STATS 4M03: Multivariate Analysis

Final Project

Diabetes Dataset

Submitted to

Dr. Eman M.S. Alamer

Department of Mathematics and Statistics

McMaster University

Hamiltion, Ontario, Canada L8S 4K1

November 22, 2024

Reported by

Xinyi Chen (400326045)

Tonu Xu(400370837)

Rayyan Kazim(Student ID)

Safi Khan (400402095)

First last (Student ID)

1 Introduction

1.1 Abstract

The study of diabetes is vital in understanding the progression of the disease and identifying key predictors. Throughout this paper, we perform data analysis on the diabetes dataset, primarily focusing on leveraging various statistical methods that we learned in STATS 4M03\6M03: Multivariate Analysis. By using a variety of methods, our goal is to predict the onset of diabetes from detailed medical diagnostic measurements based on several contributing health factors — with the aim of uncovering patterns and relationships between various clinical and lifestyle factors. Through this analysis, we hope to emphasize actionable insights for clinical decision-making and provide preventive strategies.

1.2 The Data

We will study the Diabetes Dataset from Rahman (2024), which contains 768 rows x 9 columns. Each column represents various health diagnostic metrics for predicting diabetes and each row corresponds to a unique patient record. Table 1 showcases a description for each of the columns in the dataset, and we will be using R as our main computing software (R Core Team, 2024).

Column	**Description Of Column**
Pregnancies	Integer: Number of times the patient has been pregnant.
Glucose	Integer: Plasma glucose concentration (mg/dL) after a 2-hour oral glucose tolerance test.
BloodPressure	Integer: Diastolic blood pressure (mm Hg).
SkinThickness	Integer: Triceps skinfold thickness (mm).
Insulin	Integer: 2-hour serum insulin (mu U/ml).
BMI	Float: Body mass index, defined as weight in kg/(height in m)$\hat{2}$.
DiabetesPedigreeFunction	Float: A score indicating genetic predisposition to diabetes based on family history.
Age	Integer: Age of the patient (in years).
Outcome	Binary: Target variable where 1 indicates diabetes, and 0 indicates no diabetes.

Table 1: Description of the Diabetes Dataset

1.2.1 Exploratory Data Analysis (EDA)

Figure 2 illustrates that there are a lot more individuals in the dataset who do not have diabetes. Figure 1 showcases that the ages listed in this dataset follow a right-skewed distribution, where majority of the individuals are aged 20-30.

Figure 3 illustrates the relatively high correlation between "SkinThickness," "BMI," and "Insulin." We also note that "Glucose" is reasonably correlated with "Insulin," "BMI," and "Age." Furthermore, "Age" is highly correlated with "Pregnancies."

The true label of this dataset is "Outcome," whereas the other variables are considered the predictors. Since our dataset is not that large, we determined that we should use all of the predictors as they all show reasonably high correlation with each other.

We cannot apply factor analysis on our dataset as it is not normally distributed. This can be confirmed by the Shapiro-Wilk test and normal QQ plot (R Core Team, 2024).

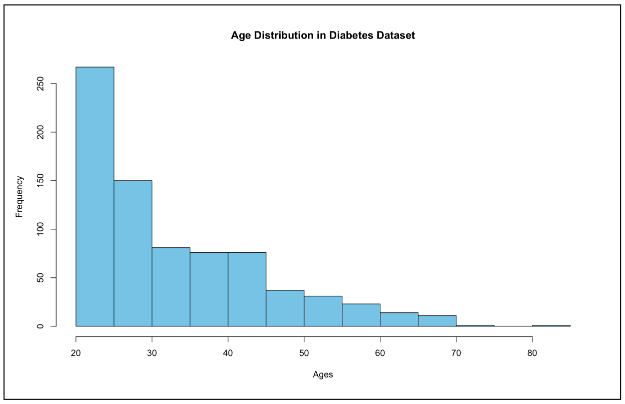

Figure 1: Age Distribution of the Dataset

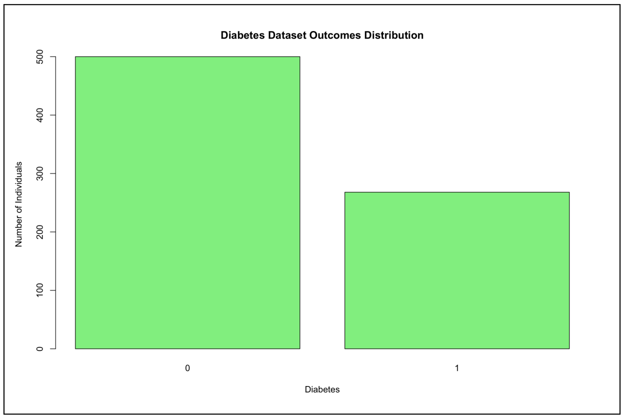

Figure 2: Distribution of Outcomes

1.2.2 Data Preparation

We prepared the data by scaling all the columns in the dataset except for column nine which is the response variable, "Outcome." We will split the data as follows: 75% for training, whereas the rest will be used for testing.

