increase to over 600 million by 2040 (Ong et al. 2023). It is acknowledged that diabetes is one of the most common health issues in the globe (Mpofu et al. 2024). Diabetes a non-communicable disease, requires smart ways to manage it (Maguraushe and Ndayizigamiye, 2024). Smart technologies for monitoring diabetes such as the Internet of Things are under-researched for continents such as Africa. Yet, it has been observed that this continent will have one of the main rates of diabetes prevalence and that its resources are insufficient to manage this illness (Mutunhu et al. 2022). In Zimbabwe, the diabetes prevalence stands at approximately 9.6%, surpassing the global average of 8.5% (Arokiasamy et al. 2021). Several factors contribute to the high prevalence of diabetes in Zimbabwe, factors as poverty, limited healthcare options, and unhealthy diets rich in processed carbs and fats (Mureyi et al. 2022). As a result, diabetes is a key public health issue in Zimbabwe, with a high burden of morbidity and mortality. Unfortunately, the affordability of diabetes management creates a significant obstacle to controlling and preventing the disease (Mureyi et al. 2022). There is a significant need for improved predictive models, personalized treatment plans, and analysis of health data in Zimbabwe (Gona et al. 2021). Currently, there is a lack of effective predictive models for diabetes risk, which makes it difficult to target prevention and treatment efforts (Sun et al. 2022). There is also a lack of personalized treatment plans, therefore patients often do not receive the optimal care for their needs (Mureyi et al. 2022). Additionally, technologies such as the Quantified-Self, have not been comprehensively researched in fostering diabetes self-care because of awareness issues (Mutunhu et al. 2024). Moreso, there is limited access to health data, which makes it difficult to identify trends and make informed decisions about diabetes management and prevention (Mureyi et al. 2022). Therefore, addressing these gaps might lead to significant improvements in diabetes outcomes in Zimbabwe. Improved diabetes management and prevention could have significant benefits for individuals, communities, and the health system in Zimbabwe (Ernersson et al. 2023). The World Bank (2021) reports that the prevalence of diabetes in Zimbabwe is 2.1% for people aged 20 to 79. The most recent data from the World Health Ranking (2020) revealed that diabetes-related fatalities in Zimbabwe reached 3344 in 2020, making it the country's sixth most common cause of death.

Poised with this global threat this research proposes the use of machine learning algorithms for predicting diabetes. Machine learning is growing in popularity as more research is conducted in the field of medicine. At this point, clinical decision support systems can incorporate it because of its sufficient effectiveness (Mukura and Ndlovu 2023) . Given the importance of early diabetes detection and the abundance of machine learning algorithms available, a thorough comparative study is necessary to determine the best method for this predictive task. This research aims to provide significant insights into the strengths and weaknesses of random forest, Naive Bayes, XGBoost, decision trees, and support vector machines in the context of diabetes prediction by evaluating performance and time complexities for each technique. In addition to adding to the body of knowledge in the field of machine learning for healthcare applications, the comparative analysis's conclusions will be useful in assisting researchers and healthcare professionals in choosing the best algorithm for diabetes risk assessment. Ultimately, this study's findings may lead to better health outcomes for those who are at risk by facilitating prompt interventions, enhancing early identification, and improving overall diabetes care.

The paper is organized to flow logically: Section 1 lays the groundwork with background information, Section 2 dives into previous research, Section 3 details the chosen methods, Section 4 explains how data was collected, and the final section presents the findings and analysis.

2. Literature Review

Predicting diabetes is a difficult undertaking that necessitates examining several variables and how they interact. With the advent of machine learning algorithms that can handle big datasets, capture non-linear correlations, and produce reliable predictions, they have attracted a lot of attention in this field. Lai et al. (2019) built a model to predict diabetes risk in Canadians. This model analyses patient data like age, gender, blood pressure, cholesterol levels, and blood sugar levels to identify those most likely to develop the disease. Deep comprehension of medical data, prognosis, and disease diagnosis are made possible by data mining applications in the healthcare industry (Tripathi et al. 2022). For example, it could offer a clearer knowledge of the relationship between several chronic illnesses, like diabetes, which is a main health issue and a leading cause of death (Sharma et al. 2016).

Data mining techniques play a major role in data analysis. By combining machine learning, artificial intelligence, statistics, and database approaches, data mining is the process of identifying patterns in enormous data sets (Song et al. 2021). Researchers are currently focusing a lot of effort on creating AI-based instruments and procedures appropriate for tracking and managing chronic illnesses. In particular, ML models have been extensively used to

estimate the likelihood of a disease developing under different assumptions or risks (Standl et al. 2019). Support vector machines, artificial neural networks, and decision trees are some potent machine-learning technologies that can be applied to this problem. Priyadarsini and Titus (2020) classified the patients into diabetic and non-diabetic groups using supervised machine learning algorithms whilst Rama and Prasad (2022) forecasted the beginning of diabetes, used an ensemble-supervised learning strategy. Five popular classifiers were chosen for the ensembles, and the outputs were aggregated using a meta-classifier. Carpinteiro et al (2023) study explored how well different machine learning techniques (SVM, Logistic Regression, and ANN) can predict diabetes early on. The ultimate goal was to develop a smart system for diabetes prediction using patient data, similar to the approach proposed by Sakib et al. (2021) who suggested data mining for this purpose.

