Predicting Diabetes: Identifying Key Risk Factors Using Machine Learning Models

CSC3031 - Applied Data Science

Student Number - 720017170

Table of contents

Link to GitHub Repository = https://github.com/JoshLG18/Diabetes-Prediction

1. Introduction

Diabetes is a major global health issue with widespread implications for healthcare systems (Yang, Liu and Zhao, 2024). It can lead to medical complications, including cardiovascular disease, kidney failure, and neuropathy. Due to the increasing instances of diabetes, there is a need for reliable early detection methods to inform preventative healthcare strategies (Hoppe, Muller and Becker, 2024). Understanding the most effective predictors of diabetes can help healthcare practitioners prioritize resources for high-risk individuals. Furthermore, there has been research showing that many chronic conditions share common risk factors (Da Silva, Martins and Sousa, 2024). Finding out these common predictors can help improve management techniques.

This study aims to identify key predictors of diabetes by comparing three machine learning models; logistic regression (LR), Random Forest (RF), and support vector machines (SVM). Each model will complete different objectives; LR for interpreting predictor significance and calculating odds ratios, RF for identifying the most important factors, and SVM for identifying non-linear relationships and evaluating the performance of the prediction.

The analysis is expected to identify the most important predictive factors for diabetes and evaluate each model's classification accuracy. Conclusions from this study may contribute to the development of intervention strategies for diabetes and support broader healthcare goals to decrease the condition's impact.

2. Methods

The aim of this study is to determine the most significant predictors of diabetes and how accurately the given factors predict diabetes using Python.

The research questions are:

1. Which factors are the strongest predictors of diabetes?
2. How accurately can these factors classify diabetes status?
3. Which model most accurately classifies diabetes status?

The first step of this research was to prepare the data for any analysis that was to be conducted. This involved handling missing valued by replacing them with means for the variable to maintain the size of the dataset as 49% of the rows has missing values. Continuous variables were normalised to ensure comparability across the predictors. The next step of the study was exploratory data analysis (EDA). This allows an understanding of the data set and the identification of any potential patterns or outliers. Descriptive statistics were calculated to describe the data e.g. mean, median, and standard deviation, and distributions were assessed using bar charts showing the distribution of diabetes status grouped by key predictors.

Three models were selected to assess the research questions. LR was used to provide odds ratios that can quantify the relationship between predictors and diabetes risk. RFs allow the importance of the predictors to be ranked. SVMs evaluate non-linear relationships in the data and how well the predictors can classify between Diabetic and Non-Diabetic. The dataset was split into a train and test set with 80% of the data used for training and 20% used for testing.

To evaluate each of the model's performance a variety of metrics were used. Accuracy was calculated to measure the overall proportion of correctly classified individuals. Precision was calculated to show the proportion of predicted positive cases that were correct, and recall was used as it assesses the proportion of actual positive cases that were identified correctly. An F1 score was also calculated which is the mean of both precision and recall. When evaluating the importance of the predictor odds ratios were calculated within LR to quantify the effect each variable has on the risk of diabetes. Within the RF Model, feature importance scores were generated which allows ranking of the predictors based on their contribution to the model's accuracy.

3. Data

The dataset used in this project was sourced online from Kaggle (Rahman, 2024). This dataset provided medical attributes from female patients along with an outcome variable that indicates whether the patient has diabetes or not. The attributes that were collected were; The number of times the patient had been pregnant (Pregnancies), Plasma glucose concentration after a 2-hour oral glucose tolerance test (Glucose), Diastolic blood pressure in mm Hg (BloodPressure), Tricep skinfold thickness in mm (SkinThickness), 2-hour serum insulin in mu U/ml (Insulin), Body mass index (BMI), The likelihood of diabetes based on family history (DiabetesPedigreeFunction), and age of the patient in years (Age).

To prepare this dataset for the analysis, missing values required handling, and continuous variables needed to be scaled to standardize their ranges. This was done by replacing all missing values with the mean for the variable. Scaling was completed by performing z-score normalization which transformed each value into the number of standard deviations away from the mean.

Table 1: Summary Statistics

Variable	N	Mean	Median	SD	Min	Max
Pregnancies	768.0	3.8	3.0	3.4	0.0	17.0
Glucose	768.0	121.7	117.0	30.4	44.0	199.0
BloodPressure	768.0	72.4	72.2	12.1	24.0	122.0
SkinThickness	768.0	29.2	29.2	8.8	7.0	99.0
Insulin	768.0	155.5	155.5	85.0	14.0	846.0
BMI	768.0	32.5	32.4	6.9	18.2	67.1
DiabetesPedigreeFunction	768.0	0.5	0.4	0.3	0.1	2.4
Age	768.0	33.2	29.0	11.8	21.0	81.0

3.1 Descriptive Statistics

Table 1 presents the summary statistics of the key variables used in the analysis of diabetes prediction. The dataset comprises 768 instances, with variables such as Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age.

Glucose shows a mean of 121.7 and a standard deviation of 30.4, with values ranging from 44.0 to 199.0. This wide range highlights the variability in glucose levels within the dataset. BMI has a mean of 32.5, with a standard deviation of 6.9, and values between 18.2 and 67.1. The Insulin variable exhibits a high standard deviation of 85.0, with a mean of 155.5, reflecting the substantial variability in insulin levels across the sample. The Age variable shows a mean of 33.2 and a standard deviation of 11.8, with values ranging from 21.0 to 81.0. The DiabetesPedigreeFunction, which represents a family history of diabetes, has a mean of 0.5, suggesting that, on average, this factor is somewhat neutral in its influence, with values spanning from 0.1 to 2.4.

These summary statistics provide an overview of the data distribution and variability, helping inform subsequent analyses and model predictions. The large spread in variables like Insulin and Glucose suggests that diabetes prediction models could benefit from understanding how these variables interact with each other and the outcome variable.