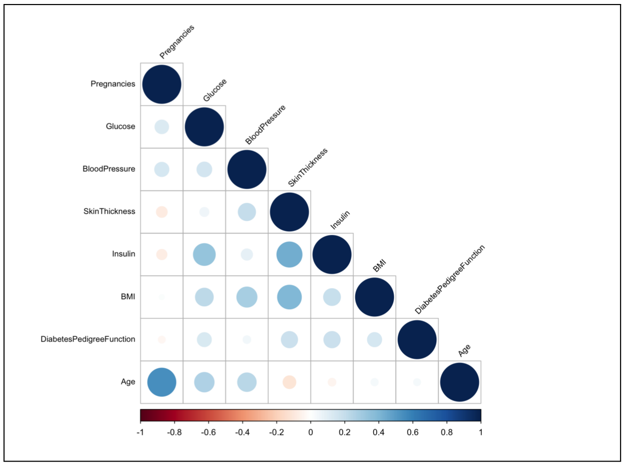

Figure 3: Correlation of Variables

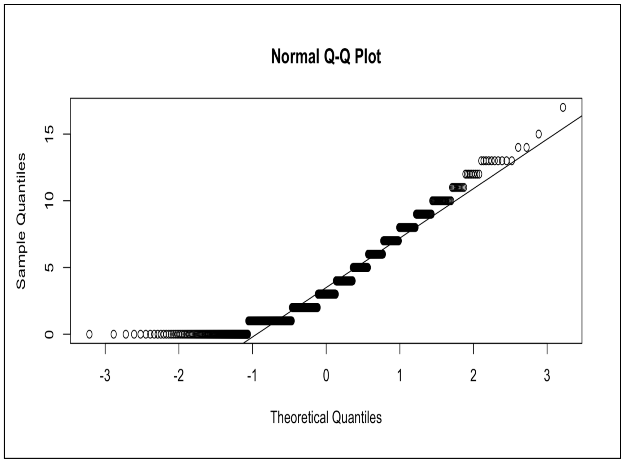

Figure 4: Normal QQ-Plot

2 Methodology

2.1 k-Nearest Neighbours

We executed the k-Nearest Neighbours algorithm, a non-parametric, supervised learning classifier that uses proximity to make classifications about the grouping of a dataset. (Peterson, 2009). Specifically, the algorithm takes an unlabelled observation and assigns it to the class that has the most labelled observations within its neighbourhood. (Alamer, 2024a). Additionally, we note that the optimal k-value will result in the best classification rate, based on the labelled points that are being treated as unlabelled by the algorithm (Alamer, 2024a).

2.2 Random Forest Classifiers

We also utilized the Random Forest Classifiers algorithm on our dataset (Zhou, 2012). Random Forest classifiers is a bootstrapping sampling method that combines the results of multiple decision trees to draw on a conclusion. The algorithm is an extension of the bagging algorithm which creates uncorrelated decision trees. For each tree, a random sample of $\mathcal{M}$ is taken at each decision tree split, M (Alamer, 2024b).

2.3 Binary Logistic Regression

We performed binary logistic regression on our dataset to explore the relationship and significance between our predictor variables and the binary response variable, “Outcome" (Faraway, 2016). We identify the most significant predictors using backward elimination, a step-wise technique employed in regression to minimize the risk of overfitting. This means that predictors whose p-value is greater than $\alpha = 0.05$ level of significance will be removed and we will run the algorithm again, repeating this process until we end up with only statistically significant predictors in our model.

2.4 Boosting

Boosting is one of the ensembles methods which combines multiple weak learners, typically decision trees, to create a strong predictive model (Chen, 2015; Friedman, 2001). For this analysis, we used XGBoost (Extreme Gradient Boosting), which is efficient and well optimized for large datasets, to predict the presence of diabetes based on our predictor variables. Missing or zero values were present in some predictors (e.g., insulin), which could influence the model’s performance.