Hasan et al. (2020) and Zou et al. (2018) highlighted the vast amount of research on data mining techniques for diabetes detection and prediction. This field encompasses developing new techniques, analyzing their performance, and reviewing existing methods. Similarly, Zhu et al. (2021) conducted a comprehensive review of data mining in diabetes research. These studies demonstrate that data mining is a crucial tool for managing blood sugar levels (glycemic control) and holds great promise for future advancements. Predictive models for diabetes can play a key role in early detection and treatment. These models, like those developed using logistic regression Daghistani and Alshammari (2020) or for undiagnosed cases Ong et al (2023) can identify individuals at high risk of developing the disease. This early identification allows doctors to screen for pre-diabetes and choose the most appropriate treatment plan for each patient (Liu et al. 2023).

Based on the analysed research this research acknowledges the works done by previous research however, there are still missing gaps in the literature. The following gaps were discovered in the previously reviewed literature on diabetes treatment management. The study by Priyadarsini and Titus (2020) lacks evaluation of novel machine learning algorithms for diabetes prediction. Ong et al (2023) extracted the techniques used for the evaluation of machine learning methods developed for the prediction of diabetes complications such as the temporality of prediction, the risk of bias, and validation metrics. The objective was to prove whether machine learning exhibited discrimination ability to predict and diagnose diabetes. Although this ability was confirmed, the authors did not report which machine-learning model produced the best results. Consequently, another study by Zhu et al. (2021) lacks data availability, which can make deep learning model construction and assessment more difficult. Interpretability and transparency are important for acceptability and confidence in healthcare settings, so this can be concerning. Thus, there is a need for research to investigate and create methods to enhance deep learning models' interpretability in diabetes care. One way to do this could be to look at ways that show and explain how deep learning models learn representations and make decisions. This study will utilize five machine learning techniques that offer insights into the characteristics and elements that influence the predictions to counter these aforementioned shortcomings.

3. Methods

Cross-Industry Standard Process (CRISP-DM) data mining was employed in this research. This standardized process, known as data mining, allows researchers to uncover patterns, trends, and connections hidden within large datasets (Zia et al. 2022). The CRISP-DM approach was used because of its practical experience with industrial data mining and because it is thought to be appropriate for industrial projects among related process models (Studer et al. 2021). The procedure is broken down into six steps: data preparation, business understanding, deployment, modeling, assessment, and deployment.

Predicting the diagnosis of each patient and the clinical support that will be given to them are the primary goals of the business understanding phase. Other goals include estimating the classification of a particular diabetic patient and evaluating the influence of each variable on this characterization. The “PIMA” dataset, which includes information on Indians with both positive and negative diagnoses for diabetes, was taken from the internationally renowned intensive care unit (ICU) repository after being extracted. In the following chapter, the central tendency imputation technique was used to obtain clear data after the missing data had been located. Data balancing procedures, coding, and data structuring was done to prepare the data utilised by the models. The Support Vector Machine, Random Forest, XBoost, Naive Bayes and Decision Tree models were assessed using a confusion matrix, accuracy, recall, and precision. To determine whether the models were appropriate for the dataset and development of the model, the models underwent a comprehensive evaluation.

4. Data Collection

The data for the study was from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The authors retrieved the data from Kaggle, a popular online platform. Kaggle hosts machine learning competitions where data scientists test their skills by building models for various datasets. Kaggle offers a vast collection of datasets, code examples, and tutorials, making it a valuable resource for anyone working with data analysis or machine learning. The collected data was securely stored on an encrypted laptop, and any patient-identifiable information was removed to uphold confidentiality. Feature engineering techniques were applied to identify and include the most influential features that contribute to the desired outcome.

4.1 Data quality assessment

The assessment of data quality, which was carried out utilizing the six data quality magnitudes, yielded the following results:

Relevance: The dataset underwent an evaluation to ascertain the relevance of its included features for predicting diabetes outcomes. Additionally, deliberations were undertaken regarding the potential necessity of supplementary data sources or variables to augment the predictive efficacy of the model.

Granularity: The data was aggregated at an appropriate level for the predictive modelling task.

Interpretability: The data can be easily understood and analysed by stakeholders. The data was well-organized and used consistent formats. The data types were accurate, and the variable names were clear and easy to understand, labels, and units of measurement, as well as the documentation or metadata provided with the dataset.

Timeliness: The data is up-to-date and relevant for the predictive modelling task, and it aligns with the target population or timeframe of interest**.**

Integrity: Given their origin from reputable sources, the data is considered to have minimal errors or inconsistencies.

Accuracy: Data was considered reasonably accurate since it originated from the recording institutions. It was assumed that all data underwent a quality assurance process.

5. Results and Discussion

The data was analysed with Python software through the use of machine learning algorithms. Some packages were installed before the study, such as Scikit-learn, NumPy, and Pandas, to help with modelling and data manipulation. The features and organization of the dataset were discovered using exploratory data analysis utilizing techniques like describe(), info(), and shape(). A basis for more data analysis and modelling was established by these actions.

