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

In this paper, we will study the Diabetes Dataset. 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, with features capturing key medical attributes. Table 1 showcases each of the columns in the dataset and a description of each of the columns. We will be using R R Core Team (2024) as our main computing software

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 without diabetes than with 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. 75% of the data will be used for training, whereas the rest will be used for testing.

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. For dimensionality reduction, we would use principal component analysis. We will use 3 principal components.

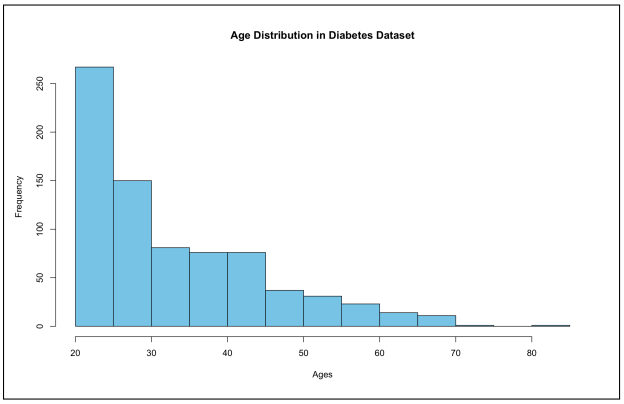

Figure 1: Age Distribution of the Dataset

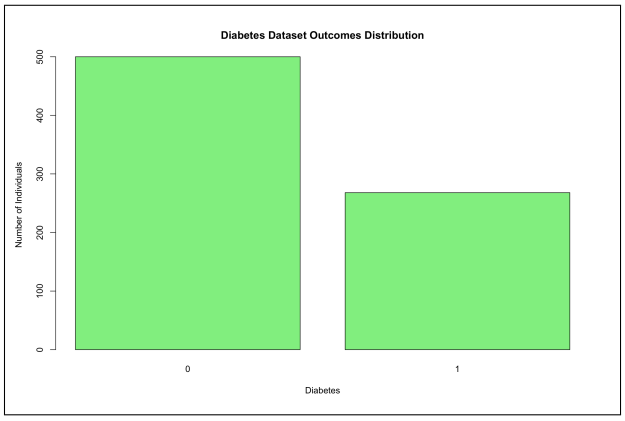

Figure 2: Distribution of Outcomes

1.2.2 Data Preparation

We write about the training and testing here, and how we used column 9 as a label. We also scaled the dataset.

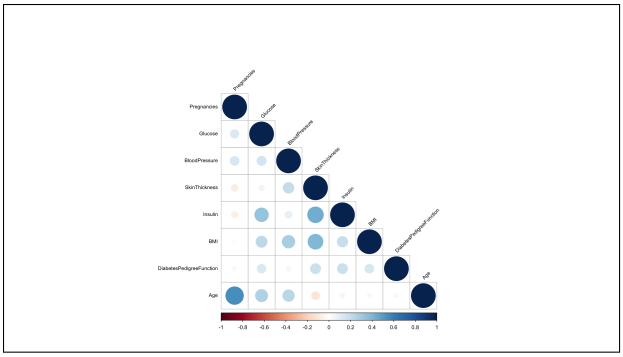

Figure 3: Correlation of Variables

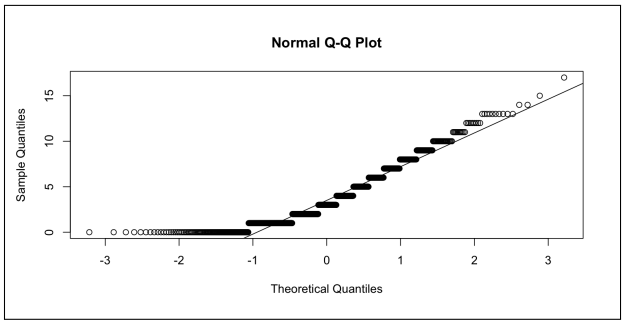

Figure 4: Normal QQ-Plot

2 Methodology

2.1 Supervised learning Analysis

In this section, we perform supervised learning analysis using Classification trees: k-Nearest Neighbours and the following ensembles method: Random Forests Classifiers.Zhou (2012)

2.1.1 k-Nearest Neighbours

Firstly, we executed the k-Nearest Neighbours algorithm Peterson (2009) on our dataset. The algorithm is a non-parametric, supervised learning classifier that uses proximity to make classifications about the grouping of a dataset.

2.1.2 Random Forest Classifiers

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

2.2 Binary Logistic Regression

We performed binary logistic regression Faraway (2016) on our dataset to explore the relationship and significance between our predictor variables and the binary response variable “Outcome." The initial model uses all eight predictor variables.The objective of using binary logistic regression was to identify the most significant predictors using backward elimination, a step-wise technique employed in regression. 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 — minimizing the risk of over-fitting.