Model Parameters; Learning rate (`eta`): 0.1, Maximum tree depth (`max_depth`): 6, Evaluation metric: Area Under the ROC Curve (AUC), and Number of boosting rounds (`nrounds`): 100

3 Discussion

3.1 Binary Logistic Regression

From (Table 2), we see that the coefficients for “BloodPressure,” “SkinThickness,” “Insulin,” and "Age" have high p-values, indicating that at $\alpha = 0.05$ level of significance, they are not significant in predicting the "Outcome". At this step, the full binary logistic regression model outputs a Misclassification Rate (MCR) of 0.3697917.

Table 3 showcases the output from the reduced Binary Logistic Regression Model. We observe that the p-values for all the predictors in the model are less than $\alpha = 0.05$, indicating that they’re statistically significant in predicting the response variable, "Outcome." Furthermore, the MCR has decreased to 0.2395833, i.e. $\sim 76\%$ of the the individuals predicted to have or not have diabetes

Coefficients	**Estimate**	**Std. Error**	**Z-Value**	**Pr(>|Z|)**
(Intercept)	-0.863576	0.111504	-7.745	9.57E-15
Pregnancies	0.411598	0.128257	3.209	1.33E-03
Glucose	1.003291	0.133428	7.519	5.51E-14
BloodPressure	-0.15522	0.122145	-1.271	0.2038
SkinThickness	-0.008376	0.123519	-0.068	0.94594
Insulin	-0.160385	0.120504	-1.331	1.83E-01
BMI	0.699525	0.135793	5.151	2.59E-07
DiabetesPedigreeFunction	0.295516	0.114215	2.587	0.00967
Age	0.212147	0.127142	1.669	0.0952

Table 2: Binary Logistic Regression Full Model Output

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	Odds Ratio
(Intercept)	-8.118918	0.730397	-11.116	<2e-16	—
Pregnancies	0.154186	0.032151	4.796	1.62E-06	1.167
Glucose	0.030679	0.003781	8.113	4.92e-16	1.031
BMI	0.080086	0.016048	4.990	6.03e-07	1.083
DiabetesPedigreeFunction	0.863621	0.336274	2.568	0.0102	2.372

Table 3: Binary Logistic Regression Reduced Model Output with Odds Ratio

were correctly classified. showcasing an improvement in the performance of the model.

Furthermore, Table 3 showcases the Odds Ratios (OR) for the predictors in the model. Note that each OR is greater than 1, suggesting increased likelihoods of the presence of diabetes with respect to the predictors. Specifically, the OR for ‘Pregnancies’ is 1.167, indicating that females with a history of pregnancies are 1.167 times more likely to have diabetes than those without. Most significantly, the OR for ‘DiabetesPedigreeFunction’ is 2.372, meaning that females who indicate a genetic predisposition to diabetes based on family history are 2.372 times more likely to develop diabetes than those without a genetic predisposition. The OR for the other predictors can be interpreted similarly.

3.2 Boosting

Boosting achieved an accuracy of 73.44%, with an AUC of 0.802. The AUC is the area under the ROC curve (Figure 5). The ROC Curve is a plot of the model’s Sensitivity (True Positive Rate) against the Specificity (True Negative Rate, showcasing how well the model distinguishes between classes across all possible thresholds (Fawcett, 2006). The AUC is the area under the ROC curve, summarizing the model's overall ability to discriminate between positive and negative classes. An AUC of 0.5 represents random guessing, while an AUC closer to 1.0 indicates excellent performance (Robin et al., 2011; Fawcett, 2006).

