Which factors are the strongest predictors of diabetes, and how accurately can these predictors classify diabetes outcomes?

Word Count - 2282

720017170

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, random forests, and support vector machines (SVM). Each model will complete different objectives; logistic regression for interpreting predictor significance and calculating odds ratios, random forest 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.

Methods

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

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. Logistic regression was used to provide odds ratios that can quantify the relationship between predictors and diabetes risk. Random Forests 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 Logistic Regression to quantify the effect each variable has on the risk of diabetes. Within the Random Forest Model, feature importance scores were generated which allows ranking of the predictors based on their contribution to the model's accuracy.

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.

Results and Discussion

Table 1: Summary Statistics

Variable	N	Mean	Median	SD	Min	Max
Pregnancies	768	3.8	3	3.4	0	17
Glucose	768	122	117	30	44	199
BloodPressure	768	72	72	12	24	122
SkinThickness	768	29	29	8.8	7	99
Insulin	768	156	156	85	14	846
BMI	768	32	32	6.9	18	67
DiabetesPedigreeFunction	768	0.47	0.37	0.33	0.078	2.4
Age	768	33	29	12	21	81

Table 1 shows the summary statistics for the cleaned dataset, consisting of 392 individuals. Metrics including mean, median, standard deviation (SD), minimum, and maximum values will give an overview of the factors that contribute to diabetes. These metrics can show the factor's distribution and variability allowing understanding of the dataset.

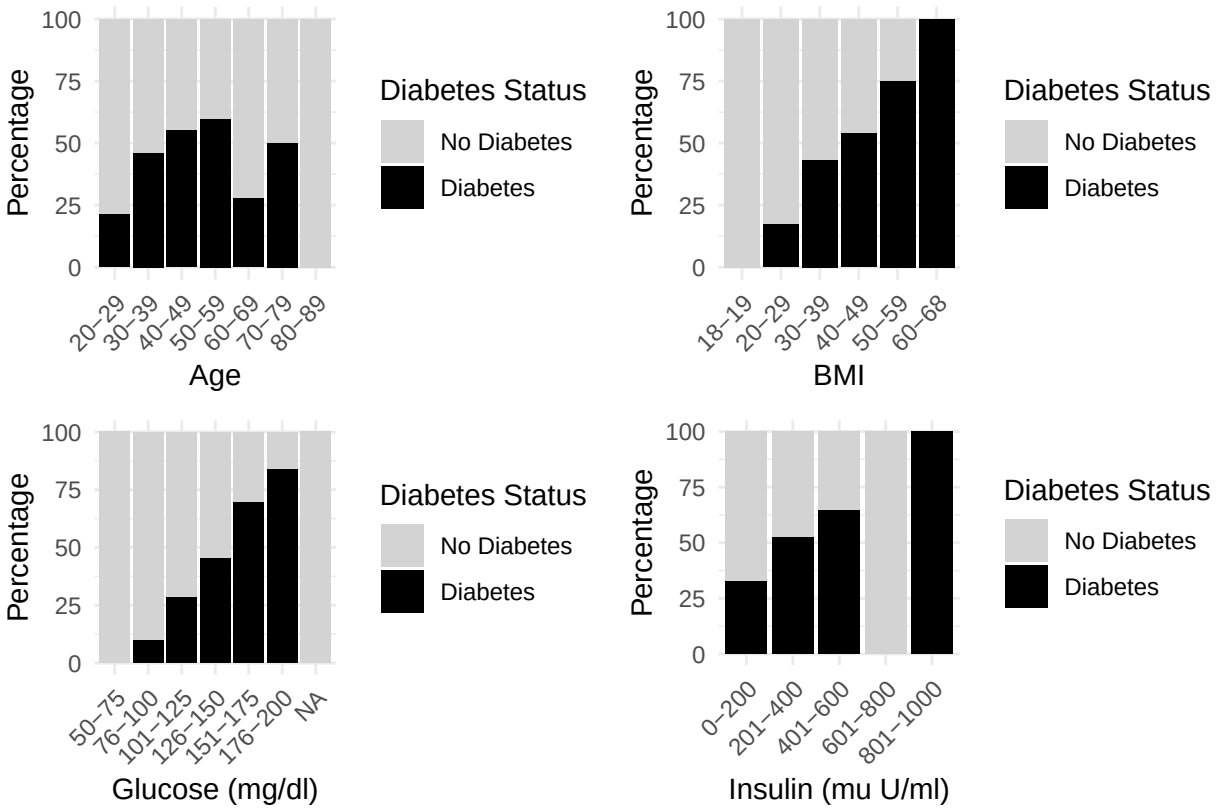


Figure 1: Distribution of Diabetes Status Across Key Factors

Figure 1 shows the percentage of diabetes cases across four key factors; age, BMI, glucose, and insulin. 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).