5.1 Results

The correlation heatmap (Figure 1) offers important information on the variables impacting diabetes. The correlation shows the movement trend of two values about each other, as shown in Figure 1 the correlation coefficient matrix between the features is shown. Most of the features have a positive correlation with the outcome value. Older women tend to have had more pregnancies, and this can increase their risk of type 2 diabetes. There are a few reasons for this. First, some women develop gestational diabetes during pregnancy, which can raise their future risk of type 2 diabetes. Second, pregnancy can make the body less sensitive to insulin (insulin resistance). Over time, with more pregnancies, this reduced sensitivity can build up and increase the risk of type 2 diabetes. Finally, as women age, their bodies naturally become less efficient at processing sugar (glucose metabolism). All these factors combined can significantly raise the risk of type 2 diabetes in women who have had multiple pregnancies later in life.

Another significant finding is the moderate correlation between BMI and skin thickness. This implies that as an individual's body mass index increases, it is often accompanied by a higher percentage of body fat. This accumulation of subcutaneous fat can lead to thickening of the skin's layers, particularly the dermis. While the correlation between BMI and skin thickness is not absolute, as individual variations exist, the underlying physiological mechanisms provide a reasonable explanation for the moderate positive relationship observed between these two factors. The degree of this correlation can be influenced by age, genetics, and other individual characteristics. Additionally, these findings show that the dependent variable class moderate correlation with Glucose. This means that as diabetes progresses or worsens, there is a corresponding increase in glucose levels, but the relationship is not perfectly linear.

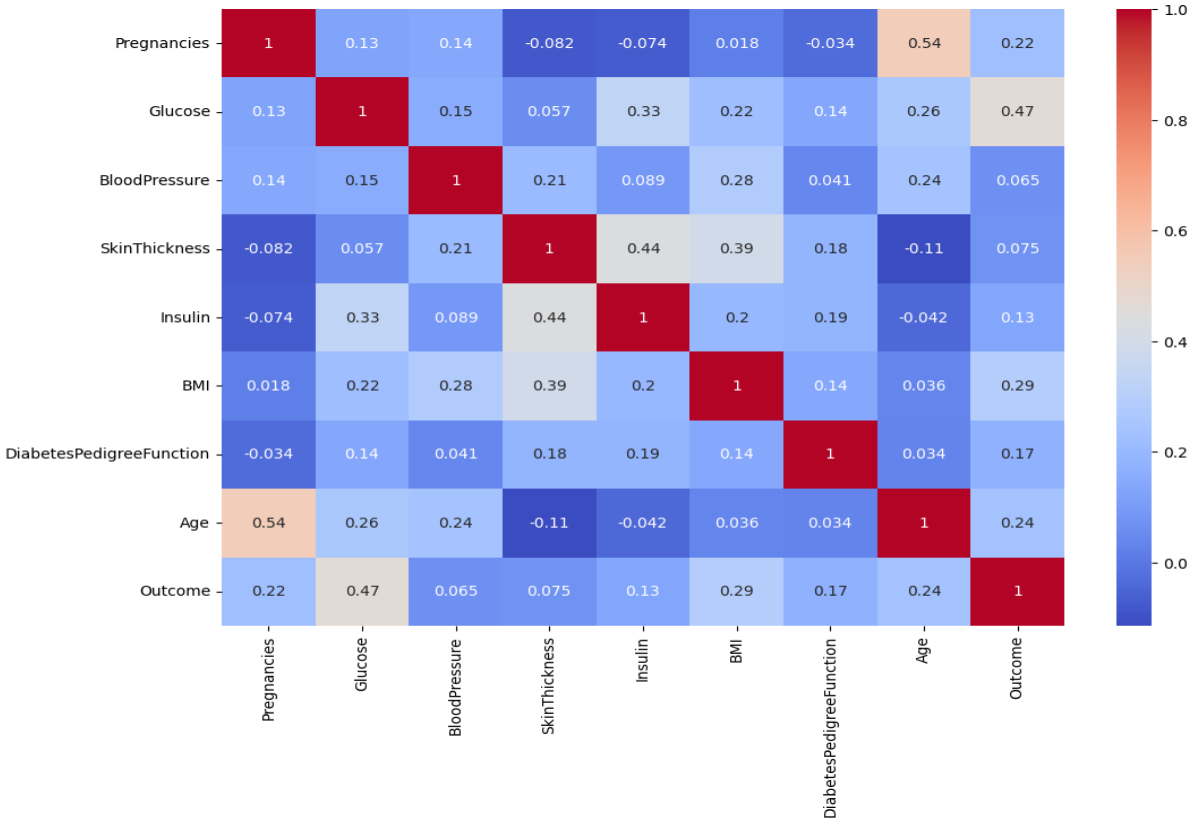

Figure 1. Correlation Heatmap

The findings show that blood pressure has a weak correlation, which means it might be less significant for prediction. High blood glucose levels are a primary indicator of diabetes, so this strong correlation is expected and reinforces the importance of glucose level as a predictor. If the glucose feature shows a high positive correlation with 'diabetes', it indicates that higher glucose stages are connected with an improved likelihood of diabetes. Additionally, a weak positive is found between glucose and insulin levels. A high positive correlation between glucose and insulin might indicate that these two features are related, possibly due to the body's response to regulating blood sugar levels. While both are important predictors, this could lead to multicollinearity, which might affect model performance if both are included without consideration. This suggests that factors such as blood pressure, and body mass index may have varying degrees of impact on diabetes, with other variables like blood age and glucose playing more substantial roles.