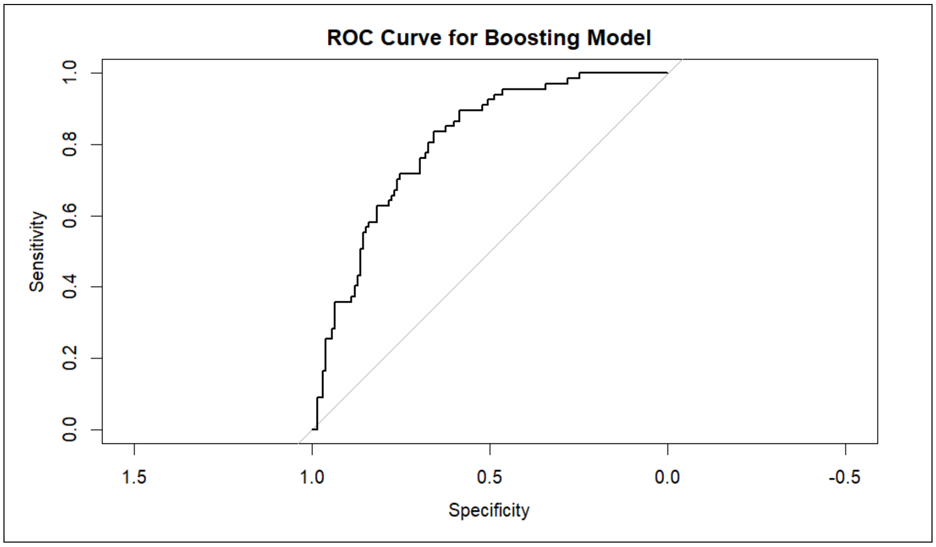

Figure 5: ROC Curve for Boosting Model

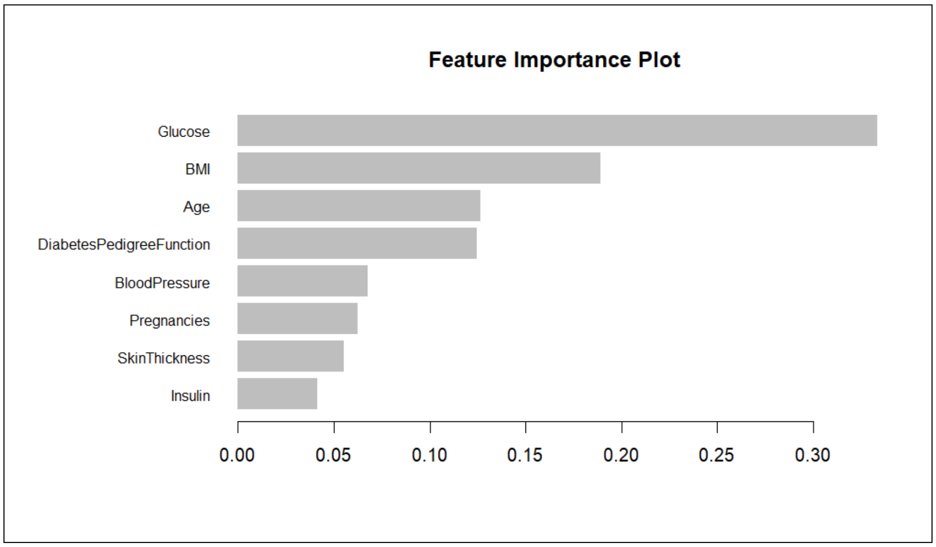

Figure 6: Feature Importance Plot

Boosting proved to be an effective method for diabetes classification through achieving an AUC of 0.802 and identifying Glucose, BMI, and Age as the most critical predictors in Figure 6. However, the moderate specificity given by the model (58.21%) suggests potential overemphasis on predicting positive cases, leading to false positives. Future work could improve model balance through hyperparameter tuning or additional data preprocessing techniques.

3.3 k-Nearest Neighbours

After running tune.knn() with 5-fold cross validation to get the parameters for the k-nearest neighbours algorithm (k-NN), we found that the best value for k is 3. which can be observed in Figure 7. After inputting $k = 3$ and running the k-NN algorithm, we found that the MCR given by k-NN classification is 0.2916667, i.e. 71% of the individuals with or without diabetes were correctly classified.