Random Forest Model

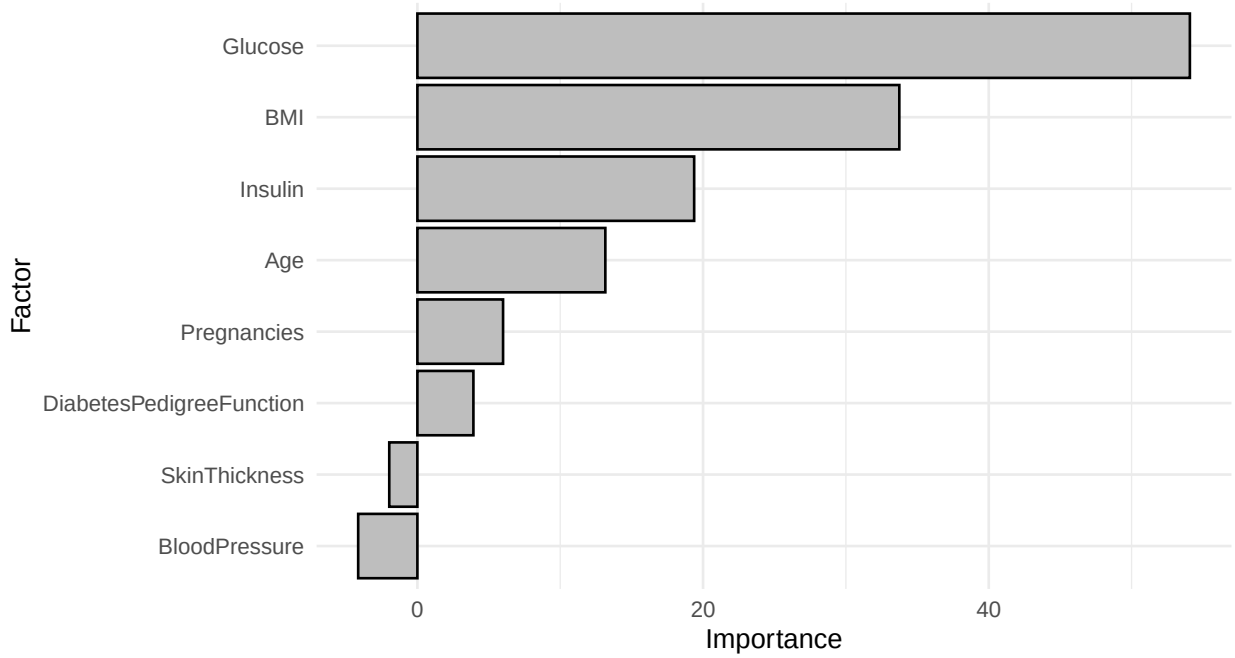
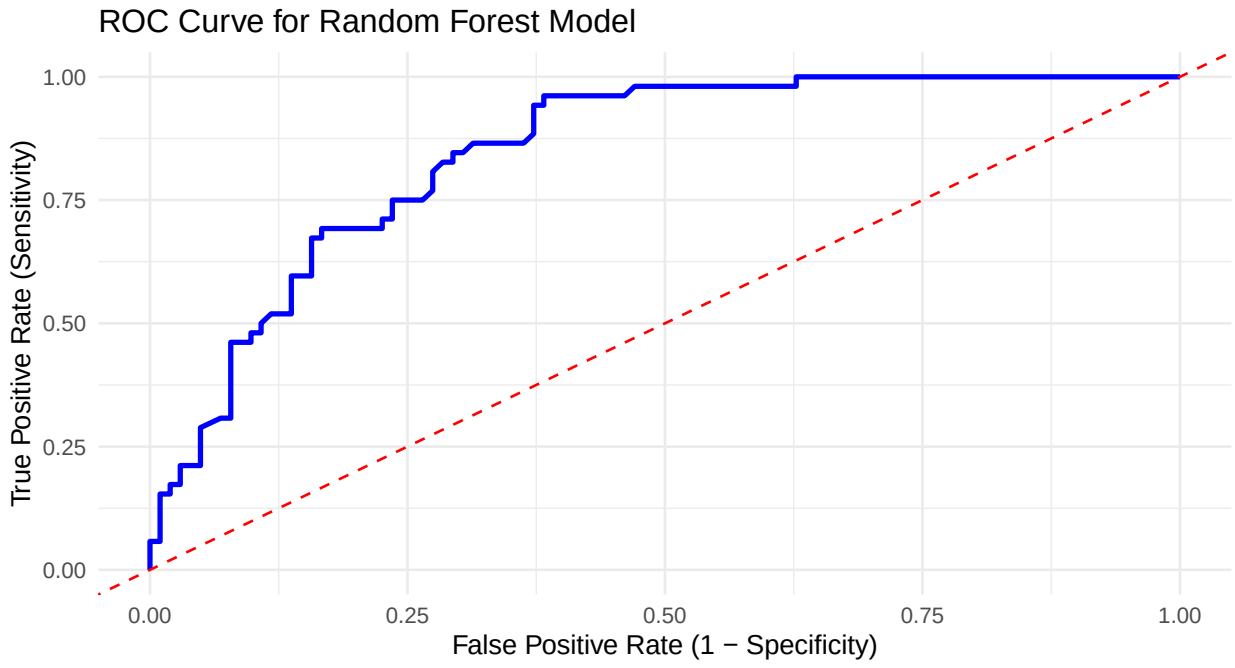