5.2 Proposed Improvements

The findings suggest that health institutions can target screening and intervention programs should be developed to identify high-risk individuals, particularly women with a history of gestational diabetes or multiple pregnancies. These programs can incorporate personalized risk assessment models that account for the complex interplay between variables like age, pregnancy history, glucose, and insulin levels. By incorporating these models into clinical decision-making tools, doctors can get help in suggesting the best strategies for preventing and managing diabetes early on, based on strong evidence and tailored to each patient's specific needs.

5.2 Validation

To assess the predictors, five models were taken into consideration in this study: the random forest regressor, XBoost, for easier training and evaluation, the researchers split the data into two sets: training and testing. This allows the models, including Naive Bayes, decision trees, and support vector machines, to learn from the training data and identify patterns that might predict diabetes. The accuracy of these models was then assessed using the testing data. To achieve this, the data was divided at a ratio of 80% for training and 20% for testing. We were able to gain important insights into the relative value of features by utilising the random forest regressor, which can help direct the model creation process (Figure 2).

```
Feature ranking:
1. feature 1 (0.24696871404503518)
2. feature 5 (0.18933566596650672)
3. feature 7 (0.1498736756376346)
4. feature 6 (0.1250979546042302)
5. feature 2 (0.08410171278913563)
6. feature 4 (0.07016243789046246)
7. feature 3 (0.06790966437884513)
8. feature 0 (0.0665501746881501)
```

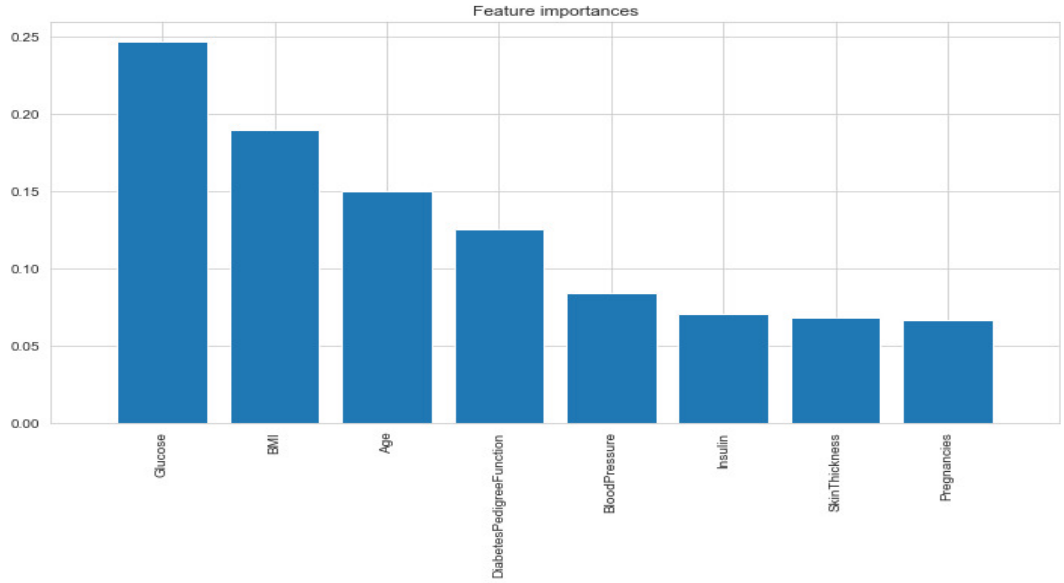

Figure 2. Feature importance for predicting diabetes

The findings show that glucose is a significant feature in predicting diabetes. Glucose levels are closely monitored in individuals with diabetes to evaluate the effectiveness of treatment interventions such as medication, diet, and exercise. Changes in glucose levels delivered an appreciated understanding of the success or failure of the treatment plan. Additionally, age and body mass index were also found to be significant variables in predicting diabetes. As individuals get older, their bodies may become less efficient in producing or utilizing insulin, leading to a higher likelihood of developing diabetes. Age also contributes to the cumulative effect of other risk factors and the overall metabolic changes that occur over time. Age and BMI also play roles in the progression of diabetes and the development of associated difficulties. Older age and higher BMI are connected with an advanced risk of adverse outcomes such as cardiovascular diseases, kidney disease, and diabetic retinopathy. Including age and BMI as features in predictive models helps capture the impact of these factors on the overall disease trajectory and prognosis.

5.3.3 Evaluation Results

Once a machine learning model is trained, it's important to evaluate its effectiveness. This process, called model evaluation, involves checking how well the model performs and if it meets the standards for the task. It measures the model's predictive accuracy, generalization capabilities, and overall effectiveness in solving the intended problem. Model assessment is crucial to regulate how well the prototype is performing and whether it meets the desired criteria for deployment or further improvement. The following graphs show the ROC curve for five different classifiers to compare their performance.