3.2 Correlation Analysis

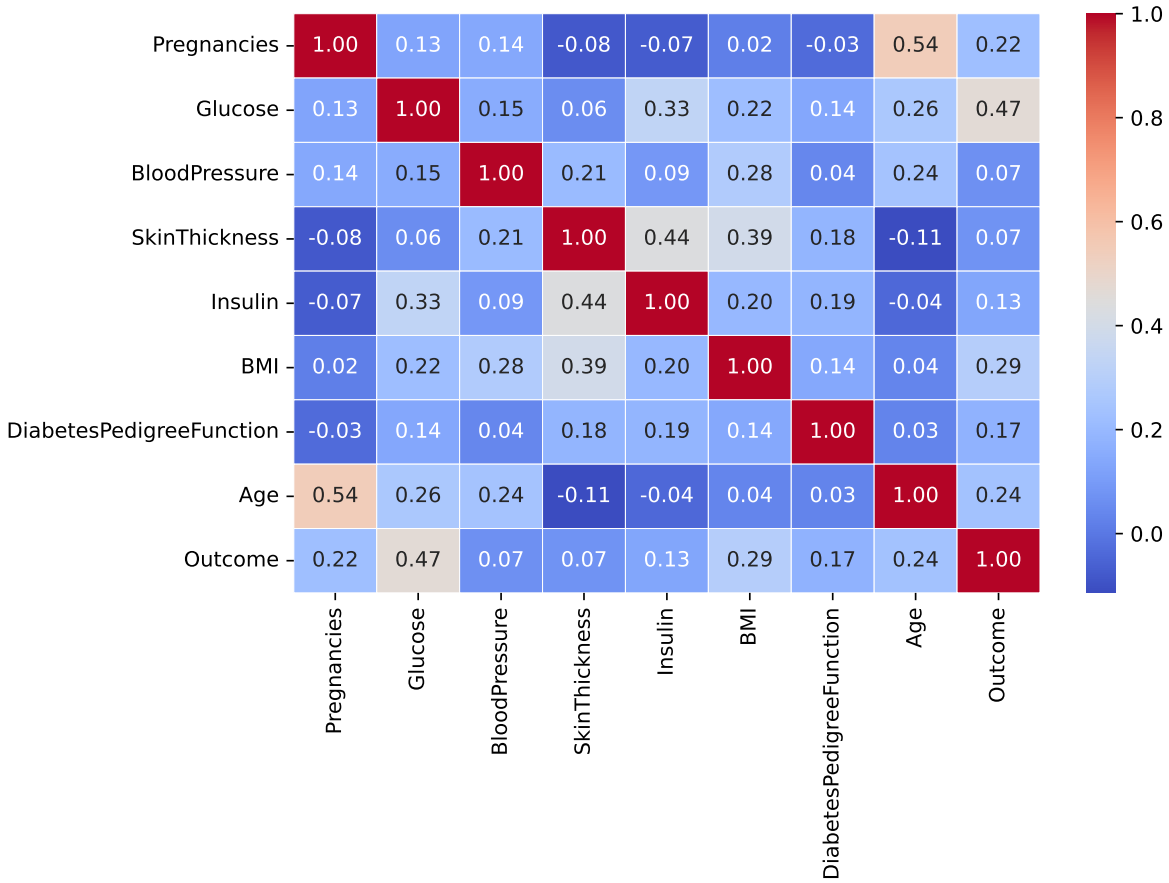

Figure 1: Correlation Matrix for Diabetes Dataset

In the correlation analysis, several key predictors of diabetes are evaluated, providing insights into how strongly each factor correlates with the outcome variable, "diabetes." The correlation matrix reveals a notable positive correlation between Glucose ($r = 0.47$) and BMI ($r = 0.29$) with diabetes outcome, showing that as glucose and BMI increase, so does the likelihood of diabetes. Notably, Age ($r = 0.24$) and Pregnancies ($r = 0.22$) are also correlated with diabetes, although to a lesser extent.

The correlation between Insulin ($r = 0.13$) and diabetes, while positive, is weaker compared to other predictors. Variables like SkinThickness, BloodPressure, and DiabetesPedigreeFunction show low correlation values with the outcome, suggesting their limited direct relationship with diabetes in the current dataset. These insights reflect the important role of glucose and BMI as strong predictors of diabetes, which are consistent with existing literature. The findings highlight the potential of glucose and BMI in developing predictive models for diabetes, while

also indicating the need for additional variables and enhanced modeling techniques for better prediction.

3.3 Distribution Analysis

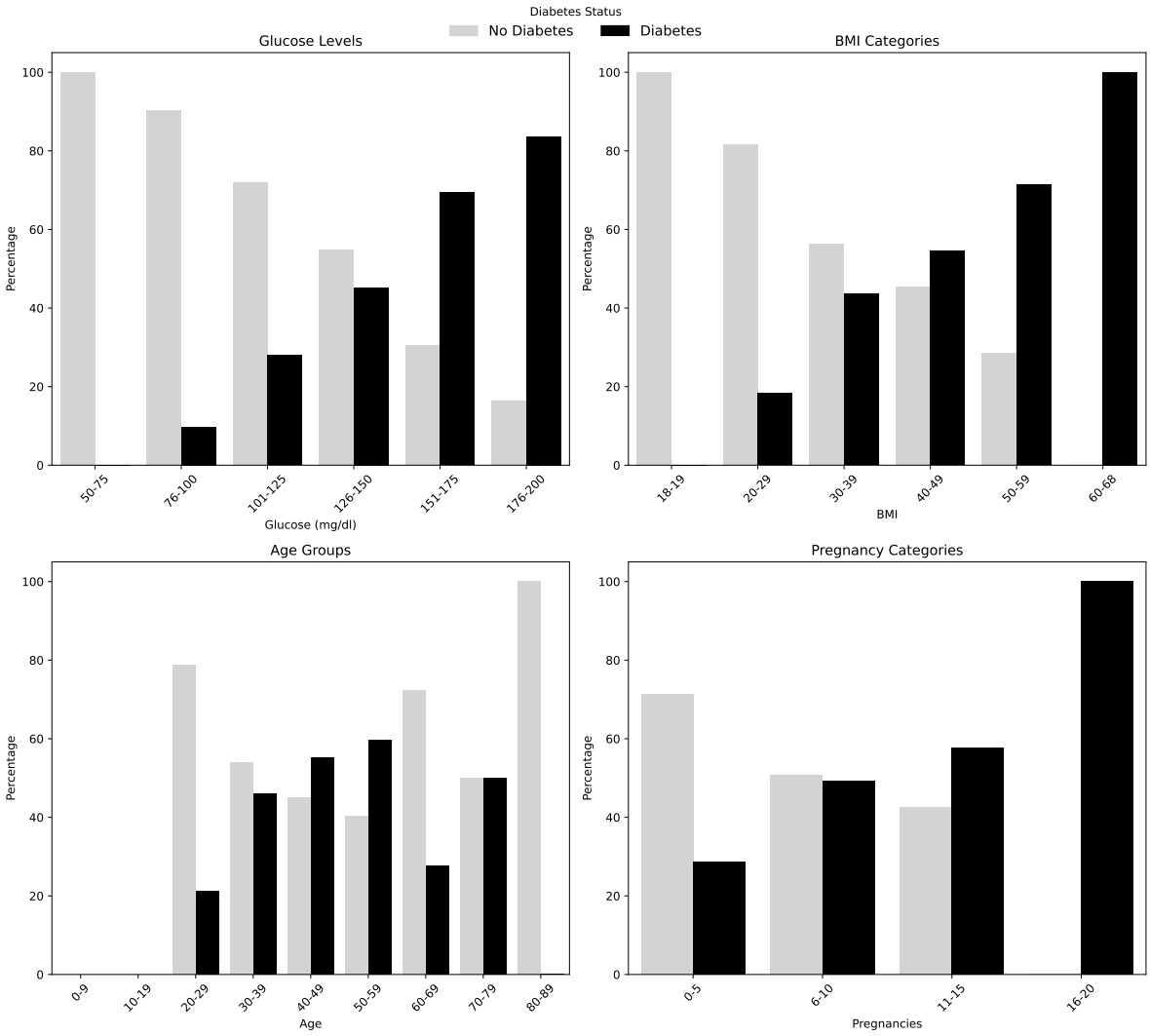

Figure 2: Distribution of Diabetes Status Across Key Factors

Figure 2 shows the percentage of diabetes cases across four key factors; age, BMI, glucose, and pregnancies. All factors show that with increases in the factor, there is an increase in the percentage of cases that have diabetes. This aligns with existing literature. Age has been shown to be a significant factor, as older individuals are at higher risk of developing diabetes due to declining pancreatic function and increased insulin resistance (Dominguez and Gonnelli,

2024). BMI is correlated with diabetes risk, with obesity being a major contributor to insulin resistance and type 2 diabetes development (Guh *et al.*, 2009). Elevated glucose levels are central to diabetes diagnosis, as they reflect poor glucose metabolism (American Diabetes Association, 2021). Finally, high insulin levels indicate insulin resistance, which is a hallmark of diabetes progression (DeFronzo, 2005).