Figure 2: Importance of Features Based on Mean Decrease in Accuracy

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



AUC: 0.846

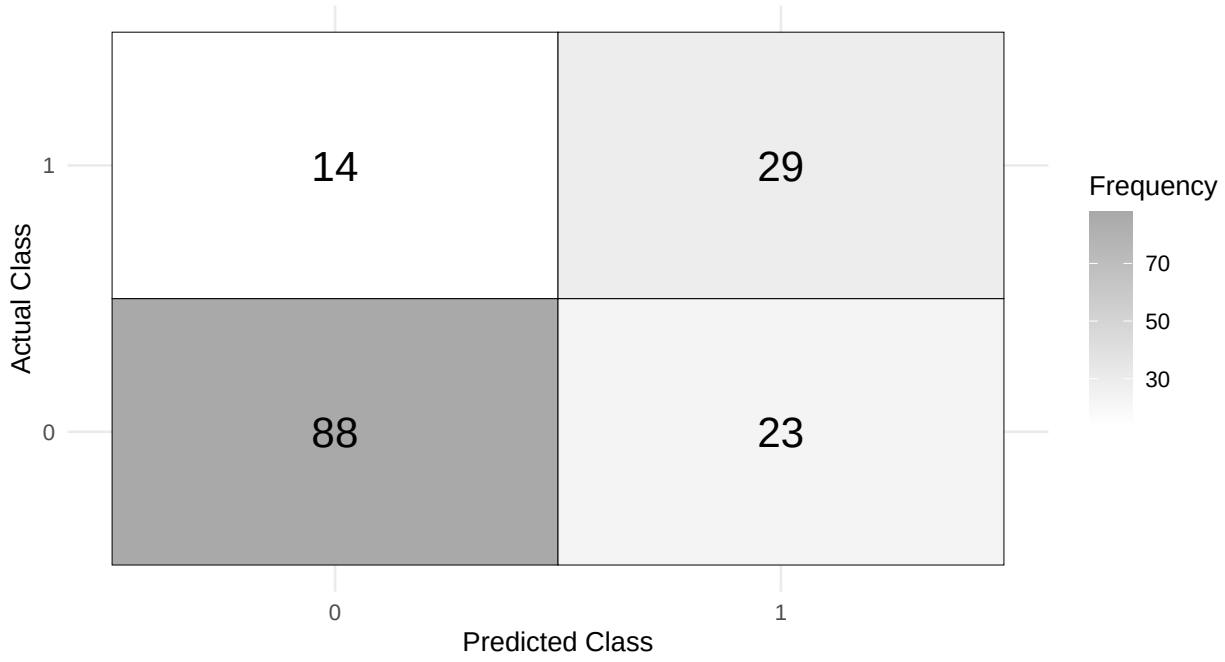


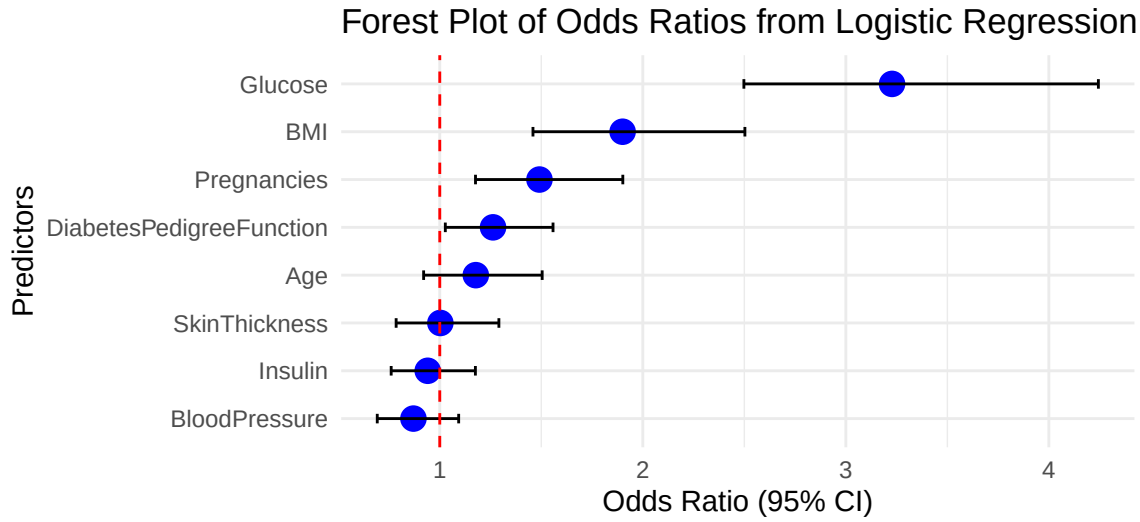
Figure: Confusion Matrix for the Random Forest Model

The random forest model trained in this study highlights glucose, age, and insulin as the most significant predictors of diabetes (Figure 2). These findings are consistent with existing literature, reinforcing the understanding of these variables as critical indicators in diabetes diagnosis.

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 random forest 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.