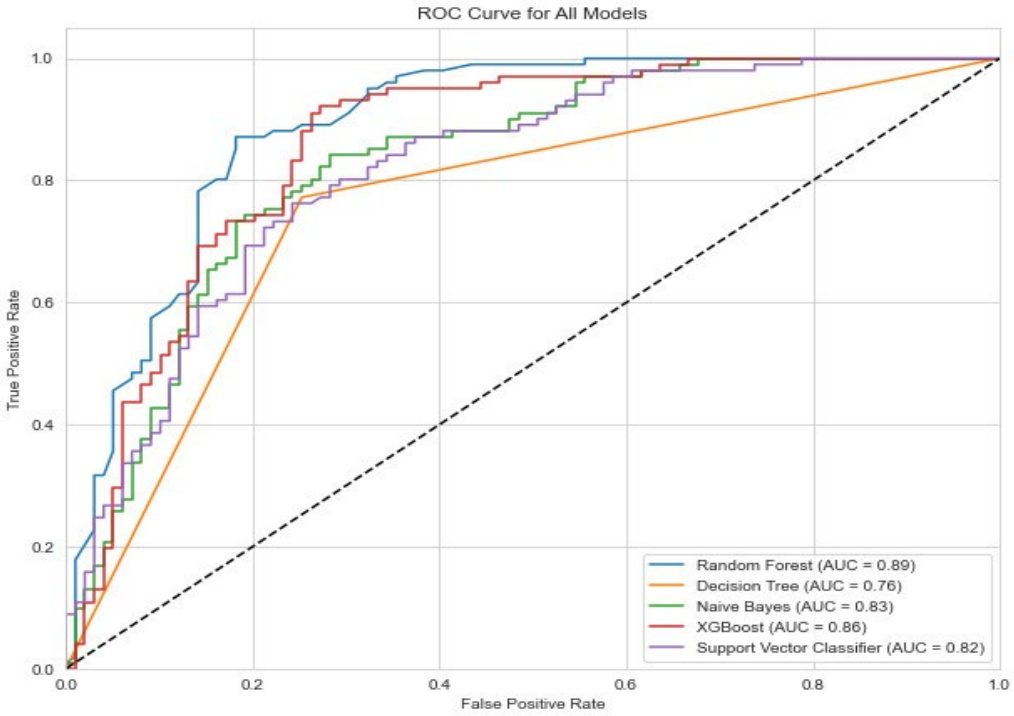

Figure 3. ROC curve for five different classifiers

The results show that random forest is out-competing the other classifiers with a higher area under the curve. Based on the results indicate that the random forest model performed better than other models at classifying data (Figure 3). This is because it achieved a higher area under the curve (AUC) score under the curve (AUC), which suggests that the random forest algorithm is well-suited for classifying whether the person is diabetic or not. AUC, or Area Under the Curve, is a popular way to measure how well a model distinguishes between two groups in a binary classification task. A higher AUC score means the model is better at differentiating between the positive and negative cases, indicating stronger overall performance.

Time complexities for optimal analysis of the model performances.

To provide an optimal analysis of model performances, several time complexities need to be considered. Here are the time complexities for common evaluation metrics used in model performance analysis: Time complexities may vary depending on the specific implementation, the size of the dataset, and the computational efficiency of the algorithms used. Additionally, other factors such as feature extraction, preprocessing, and model training time should also be considered for a comprehensive analysis of model performances.

```
model_results_df = pd.DataFrame(model_results)
print(model_results_df)
```

```
                                                    Random Forest  \
Classification Report                precision    recall  f1-score   ...
Confusion Matrix                                    [[77, 22], [12, 89]]
Training Time                                                0.448807

                                                    Decision Tree  \
Classification Report                precision    recall  f1-score   ...
Confusion Matrix                                    [[74, 25], [23, 78]]
Training Time                                                0.0155323

                                                      Naive Bayes  \
Classification Report                precision    recall  f1-score   ...
Confusion Matrix                                    [[78, 21], [25, 76]]
Training Time                                                0.0160494

                                                          XGBoost  \
Classification Report                precision    recall  f1-score   ...
Confusion Matrix                                    [[73, 26], [10, 91]]
Training Time                                                0.433631

                                                Support Vector Classifier
Classification Report                precision    recall  f1-score   ...
Confusion Matrix                                    [[74, 25], [24, 77]]
Training Time                                                0.293686
```

Figure 4. Time complexities for optimal analysis of the model performances.

It is important to note that reducing time complexities should not compromise the accuracy and reliability of the analysis (Figure 4). Accurate performance evaluation is crucial to ensure robust and trustworthy results. Therefore, results show that the process of training a random forest involves growing multiple decision trees simultaneously. However, the training time for individual decision trees and the subsequent combination of their predictions can still contribute to a longer overall runtime than the other four classifiers.