4. Results and Discussion

4.1 Random Forest Model

The RF model trained in this study highlights glucose, age, and insulin as the most significant predictors of diabetes, illustrated in Figure 3. These findings are consistent with existing literature, reinforcing the understanding of these variables as critical indicators in diabetes diagnosis.

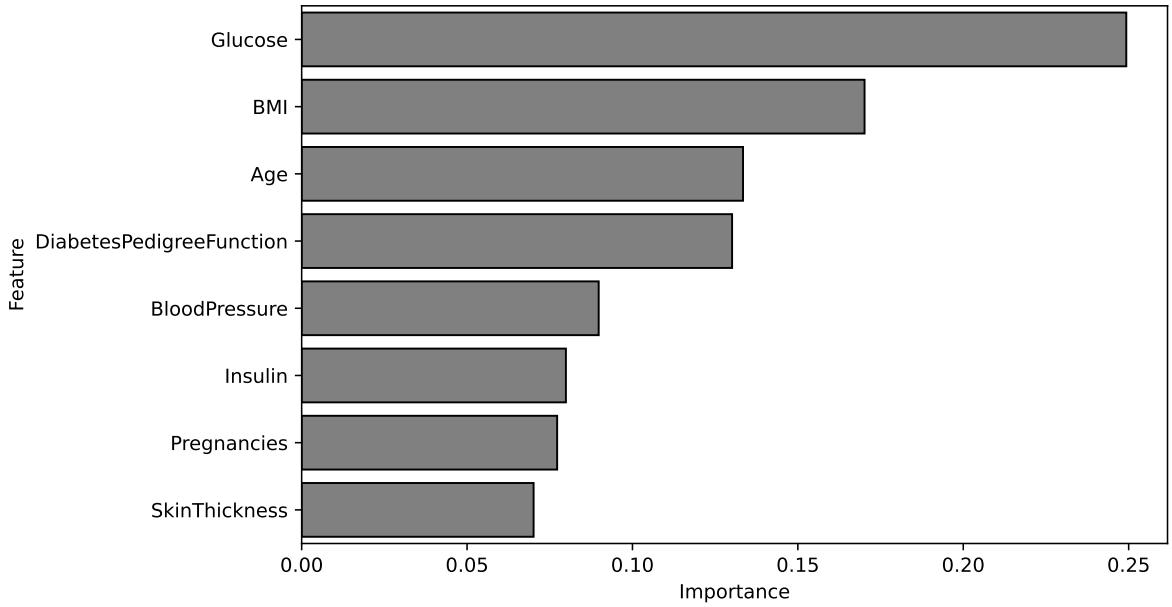

Figure 3: Importance of Features Based on Mean Decrease in Accuracy

Glucose being the most significant predictor of diabetes is consistent with prior research which shows serum glucose levels are often used for diagnosing diabetes due to their role in metabolic processes. This is proven by clinical diagnostic tools like fasting plasma glucose and oral glucose tolerance tests which use glucose as the primary marker (American Diabetes Association, 2021). BMI was shown to be the second most important factor within our model, this is supported a study by Moghaddam *et al.* (2024) showed it to be the most important factor within its RF model showing that BMI is a very important predictor of diabetes.

On the other hand, a factor that is usually identified to be large contributors to diabetes; body fat, measured by skin thickness, showed lower importance scores in this model. This could be attributed to multicollinearity with glucose since glucose levels may mediate their relationship with diabetes (Wang *et al.*, 2024). Blood pressure is often associated with diabetes shown by a study by Noroozi, Azizi and Khani (2024) which showed that high and fluctuating blood pressure are significant contributors to diabetes. However, this was indicated to have the lowest importance score in this model.

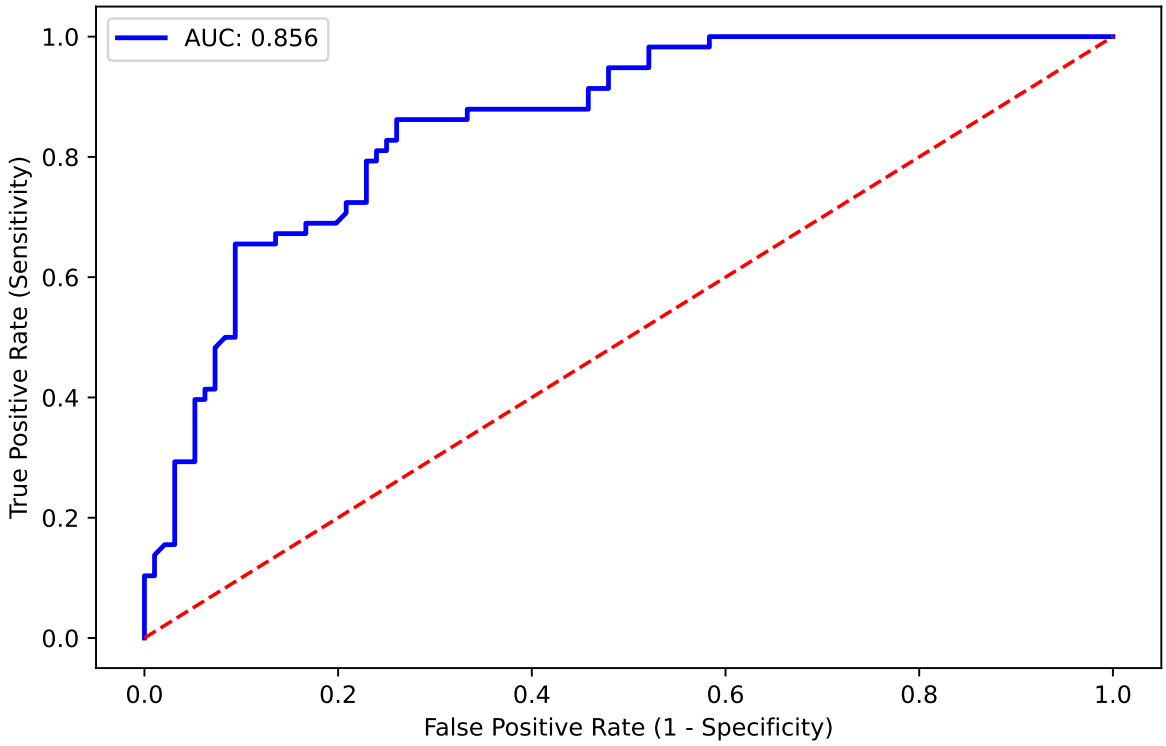

Figure 4: ROC Curve for Random Forest Model

The ROC curve (Figure 4) for the RF model assesses its classification performance in identifying diabetic and non-diabetic individuals. The model achieved an AUC of 0.856, demonstrating strong predictive power by accurately distinguishing between the two classes 85.6% of the time. The sharp rise at lower false positive rates indicates good sensitivity, meaning the RF model successfully identifies most true diabetic cases while keeping false positives at a minimum. However, as the curve flattens, there is a trade-off between sensitivity and specificity, suggesting that the model achieves a balance between detecting diabetic cases and minimizing misclassifications. This balance between sensitivity and specificity is further supported by the model's F1 score of 0.827, indicating a good harmonic mean between precision and recall.