2.3 Boosting

Boosting Chen (2015) Friedman (2001) is one of the ensembles methods which combines multiple weak learners, typically decision trees, to create a strong predictive model. For this analysis, we used XGBoost (Extreme Gradient Boosting), which is efficient and optimized for large datasets, to predict the presence of diabetes based on our predictor variables.

The model was trained on 75% of the data, leaving 25% for testing. The binary outcome variable (1 = diabetes, 0 = no diabetes) was predicted using features such as glucose levels, BMI, and age. 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 Curve (AUC)
- Number of boosting rounds (`nrounds`): 100

3 Discussion

3.1 Binary Logistic Regression

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

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.

Coefficients:	Estimate	Std. Error	Z-Value	Pr(>|Z|)
(Intercept)	-7.90218	0.71505	-11.051	<2e-16
Pregnancies	0.14734	0.03174	4.642	3.45E-06
Glucose	0.03135	0.00375	8.359	<2e-16
BMI	0.08415	0.01597	5.269	1.37E-07
DiabetesPedigreeFunction	0.863621	0.336274	2.568	0.0102
Age	0.029419	0.010701	2.749	0.00598

Table 3: Binary Logistic Regression Reduced Model Output

Table 3 contains the output from the Reduced Binary Logistic Regression Model. We observe that at $\alpha = 0.05$ level of significance, every coefficient in the model is statistically significant in predicting diabetes Outcome. This is further reinforced given that the reduced model outputs an MCR of 0.2395833, i.e. 76% of individuals who were predicting to have, or not develop diabetes

were correctly classified, indicating enhanced model performance. Consequently, we conclude that "Pregnancies," "Glucose," "BMI," and "DiabetesPedigreeFunction" all have a significant influence on the “Outcome”. This suggests a strong association between these factors and the risk of diabetes in females.

3.2 Boosting

The model achieved an accuracy of 80.5% on the test set, with an AUC of 0.87. The AUC, derived from the ROC curve (Figure 5), indicates that the model performs well in distinguishing between diabetic and non-diabetic individuals.

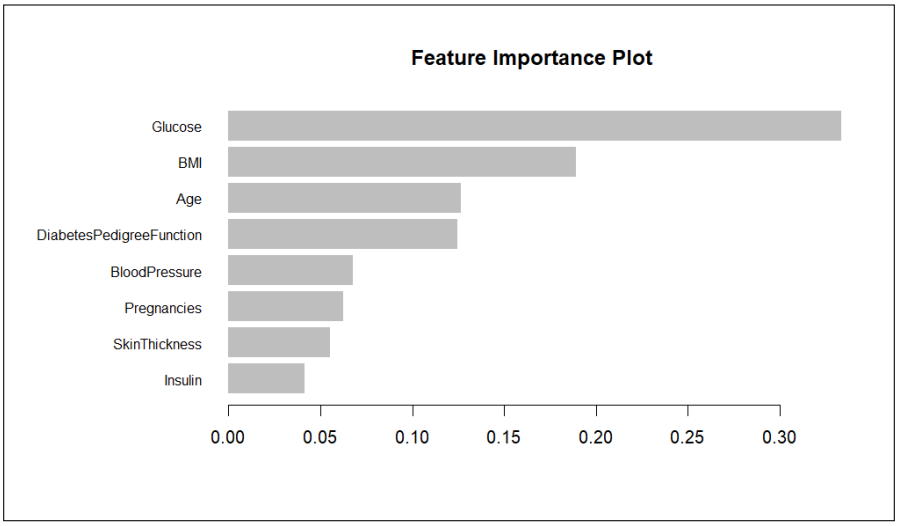

Figure 5: ROC Curve for Boosting Model

The feature importance plot (Figure 6) reveals that glucose is the most significant predictor, followed by age, BMI, and genetic predisposition (DiabetesPedigreeFunction). These results align with established medical understanding of diabetes risk factors.

Boosting effectively identified significant predictors of diabetes, particularly glucose and BMI. However, the model’s performance may be affected by missing data and the dataset’s limited size. Future work could address these limitations by imputing missing values and validating the model on larger datasets. Overall, XGBoost proved to be a robust method for this binary classification problem.