Classification reports the classifiers

Because they offer a thorough assessment of a machine learning classifier's performance, classification reports are significant findings in modelling. For every class in the classification job, they offer precise, recall, F1-score, and support metrics in detail.

```
Classification Report for Random Forest:

              precision    recall  f1-score   support

           0       0.87      0.78      0.82        99
           1       0.80      0.88      0.84       101

    accuracy                           0.83       200
   macro avg       0.83      0.83      0.83       200
weighted avg       0.83      0.83      0.83       200


------------------------------------------------------------


Classification Report for Decision Tree:

              precision    recall  f1-score   support

           0       0.76      0.75      0.76        99
           1       0.76      0.77      0.76       101

    accuracy                           0.76       200
   macro avg       0.76      0.76      0.76       200
weighted avg       0.76      0.76      0.76       200
```

Figure 5. Classification report for Random Forest and Decision Tree

These measures aid in evaluating the model's accuracy in classifying instances into various groups, spotting biases or imbalances in the predictions, and comprehending the trade-off between recall and precision (Figure 5).

The results show that the random forest model performed very well. It achieved a high F1 score, which means it effectively identified a large number of true diabetes cases while minimizing mistakes like labelling healthy people as diabetic (false positives) or missing actual cases (false negatives). This balance between accuracy and completeness is crucial. Compared to decision trees, the random forest classifier seems to be more consistent in correctly identifying cases across different types of diabetes (Figure 6).

```
Classification Report for Naive Bayes:

              precision    recall  f1-score   support

           0       0.76      0.79      0.77        99
           1       0.78      0.75      0.77       101

    accuracy                           0.77       200
   macro avg       0.77      0.77      0.77       200
weighted avg       0.77      0.77      0.77       200


------------------------------------------------------------


Classification Report for XGBoost:

              precision    recall  f1-score   support

           0       0.88      0.74      0.80        99
           1       0.78      0.90      0.83       101

    accuracy                           0.82       200
   macro avg       0.83      0.82      0.82       200
weighted avg       0.83      0.82      0.82       200
```

Figure 6. Classification report for Naïve Bayes and XBoost

The results indicate that XBoost outperforms other models in predicting whether a patient is diabetic or not, it suggests that XBoost achieves higher precision and recall compared to these classifiers. A higher precision in XBoost means that it has a lower rate of misclassifying non-diabetic patients as diabetic compared to the other classifiers. This indicates that when XBoost predicts a patient as diabetic, there is a higher likelihood that the patient is truly diabetic. For medical diagnosis to prevent needless treatments, more precision is essential or interventions for patients who are not diabetic. By outperforming other classifiers in terms of precision and recall, XBoost demonstrates its effectiveness in accurately classifying instances and predicting diabetes. However, it is essential to thoroughly analyse the specific dataset and carefully interpret the results to understand the strengths and limitations of XBoost in the given context.

```
Classification Report for Support Vector Classifier:

              precision    recall  f1-score   support

           0       0.76      0.75      0.75        99
           1       0.75      0.76      0.76       101

    accuracy                           0.76       200
   macro avg       0.76      0.75      0.75       200
weighted avg       0.76      0.76      0.75       200


------------------------------------------------------------
```

Figure 7. Classification report for Support Vector Classifier

The results obtained from comparing the F1 scores of the random forest classifier with other models indicate that the random forest classifier performs well in accurately classifying instances across all classes (Figure 7). The random forest model achieved a high F1-score, which is like a scorecard for how well it identifies diabetes. This score shows the model is good at finding true cases of diabetes (high proportion of true positives) while minimizing mistakes like labelling healthy people as diabetic (false positives) and missing actual cases (false negatives). This balanced approach makes the model a strong candidate for real-world use. This suggests that the random forest classifier outperforms the other classifiers in accurately predicting whether a patient has diabetes. The ensemble approach of the random forest algorithm, where multiple decision trees are combined, contributes to its enhanced performance and robustness. These findings highlight the effectiveness of the random forest classifier as a powerful tool for diabetes prediction compared to the alternative classifiers.

6. Conclusion

The objective of this reasearch was to detect patterns in diabetes data that can help to improve diabetes care. The analysis presented in the study offers crucial insights into the interplay of various factors influencing diabetes. The positive relationship between age and the number of pregnancies underscores the compounding effects of gestational diabetes history, pregnancy-induced insulin resistance, and age-related fluctuations in glucose metabolism, all of which contribute to the increased risk of type 2 diabetes over time. Additionally, the moderate positive correlation between BMI and skin thickness suggests that higher BMI is often accompanied by increased subcutaneous fat accumulation and skin thickening. Notably, glucose level demonstrates a strong positive correlation with the diabetes outcome, reinforcing its importance as a key predictor. However, the moderate positive correlation between glucose and insulin levels indicates potential multicollinearity, which should be considered when including both features in the predictive model.

Based on these findings, the study proposes that health institutions develop targeted screening and intervention programs, particularly for high-risk women with a history of gestational diabetes or multiple pregnancies. These programs should incorporate personalized risk assessment models that integrate variables like age, pregnancy history, glucose, and insulin levels to guide evidence-based prevention and management strategies. The validation of predictive models shows that the random forest classifier performs as a potent tool for diabetes prediction. Therefore, the random forest model is the preferred approach for diabetes prediction based on the given dataset and analysis, highlighting its potential as a powerful tool for accurate and reliable diabetes prediction in healthcare settings.