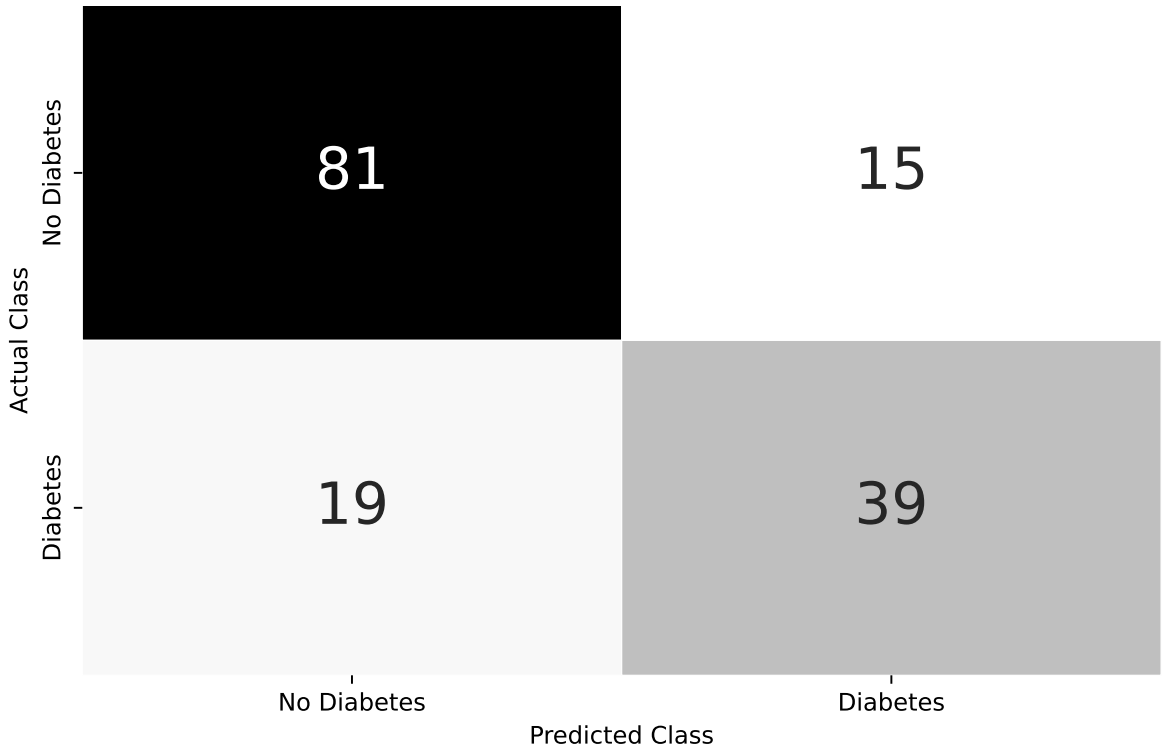

Figure 5: Confusion Matrix for Random Forest Model

The confusion matrix (Figure 5) for the RF model provides a detailed comparison of actual versus predicted diabetes status. In this case, the model correctly predicted "No Diabetes" for 81 instances (True Negatives), and correctly identified "Diabetes" for 39 instances (True Positives). However, there were 15 False Positives, where the model incorrectly predicted "Diabetes" when the actual class was "No Diabetes," and 19 False Negatives, where the model missed predicting "Diabetes" in actual "Diabetes" cases.

4.2 Logistic Regression Model

The LR model shows a quantitative understanding of how the key predictors affect diabetes risk. Odds ratios were calculated allowing an easy interpretation of the relationships between the factors and diabetes risk. The odds ratio indicates the increase in the risk of diabetes for a one-unit increase in that variable.

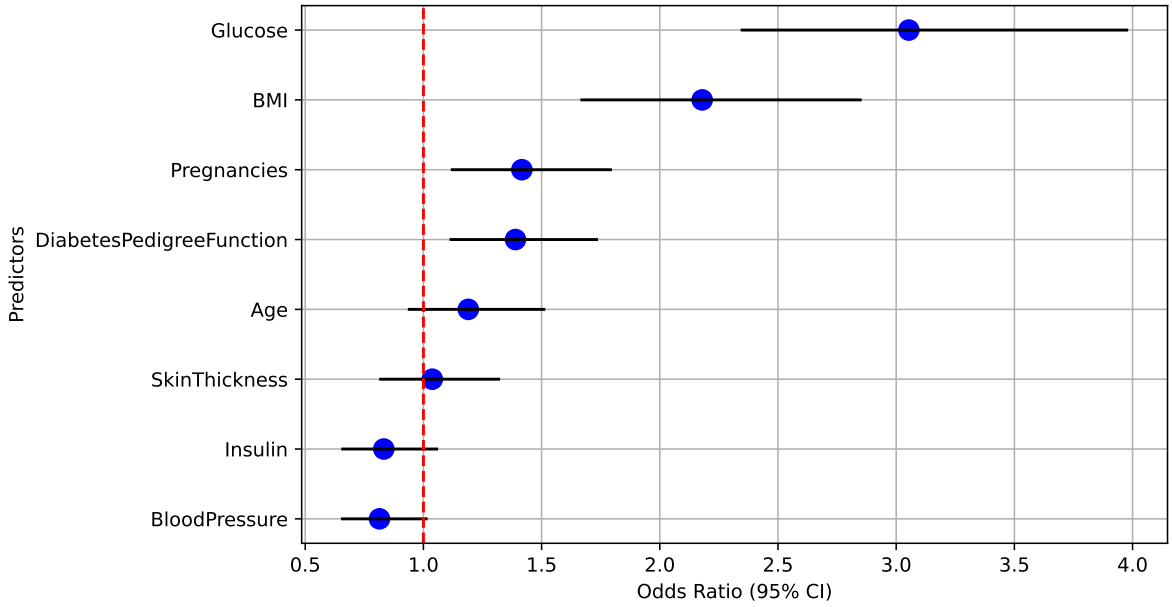

Figure 6: Forest Plot of Odds Ratios from Logistic Regression

The model was similar to the RF model as glucose and BMI were also identified as the two most important predictors of diabetes (Figure 6). The odds ratio of glucose shows a strong, positive association with diabetes risk, with an odds ratio of 3.2 indicating that with a one-unit increase in 2-hour serum glucose levels, the risk of diabetes increases by 320%. These findings are supported by literature which states that glucose has a central role in hyperglycemia and insulin resistance (Lu *et al.*, 2024). Flowers, Martinez and Turner (2024) also supports these findings, indicating the importance of glucose control in reducing diabetes risk.