Logistic Regression Model



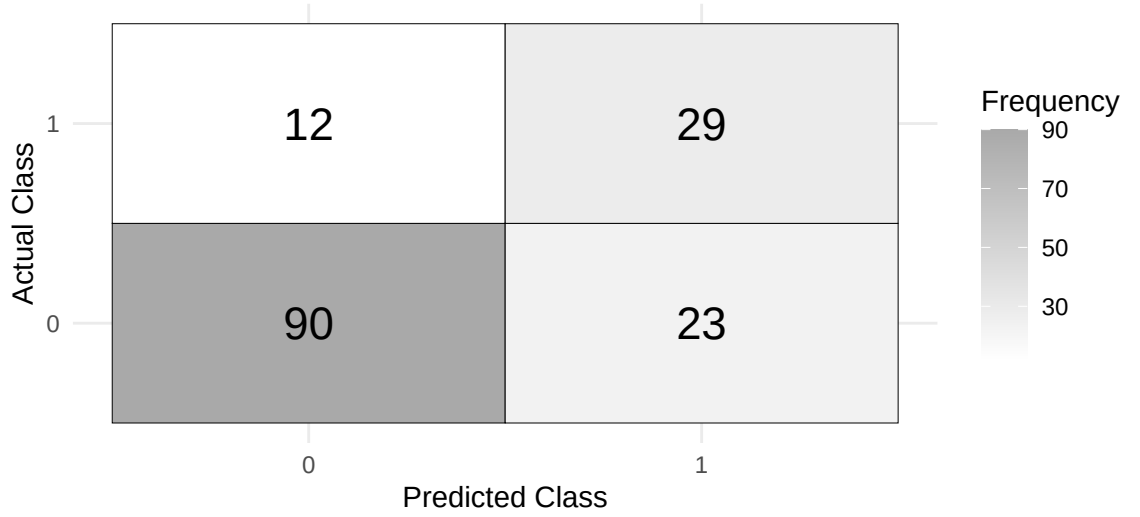
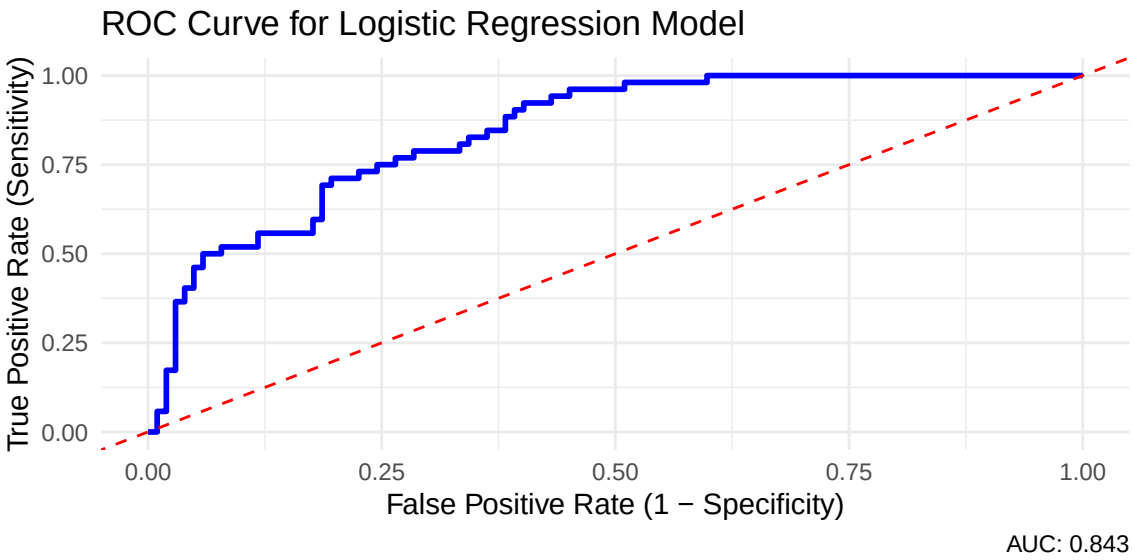


Figure: Confusion Matrix for the Logistic Regression Model



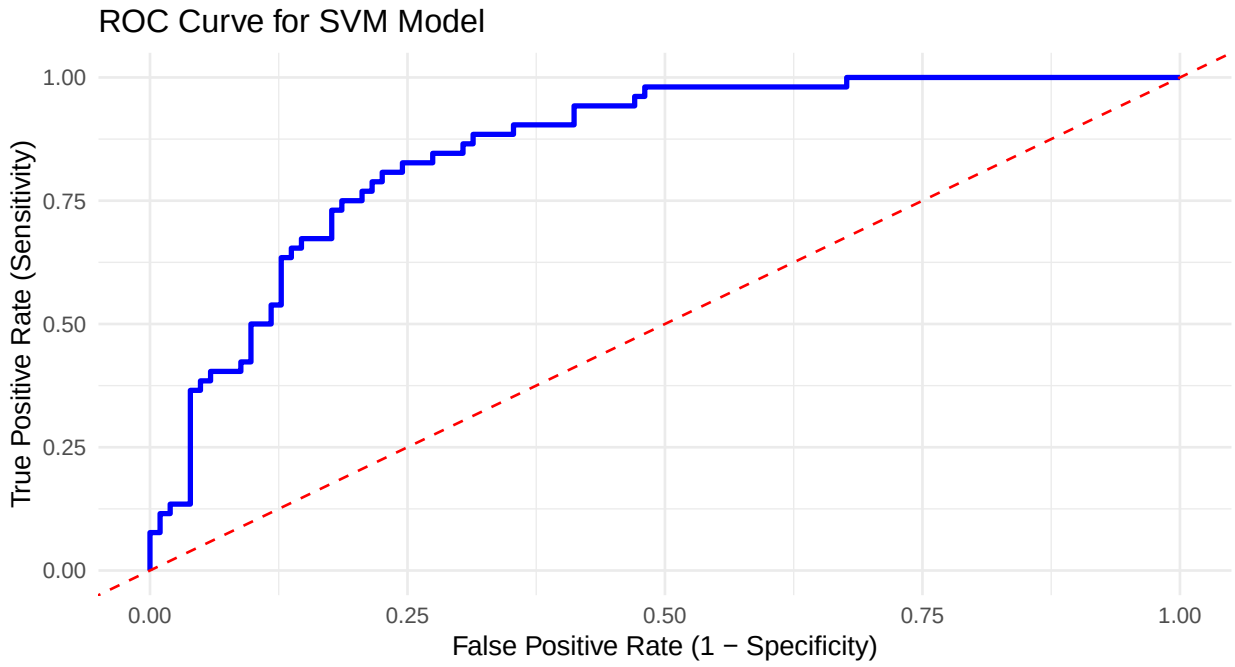
The logistic regression model shows a quantitative understanding of how the key predictors affect diabetes risk (Figure 3). Odds ratios were calculated allowing an easy interpretation of the relationships between the factors and diabetes risk. The odds ratio indicates the increase in the risk of diabetes for a one-unit increase in that variable.