References

A. J., Priyadarsini, Dr. R. J., and Titus, Dr. S., Survey on Predictive Analysis of Diabetes Disease Using Machine Learning Algorithms. *International Journal of Computer Science and Mobile Computing*, *9*(10), 19–27, 2020. https://doi.org/10.47760/ijcsmc.2020.v09i10.003.

Arokiasamy, P., Salvi, S., & Selvamani, Y., Global Burden of Diabetes Mellitus. In I. Kickbusch, D. Ganten, & M. Moeti (Eds.), *Handbook of Global Health*, pp. 495–538., 2021. Springer International Publishing. https://doi.org/10.1007/978-3-030-45009-0_28

Carpinteiro, C., Lopes, J., Abelha, A, and Santos, M. F., A Comparative Study of Classification Algorithms for Early Detection of Diabetes. *Procedia Computer Science*, *220*, 868–873. 2023. https://doi.org/10.1016/j.procs.2023.03.117,

Daghistani, T., and Alshammari, R., Comparison of statistical logistic regression and randomforest machine learning techniques in predicting diabetes. *Journal of Advances in Information Technology*, *11*(2), 78–83, 2020. https://doi.org/10.12720/jait.11.2.78-83

Ernersson, Å., Mufunda, E., and Hjelm, K., Audit of essential knowledge of diabetes in patients with diabetes in Zimbabwe. *Pan African Medical Journal*, *45*., 2023. .https://doi.org/10.11604/pamj.2023.45.103.31770

Gona, P. N., Kimokoti, R. W., Gona, C. M., Ballout, S., Rao, S. R., Mapoma, C. C., Lo, J, and Mokdad, A. H., Changes in body mass index, obesity, and overweight in Southern Africa development countries, 1990 to 2019: Findings from the Global Burden of Disease, Injuries, and Risk Factors Study. *Obesity Science and Practice*, *7*(5), 509–524. 2021. https://doi.org/10.1002/osp4.519.

Hanlon, P., Fauré, I., Corcoran, N., Butterly, E., Lewsey, J., McAllister, D., and Mair, F. S., Frailty measurement, prevalence, incidence, and clinical implications in people with diabetes: a systematic review and study-level meta-analysis. *The Lancet Healthy Longevity*, *1*(3), e106–e116. 2020. https://doi.org/10.1016/S2666-7568(20)30014-3,

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., and Hasan, M., Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, 76516–76531. 2020. https://doi.org/10.1109/ACCESS.2020.2989857,

Maguraushe, K., & Ndayizigamiye, P., Towards a Smart Healthcare System for Non-Communicable Diseases (NCDs) Management: A Bibliometric Analysis. In M. Masinde, S. Möbs, & A. Bagula (Eds.), *Emerging Technologies for Developing Countries*, pp. 107–125, 2024. Springer Nature Switzerland

Mozaffarian, D., Dietary and policy priorities to reduce the global crises of obesity and diabetes. In *Nature Food* , Vol. 1, Issue 1, pp. 38–50, 2020. Springer Nature. https://doi.org/10.1038/s43016-019-0013-1

Mpofu, L., Ndlovu, B., Dube, S., Muduva, M., Jacqueline, F., and Maguraushe, K., Predictive Model for Hospital Readmission of Diabetic Patients. In 5th African conference on Industrial Engineering and Operations Management, South Africa, , 2024. https://doi.org/10.46254/AF05.20240252 .

Mureyi, D., Katena, N. A., and Monera-Penduka, T., Perceptions of diabetes patients and their caregivers regarding access to medicine in a severely constrained health system: A qualitative study in Harare, Zimbabwe. *PLOS Global Public Health*, *2*(3), e0000255-., 2020. https://doi.org/10.1371/journal.pgph.0000255

Mutunhu, B., Chipangura, B., and Twinomurinzi, H., Internet of Things in the Monitoring of Diabetes: A Systematic Review. *International Journal of Health Systems and Translational Medicine*, *2*(1), 1–20., 2022. https://doi.org/10.4018/ijhstm.300336.

Mutunhu Ndlovu, Belinda, and Mukura, Nyasha. " Performance Evaluation of Artificial Intelligence in Decision Support System for Heart Disease Risk Prediction". *4th Asia Pacific International Conference on Industrial Engineering and Operations Management*. 2023, https://doi.org/10.46254/AP04.20230043.

Mutunhu, B., Chipangura, B., and Singh, S., Towards a quantified-self technology conceptual framework for monitoring diabetes. *South African Journal for Science and Technology, 43(1), 69-84*. ,2024. https://doi.org/10.36303/SATNT.2024.43.1.970

Ong, K. L., Stafford, L. K., McLaughlin, S. A., Boyko, E. J., Vollset, S. E., Smith, A. E., Dalton, B. E., Duprey, J., Cruz, J. A., Hagins, H., Lindstedt, P. A., Aali, A., Abate, Y. H., Abate, M. D., Abbasian, M., Abbasi-Kangevari, Z., Abbasi-Kangevari, M., Abd ElHafeez, S., Abd-Rabu, R., … Vos, T., Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet*, *402*(10397), 203–234. , 2023.. https://doi.org/10.1016/S0140-6736(23)01301-6

Rama Krishnaiah, K., and Prasad, A. H., Enhancing Health Data Prediction with Software Engineering and Machine Learning: An Application for Health Systems. In *Turkish Journal of Computer and Mathematics Education* (Vol. 13, Issue 03), 2022.