The second most important predictor BMI was shown to increase diabetes risk by 220% with a one-unit increase. Literature supports this by showing that obesity is a contributing factor to insulin resistance by disrupting the body's glucose metabolism (Torres-Torres *et al.*, 2024). The findings from this model are like prior research and highlight the need to address obesity in public health interventions in preventing diabetes.

This model identified pregnancies as the third most influential factor in diabetes risk, contradicting the LR model which shows it as fifth. Pregnancies showed an odds ratio of 1.49, indicating that with an increase of one pregnancy, the risk of diabetes increases by 49%. Literature supports these conclusions as a meta-analysis reported that women with a history of

gestational diabetes, a common complication within pregnancies, have a significantly increased risk of developing type 2 diabetes later in life (Diaz-Santana *et al.*, 2022).

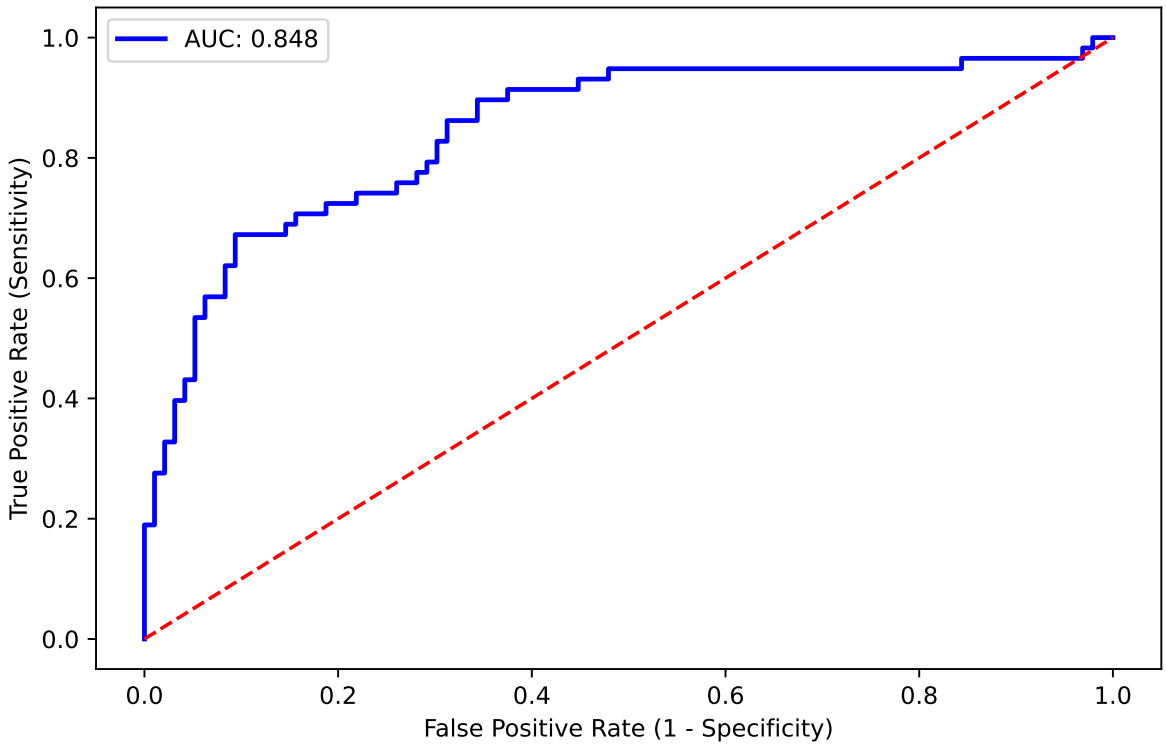

Figure 7: ROC Curve for LR Model

The ROC curve (Figure 7) for the LR model showcases its classification ability in distinguishing between diabetic and non-diabetic individuals. The model achieved a solid AUC of 0.848, indicating strong predictive capability. While it performs well, it slightly lags behind the RF model (AUC = 0.856), suggesting that more complex models offer a slight edge in prediction.

The curve demonstrates a steep rise at lower false positive rates, indicating good sensitivity and the model's ability to correctly identify diabetes cases. However, its overall performance is slightly less optimal compared to the more complex models. Improvements through techniques such as feature selection, regularization, or adding interaction terms could enhance its predictive performance. LR remains a strong baseline model in this study due to its balance of interpretability and stability, making it a reliable tool for classification tasks in healthcare.

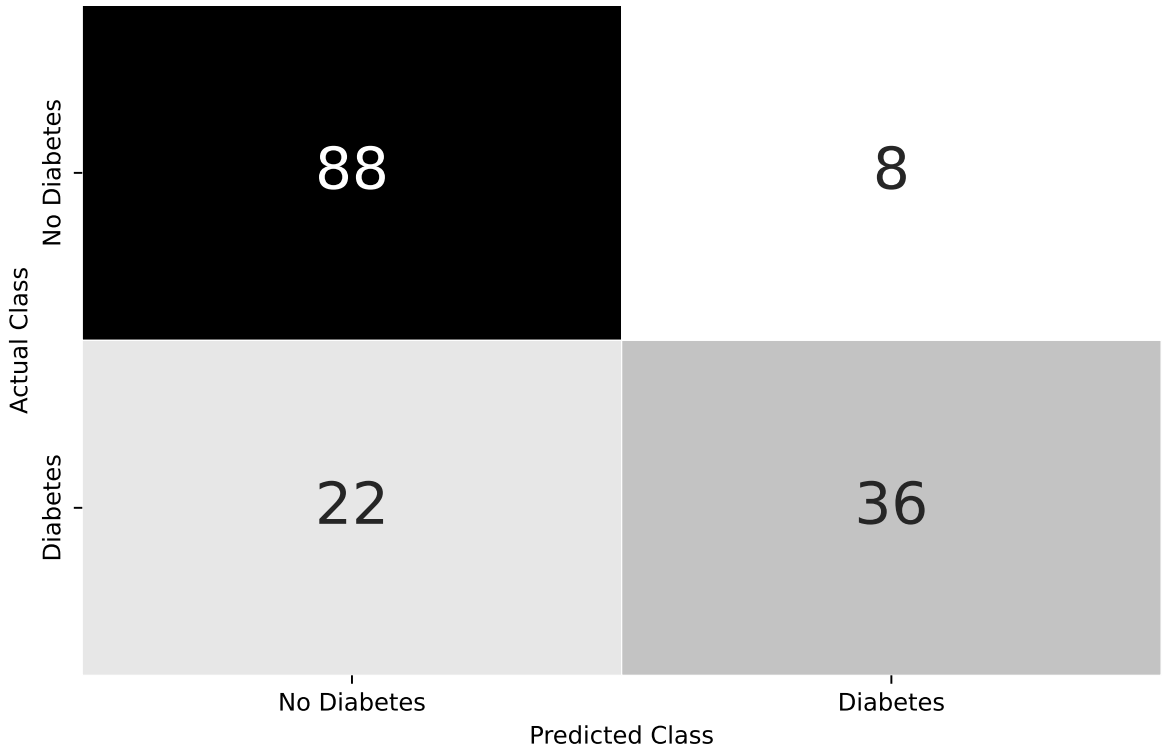

Figure 8: Confusion Matrix for LR Model

The confusion matrix (Figure 8) for the LR model provides another detailed comparison of actual versus predicted diabetes status. In this case, the model correctly predicted "No Diabetes" for 88 instances (True Negatives), and correctly identified "Diabetes" for 36 instances (True Positives). However, there were 22 False Positives, where the model incorrectly predicted "Diabetes" when the actual class was "No Diabetes," and 8 False Negatives, where the model missed predicting "Diabetes" in actual "Diabetes" cases.