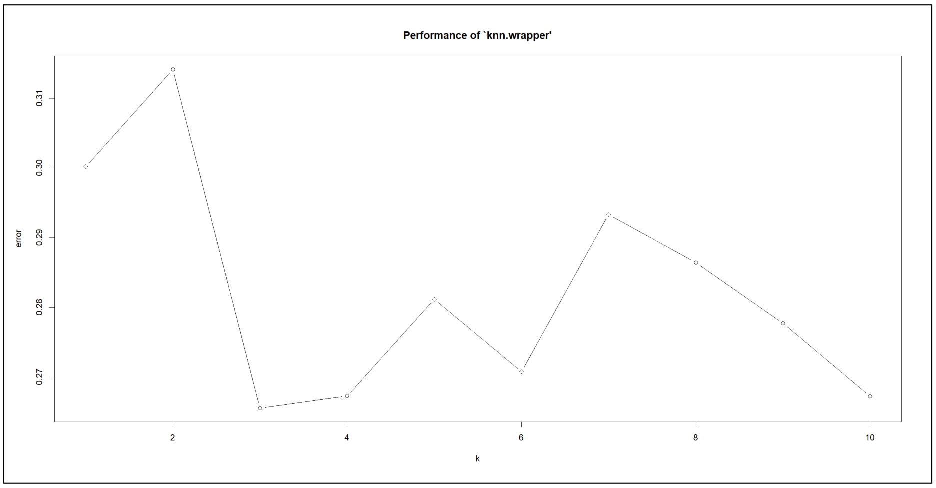

Figure 7: Output Plot From k-NN Classification

3.4 Random Forest Classifiers

After running tune.RandomForest() with 5-fold cross validation to get the best parameters for the random forest algorithm, we found that the best value for $\mathcal{M}$ is 4 and the best value of M is 200.

After tuning and inputting $\mathcal{M}$ = 4 and M = 200 into the random forest algorithm, we found that the MCR given by random forest classification is 0.28125, i.e. $\sim$ 72% of the the individuals predicted to have or not have diabetes were correctly classified. Finally, we see that Glucose and BMI are the most important variables, and excluding them from the model will lead to worse accuracy in predicting the Outcome. Conversely, SkinThickness and BloodPressure are the least important variables and excluding them from the model will not greatly affect the accuracy of the model.

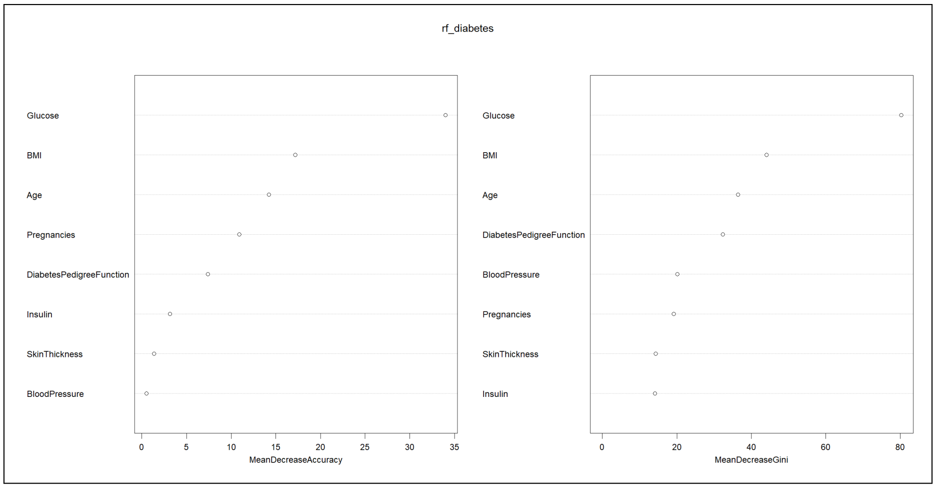

Figure 8: MeanDecreaseAccuracy Plot from Random Forest

4 Conclusion

Throughout this work, we presented various statistical Supervised Learning Analysis methods that could be used to study the Diabetes Dataset from Rahman (2024). We effectively showed the accuracy in performance of methods such as k-Nearest Neighbours classification, Binary Logistic Regression, and the ensembles methods: Random Forest and Boosting. Each of the models displayed an accuracy of $70.833\%, 76.042\%, 71.88\%$, and 73.44% respectively, indicating that Binary Logistic Regression was the best Supervised Learning Analysis method in predicting the onset of diabetes.

Furthermore, the conclusions determined by each of the various methods align with general scientific consensus on predictors for diabetes such as Glucose levels, BMI, Age, and whether an individuals family history has shown to have a predisposition to diabetes.

Ultimately, although Supervised Learning Analysis was effective in predicting the onset of diabetes based on several relevant predictors, we remain conscious of challenges such as the large sample size required for logistic regression to output stable results, choosing ‘k’ in k-NN, through computationally expensive techniques like cross validation, and the high memory usage and computational cost associated with random forest (Singh et al., 2016).

5 Bibliography

References

Alamer, E. (2024a). Lecture notes in multivariate analysis, lecture 14.

Alamer, E. (2024b). Lecture notes in multivariate analysis, lecture 16.

Chen, T. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4).

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2):1883.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahman, M. H. (2024). Kaggles predicting diabetes onset based on diagnostic measures.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77.

Singh, A., Thakur, N., and Sharma, A. (2016). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315.

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. CRC press.