Sharma, R., Singh, S. N., and Khatri, S., Medical data mining using different classification and clustering techniques: A critical survey. *Proceedings - 2016 2nd International Conference on Computational Intelligence and Communication Technology, CICT 2016*, 687–691, 2016. https://doi.org/10.1109/CICT.2016.142

Song, X., Liu, X., Liu, F., & Wang, C., Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. In *International Journal of Medical Informatics* (Vol. 151). Elsevier Ireland Ltd. https://doi.org/10.1016/j.ijmedinf.2021.104484, 2021.

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., and Müller, K. R., Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. *Machine Learning and Knowledge Extraction*, *3*(2), 392–413. 2021.https://doi.org/10.3390/make3020020,

Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., Stein, C., Basit, A., Chan, J. C. N., Mbanya, J. C., Pavkov, M. E., Ramachandaran, A., Wild, S. H., James, S., Herman, W. H., Zhang, P., Bommer, C., Kuo, S., Boyko, E. J., and Magliano, D. J., IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, *183,* 2022. https://doi.org/10.1016/j.diabres.2021.109119,

Tripathi, D., Biswas, S. K., Reshmi, S., Boruah, A. N., and Purkayastha, B., Diabetes Prediction Using Machine Learning Analytics: Ensemble Learning Techniques. *2022 2nd Asian Conference on Innovation in Technology, ASIANCON 2022*. https://doi.org/10.1109/ASIANCON55314.2022.9908975.

Wou, C., Unwin, N., Huang, Y., and Roglic, G., Implications of the growing burden of diabetes for premature cardiovascular disease mortality and the attainment of the Sustainable Development Goal target 3.4. In *Cardiovascular Diagnosis and Therapy* (Vol. 9, Issue 2, pp. 140–149). 2019. AME Publishing Company. https://doi.org/10.21037/cdt.2018.09.04,

Zhu, T., Li, K., Herrero, P., and Georgiou, P., Deep Learning for Diabetes: A Systematic Review. In *IEEE Journal of Biomedical and Health Informatics* (Vol. 25, Issue 7, pp. 2744–2757), 2021. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/JBHI.2020.3040225

Zia, A., Aziz, M., Popa, I., Khan, S. A., Hamedani, A. F., and Asif, A. R. Artificial Intelligence-Based Medical Data Mining. In *Journal of Personalized Medicine* (Vol. 12, Issue 9). MDPI. https://doi.org/10.3390/jpm12091359, 2022.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H., Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics*, *9*. , 2018. https://doi.org/10.3389/fgene.2018.00515

Biographies

Isaac Murere is a postgraduate student in Big Data at the National University of Science and Technology (NUST) in Bulawayo, Zimbabwe. He is currently a Research Consultant, He holds a Bachelor of Science Honors Degree in Statistics and Operations Research from the National University of Science and Technology (2020) and Executive Certificate Monitoring Evaluation from the University of Zimbabwe (2022).

Belinda Ndlovu is a Ph.D. in Information Systems student at UNISA. She holds an MSc in Information Systems and a BSc in Computer Science. She is a seasoned software developer and academic. She has published several papers in the fields of Data Analytics, Health Informatics, ICT4D, and 4IR.

Sibusisiwe Dube is an experienced lecturer of Information Systems and Computer Science courses. She holds a PhD in Information Systems, an MSc in Computer Science, and a BSc in Information Systems. She has been lecturing since 2004. She is also an active researcher and supervisor of Postgraduate dissertations and undergraduate student projects.

Martin Muduva is currently pursuing his PhD at the Chinhoyi University of Technology, Martin is also concurrently engaged in Master's programs in Innovation and Entrepreneurship at the Bindura University of Science Education and Mechatronics and Artificial Intelligence at the University of Zimbabwe. He holds a Master's degree in Leadership and Corporate Governance from Bindura University of Science Education, along with expertise in Big Data Analytics and Information Security. Martin's commitment to education and professional development is evident through his Certificate in Higher and Tertiary Education. His multidisciplinary background and dedication make him a valuable asset in technology, innovation, and corporate governance.

Dr. Fungai Jacqueline Kiwa holds a Doctor of Philosophy Degree in Cultural Heritage and Information Technology from Chinhoyi University of Technology, complemented by a Post Graduate Diploma in Higher Education. Additionally, she holds a Master of Science degree in Information Systems, a Bachelor of Technology (Honors) degree in Computing and Information Technology, and an Advanced Diploma in Computing and Information Technology.

With a wealth of academic achievements, Dr. Kiwa has authored 17 publications, encompassing articles, conferences, thesis, and dissertations. Currently, she is a candidate for a Master of Mechatronics and AI at the University of Zimbabwe. Her expertise extends to Artificial Intelligence, creative IoT framework designing, and intensive programming skills, particularly in Java, Python, and C++.