Compared to the RF model, the LR model demonstrates an improvement in correctly predicting "No Diabetes," with fewer False Positives (22 compared to 25 in RF). It also has a higher number of True Positives (36 compared to 33 in RF). However, the LR model still has a trade-off, with more False Negatives than RF (8 compared to 15 in RF).

4.3 Support Vector Machine Model

The final model that was implemented was an SVM model which was used to evaluate the effectiveness of using the factors for predicting diabetes.

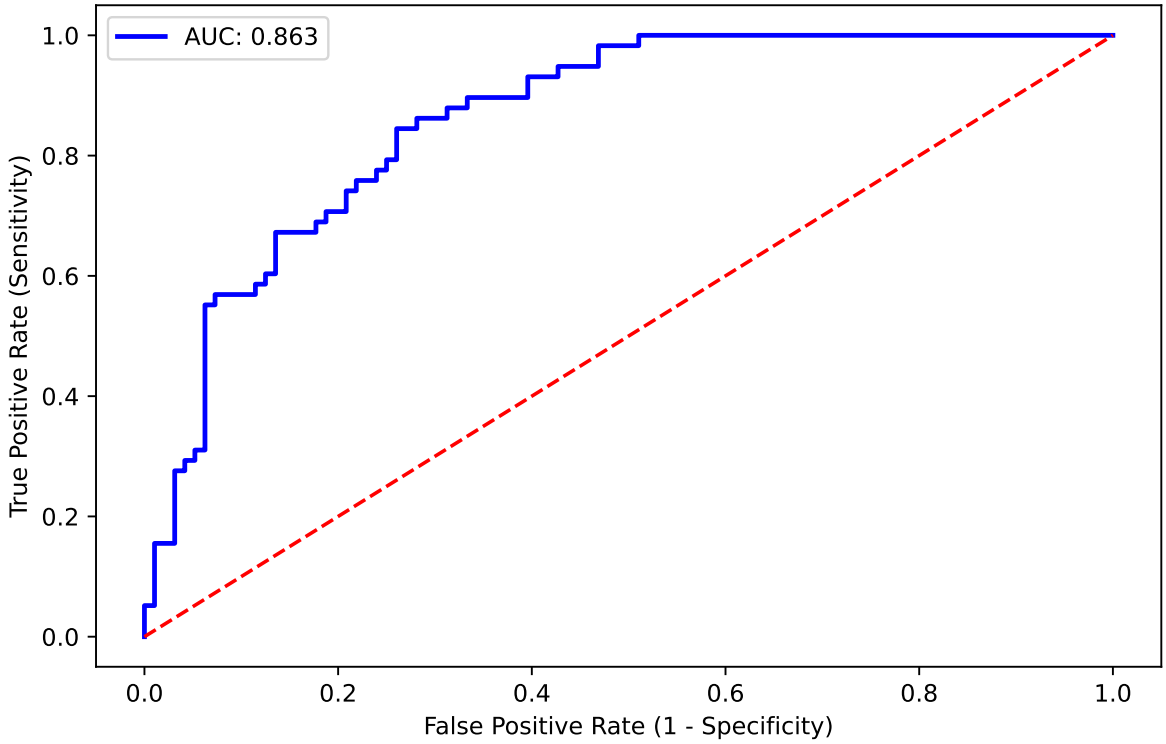

Figure 9: ROC Curve for SVM Model

The ROC curve (Figure 9) for the SVM model provides a visual representation of the model's classification performance across various threshold values. It compares the true positive rate (sensitivity) against the false positive rate (1 - specificity). In this case, the SVM model achieves an impressive AUC of 0.863, indicating robust predictive power. This value surpasses that of both LR (AUC = 0.848) and Random Forest (AUC = 0.856), which suggests that SVM outperforms these models for the given dataset.

The steep increase in the curve at low false positive rates suggests that the model effectively classifies diabetic cases with a relatively low number of misclassifications. This means it is good at identifying actual diabetes cases while minimizing false positives. However, further improvements can be achieved by tuning hyperparameters, selecting features, or experimenting with alternative kernels. All things considered, SVM demonstrates the best classification performance in this study, showing its potential for improved diabetic prediction models.

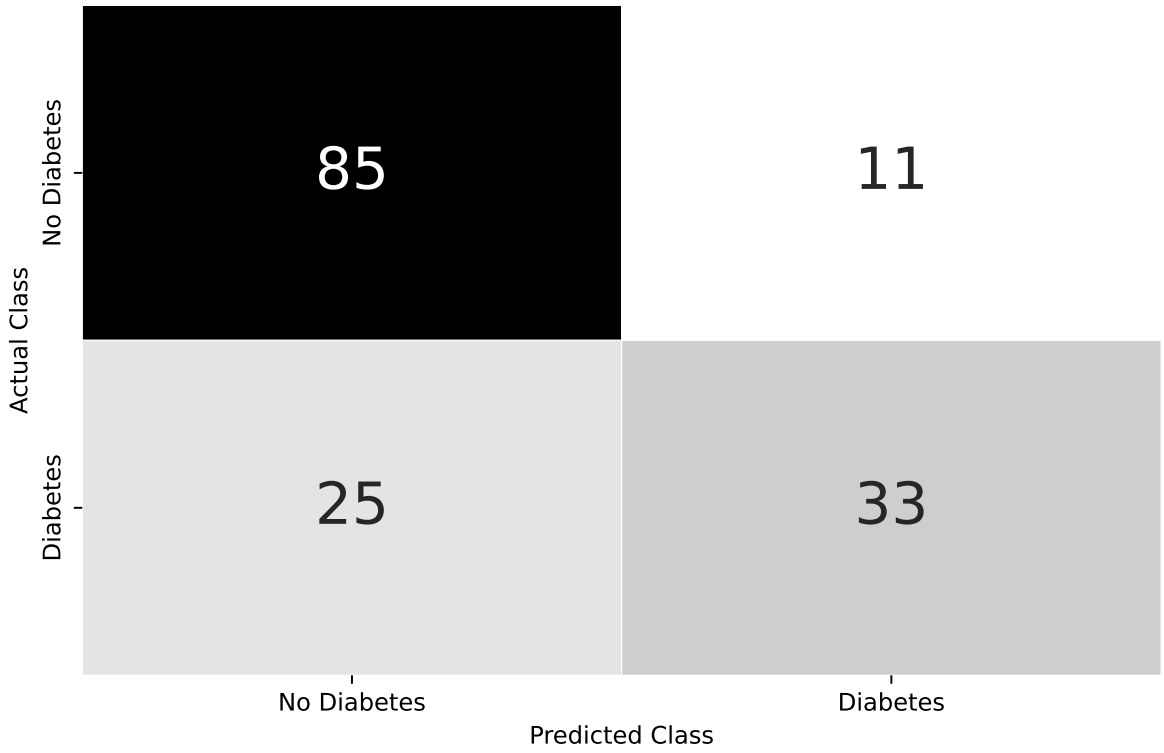

Figure 10: Confusion Matrix for SVM Model

The confusion matrix (Figure 10) for the SVM model shows a good performance with 88 true negatives (correctly predicting "No Diabetes") and 36 true positives (correctly predicting "Diabetes"). However, it still has 22 false positives (where it predicted "Diabetes" but the true class was "No Diabetes") and 8 false negatives (where it predicted "No Diabetes" but the true class was "Diabetes").