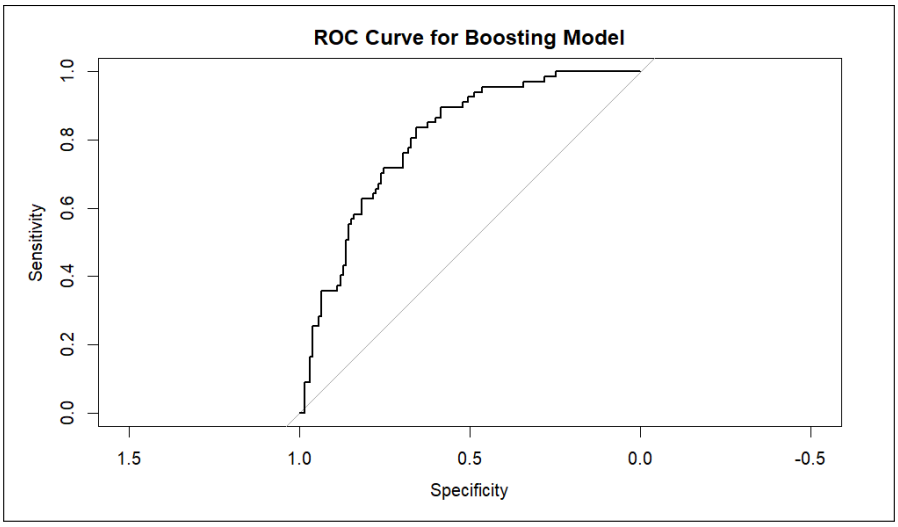

Figure 6: Feature Importance Plot

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. 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.

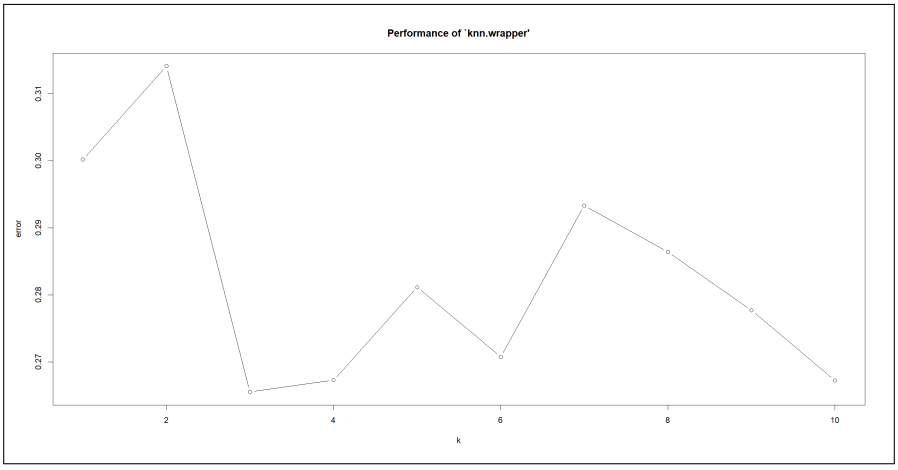

Figure 7: Output Plot From k-NN Classification

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. 72% of the the individuals with or without diabetes were correctly classified. Finally, we also observe that Glucose and BMI are the two most important variables according to the variable importance plot.

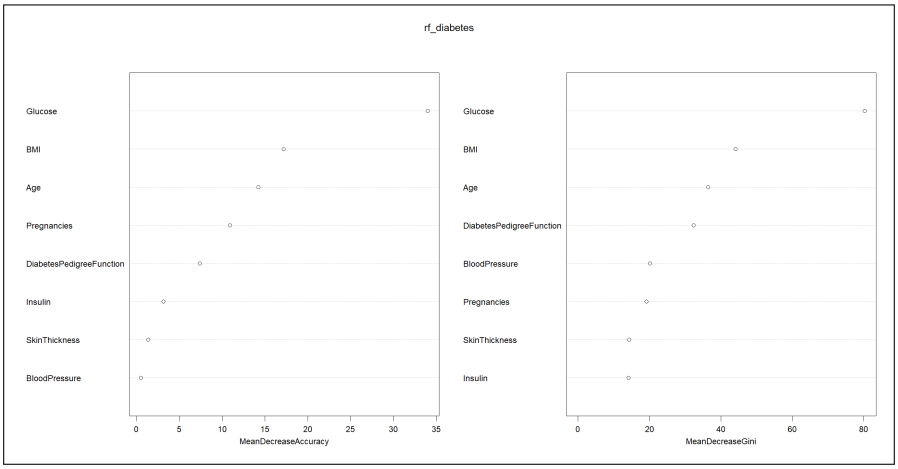

Figure 8: MeanDecreaseAccuracy Plot from Random Forest

4 Conclusion

TEMPORARY, WILL IMPROVE LATER Comparison between supervised and unsupervised learning analysis, which method performs better for this dataset, which version of machine learning analysis helps us draw better conclusions for our dataset etc.

5 Bibliography

References

Alamer, E. (2024). 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.

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.

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