The model was similar to the random forest model as glucose and BMI were also identified as the two most important predictors of diabetes. 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 over 300%. 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 56% 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 logistic regression 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).

Support Vector Machine Model



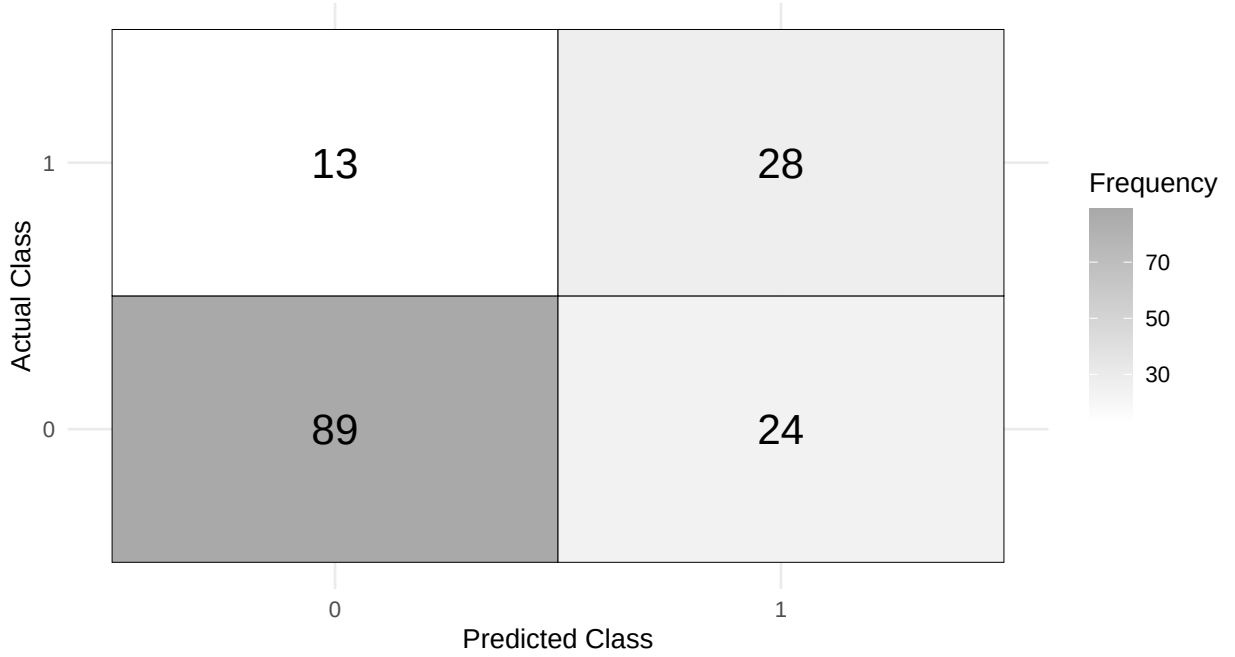


Figure 5: Confusion Matrix for the SVM model

The final model that was implemented was an SVM model which was used to evaluate the effectiveness of using the factors for predicting diabetes. This model showed good predictive ability for diabetes classification, the accuracy of this model was 0.76 indicating that the model accurately predicted 76% of the cases (Figure 4).