When comparing the SVM model to the RF and LR models, the SVM model stands out for its accuracy in correctly identifying "No Diabetes," with 88 true negatives, a significant improvement over the 81 in the RF model. However, it still has a higher number of false positives than the RF model, which had 15. In terms of true positives, the SVM model performs better than the RF model, with 36 true positives compared to 33 in RF. When compared to LR, the SVM model outperforms in both true negatives and true positives, with LR having 85 true negatives and 33 true positives. Therefore, the SVM model offers a more balanced and accurate prediction compared to RF and LR.

4.4 Model Evaluation and Comparisons

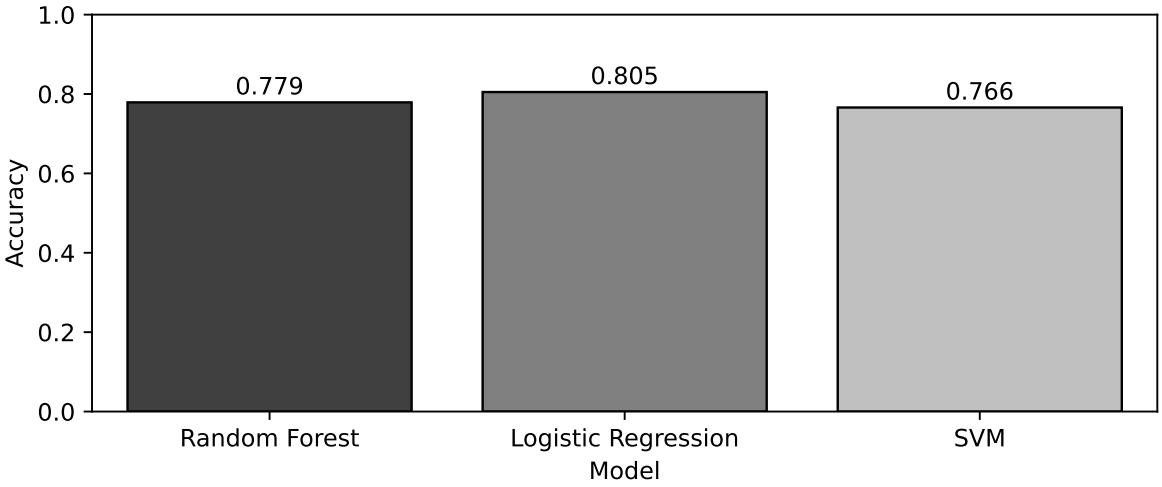

Figure ?: Accuracy for Each Model

Figure 11 shows the accuracy of each model in classifying diabetes status. In this study, LR shows the highest accuracy of 0.805 indicating that it correctly classified 81% of the cases. This model performed better than the RF and SVM which achieved accuracies of 0.779 and 0.766 respectively, indicating the need for tuning and larger samples to improve the accuracy of these more complex models and LR's simplicity could explain its increase in performance, however LR has the lowest recall and AUC values, indicating that SVMs and RFs are more applicable to diabetes onset prediction. These models are outperformed by a study by Akula, Nguyen and Garibay (2019) which showed accuracy scores of 0.81 for an SVM model and 0.82 RF model using similar factors reiterating the need for hyper parameter tuning for these models. Our LR model outperformed a study by Lai *et al.* (2019) which calculated an accuracy of only 0.76 within their model however, using different factors including age, sex, fasting blood glucose, BMI, high-density lipoprotein, triglycerides, blood pressure, and low-density lipoprotein.

Table 2: Performance Metrics for Each Model

Model	Accuracy	Precision	Recall	F1_Score	AUC
Random Forest	0.779	0.722	0.672	0.696	0.856
SVM	0.766	0.75	0.569	0.647	0.863
Logistic Regression	0.805	0.818	0.621	0.706	0.848

Table 2 presents the performance metrics for each model, highlighting their effectiveness in diabetes classification. The RF model achieved an accuracy of 0.779, indicating that it correctly classified approximately 77.9% of cases. It had a precision of 0.722, meaning that when it predicted diabetes, it was correct 72.2% of the time. Its recall of 0.672 suggests that it correctly identified 67.2% of all true positive diabetes cases. The F1 score, a balance between precision and recall, was 0.696, indicating a fairly strong trade-off between the two metrics. The AUC (0.856) shows that RF effectively differentiates between diabetic and non-diabetic cases.

The SVM model performed slightly worse in accuracy (0.766) compared to RF. While its precision of 0.75 was the highest among the three models, indicating it had the lowest false positive rate, its recall was lower at 0.569, meaning it missed a significant number of actual diabetes cases. This is reflected in its F1 score of 0.647, which is lower than both RF and Logistic Regression, emphasizing its weaker recall. However, its AUC of 0.863 was the highest, suggesting it had the best overall discriminatory ability.

The LR model had the highest accuracy (0.805) and the best precision (0.818), meaning it made the most confident correct predictions. Its recall of 0.621 was lower than RF but higher than SVM, resulting in an F1 score of 0.706, making it the most balanced model in terms of both false positives and false negatives. Its AUC of 0.848 suggests strong classification capability, though slightly lower than SVM.

These results indicate that RF provides the best balance between precision and recall, making it a robust choice for diabetes prediction. SVM excels in precision but has weaker recall, making it more prone to missing actual diabetes cases. Logistic Regression emerges as the most balanced model overall, with strong precision and competitive recall, making it a reliable choice when both false positives and false negatives are of concern.

4.5 Implications, Limitations and Future Work

This research has several important implications, particularly in improving the efficiency of diabetes screening and early intervention strategies. By identifying key predictive factors such as glucose levels, BMI, and age, machine learning models can help develop more targeted and cost-effective screening tools. These tools can enhance early detection, allowing healthcare providers to implement preventative measures before complications arise. The findings also

reinforce the importance of weight management and lifestyle interventions in reducing diabetes risk. By incorporating machine learning-based risk assessments into routine healthcare, clinicians can better prioritize high-risk individuals, leading to more effective resource allocation and improved patient outcomes.

Despite its contributions, this study has limitations that must be acknowledged. The dataset used is relatively small and limited to female participants, reducing the generalisability of the findings. A more diverse dataset that includes both genders and a wider range of ethnic and socioeconomic backgrounds would provide a more comprehensive understanding of diabetes risk factors. Additionally, the dataset does not include key lifestyle variables such as diet and physical activity, which play a significant role in diabetes development. The exclusion of these factors limits the model's ability to provide a holistic assessment of diabetes risk. Future studies should aim to incorporate a more extensive set of predictors to improve the accuracy and applicability of machine learning models in this domain.

Another limitation is the assumptions made by the models used in this study. Logistic regression assumes a linear relationship between predictors and the log-odds of diabetes, which may oversimplify complex interactions between variables. Support vector machines rely on kernel-based separability, which may not accurately capture the underlying structure of real-world health data. Random forests, while effective, can be prone to overfitting, particularly when working with smaller datasets. These methodological constraints highlight the need for further model refinement and evaluation. Additionally, this study did not employ hyperparameter tuning or cross-validation, which could have improved model performance and robustness. A more rigorous validation approach would provide stronger evidence for the generalisability of these findings.

Future research should focus on addressing these limitations. Expanding the dataset to include a broader and more diverse population is essential for improving the external validity of the models. Incorporating additional variables such as dietary habits, physical activity levels, stress, and genetic predisposition could significantly enhance predictive accuracy. Furthermore, the use of more advanced machine learning techniques, including gradient boosting methods like XGBoost and deep learning models such as artificial neural networks, could provide more powerful and flexible predictive capabilities.

Model validation and optimization should also be a priority in future studies. Implementing cross-validation techniques, such as k-fold cross-validation, would help ensure that model performance is not overly dependent on a single data split. Hyperparameter tuning using grid search or Bayesian optimization could refine model parameters to achieve better predictive accuracy. Additionally, the use of explainability techniques such as SHAP (Shapley Additive Explanations) could provide deeper insights into the relative importance of different predictors, improving the interpretability and transparency of the models.

Finally, future research should explore real-world applications of machine learning-based diabetes risk prediction. Developing practical tools such as mobile applications or AI-driven

clinical decision support systems could help integrate predictive models into routine healthcare practice. By leveraging electronic health records and wearable health data, machine learning models could provide dynamic and personalized risk assessments, further improving early detection and intervention strategies. Addressing these challenges will be critical in advancing the use of machine learning for diabetes prediction and ensuring its effectiveness in clinical and public health settings

5. Conclusion

This study utilized three machine learning models—LR, SVM, and RF—to explore the most significant predictors of diabetes and assess how accurately these models could classify diabetes outcomes. Consistent with existing literature, glucose levels and Body Mass Index (BMI) emerged as the most influential predictors across both the LR and RF models. In terms of predictive performance, the RF model outperformed the others, achieving the highest accuracy, followed by LR. These findings suggest that machine learning techniques, particularly ensemble methods like RF, hold significant potential in improving the prediction of diabetes.

However, there are limitations to this study that must be acknowledged. The dataset used was relatively small and specific to a population of women, which restricts the generalizability of the results. The lack of data diversity, particularly in terms of gender, age, and ethnicity, means that the models' performance in different populations remains unclear. Moreover, the limited sample size may have impacted the stability and robustness of the predictions, potentially affecting the reliability of the conclusions.

To build on these findings, future research should aim to expand the dataset to include a more diverse sample, encompassing a broader range of demographics. This would help determine how well the models perform across different groups and improve their generalizability. Additionally, incorporating feature interpretation methods, such as SHAP (SHapley Additive exPlanations), could provide greater insight into the decision-making process of the models, enhancing their transparency. SHAP would allow for a deeper understanding of the contribution of individual features, making it easier to interpret how certain variables, like glucose and BMI, influence diabetes predictions. Moreover, exploring other advanced techniques, such as deep learning or reinforcement learning, could lead to improvements in model performance, particularly when working with larger, more complex datasets. These advancements would be essential for optimizing machine learning models in healthcare settings, ensuring more accurate and reliable predictions.

References

Akula, R., Nguyen, N. and Garibay, I. (2019) 'Supervised machine learning based ensemble model for accurate prediction of type 2 diabetes'.

American Diabetes Association (2021) 'Improving care and promoting health in populations: Standards of medical care in diabetes', *Diabetes Care*, 44(Suppl 1), pp. S7–S14.

Da Silva, R., Martins, A. and Sousa, P. (2024) 'Shared risk factors across chronic diseases: Implications for prediction', *Chronic Disease Journal*, 48(3), pp. 120–135.

DeFronzo, F., R. (2005) 'Type 2 diabetes mellitus', *Nat Rev Dis Primers*, 1(15019).

Diaz-Santana, M.V., O'Brien, K.M., Park, Y.-M.M., Sandler, D.P. and Weinberg, C.R. (2022) 'Persistence of risk for type 2 diabetes after gestational diabetes mellitus', *Diabetes Care*, 45(4), pp. 864–870.

Dominguez, L.J. and Gonnelli, S. (2024) 'Calcium, vitamin d, and aging in humans', *Nutrients*, 16(23), p. 3974.

Flowers, K., Martinez, J. and Turner, E. (2024) 'Glucose control and its impact on reducing diabetes risk', *Diabetes Research and Clinical Practice*, 190.

Guh, D.P., Zhang, W., Bansback, N., Amarsi, Z., Birmingham, C.L. and Anis, A.H. (2009) 'The incidence of co-morbidities related to obesity and overweight: A systematic review and meta-analysis', *BMC Public Health*, 9(1), p. 88.

Hoppe, C., Muller, S. and Becker, J. (2024) 'Preventative healthcare strategies in diabetes management', *Public Health Review*, 46(1), pp. 34–50.

Lai, H., Huang, H., Keshavjee, K., Guergachi, A. and Gao, X. (2019) 'Predictive models for diabetes mellitus using machine learning techniques', *BMC Endocr. Disord.*, 19(1), p. 101.

Lu, X., Xie, Q., Pan, X., Zhang, R., Zhang, X., Peng, G., *et al.* (2024) 'Type 2 diabetes mellitus in adults: Pathogenesis, prevention and therapy', *Signal Transduct. Target. Ther.*, 9(1), p. 262.

Moghaddam, M.T., Jahani, Y., Arefzadeh, Z., Dehghan, A., Khaleghi, M., Sharafi, M., *et al.* (2024) 'Predicting diabetes in adults: Identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm', *Research Square* [Preprint].

Noroozi, F., Azizi, M. and Khani, Z. (2024) 'The role of high blood pressure in diabetes risk', *International Journal of Hypertension*, 14(2), pp. 88–99.

Rahman, M.H. (2024) 'Diabetes dataset'.

Torres-Torres, J., Monroy-Muñoz, I.E., Perez-Duran, J., Solis-Paredes, J.M., Camacho-Martinez, Z.A., Baca, D., *et al.* (2024) 'Cellular and molecular pathophysiology of gestational diabetes', *Int. J. Mol. Sci.*, 25(21).

Wang, L., Zou, J., Li, S., Tian, C., Ran, J., Yang, X., *et al.* (2024) 'Triglyceride glucose-body mass index as a mediator of hypertension risk in obstructive sleep apnoea syndrome: A mediation analysis study', *Sci. Rep.*, 14(1).

Yang, P., Liu, T. and Zhao, H. (2024) 'Diabetes and healthcare systems: Risk factors and early detection', *Health Policy and Management*, 32(5), pp. 120–138.