However, this model showed poor performance when predicting positive cases as the recall of the model was only 0.54 displaying the ability of the model to capture only 54% of the actual diabetic cases. The precision

of the model was shown to be 0.68 suggesting that the model was moderately effective at classifying diabetic cases from non-diabetic cases when predicting positive outcomes. The F1 score shows that the model lack the trade-off between precision and recall. These metrics show the need for hyperparameter tuning within more complicated models to refine their predictions as these low values may lead to false negatives which can lead to undue health risks and false positives which can put a strain on the healthcare systems.

From this model, a confusion matrix was developed (Figure 4). This figure showed that out of 41 actual diabetic cases within the test set, 28 were correctly identified as diabetic with only 13 being misclassified as non-diabetic. Within the 113 actual non-diabetic cases within the test set, only 24 were misclassified as diabetic and 89 were correctly classified. These results show that the model was very effective at classifying the diabetes outcome but with a slight bias to the majority case.

Model Evaluation

Figure 6: Accuracy for each model

Table 2: Performance Metrics for Each Model

Model	Accuracy	Precision	Recall	F1_Score	AUC
Random Forest	0.760	0.793	0.863	0.827	0.846
SVM	0.760	0.788	0.873	0.828	0.850
Logistic Regression	0.773	0.558	0.707	0.624	0.843

Figure 6 shows the accuracy of each model in classifying diabetes status. In this study, logistic regression shows the highest accuracy of 0.77 indicating that it correctly classified 77% of the cases. This model performed better than the other two models which both achieved an accuracy of 0.76, indicating the need for tuning and larger samples to improve the accuracy of these more complex models and logistic regression's simplicity could explain its increase in performance. 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 random forest model using similar factors reiterating the need for hyper parameter tuning for these models. Our logistic regression model was similar to 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.

Implications, Limitations and Future Work

This research holds many practical implications. Using the most important factors as the focus in screening tools could lead to diabetes prediction tools being more efficient and cost-effective, this can reduce medical complications and therefore strain on healthcare systems. Factors such as BMI and age show the possible use of exercise intervention schemes in reducing diabetes risk. These conclusions may be valuable as they can aid healthcare providers in implementing tailored prevention strategies using predictive models to identify high-risk individuals.

However, some limitations come with these findings. Due to the relatively small dataset, the conclusions lack some generalisability to a larger more diverse population. Furthermore, some factors such as diet and exercise which are seen to be large contributors to diabetes were excluded from this dataset. Also, the models used within the analysis have some shortcomings; logistic regression assumes a linear relationship between the factors and log-odds which could oversimplify the relationship. SVMs assume kernel-based separability which may not show the complex relationships in the data. Also, tuning methods e.g. grid search or cross-validation, were not implemented within the models which can reduce the accuracy of the models.

Future work should focus on rectifying the limitations of this current study. Any future research should include a larger more diverse dataset to ensure that the conclusions are applicable to the whole population not just females. Also, other factors such as diet and physical activity levels could've been used to show a more holistic view of the factors that affect diabetes risk. Furthermore, more advanced machine learning techniques could've been implemented such as ensemble methods, neural networks, or deep learning to enhance the accuracy of the conclusions. Finally, the models could've been implemented using cross-validation or grid search methods to increase the accuracy of the more complex models

Conclusion

This study used three machine learning models to determine the most important factors in the prediction of diabetes and how well these factors predict diabetes. Glucose and BMI were found to be the most influential predictors in both the logistic regression model and the random forest, this is supported by current literature focusing on diabetes. Logistic regression demonstrated the highest accuracy in classifying diabetes status, highlighting its effectiveness in capturing linear relationships within the data. These findings show the potential for machine learning to be an important component in diabetes prediction in the future.

However, this study was not without its own limitations. The relatively small dataset which only focused on women will restrict the ability for the conclusions to be generalized to a larger population, but the study demonstrated the potential for machine learning to be used in healthcare and future research should focus on broader datasets to make the models more accurate.

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.