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 can be found here: Diabetes Dataset. The diabetes dataset contains 768 rows x 9 columns, representing various health diagnostic metrics for predicting diabetes. 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

1.2.1 Exploratory Data Analysis (EDA)

1. The diabetes dataset consists of 768 observations and 9 variables. All variables are integers, except for "BMI" and "DiabetesPedigreeFunction", which are labelled as numeric. This dataset does not consist of any N/A values or duplicated rows.

2. A summary statistics table was used to show the mean, median, minimum and maximum values of each variable, as well as quartiles. There are a lot more observations without diabetes than with diabetes. The ages listed in this dataset follow a right-skewed distribution, where majority of the individuals are aged 20-30.

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

3. "SkinThickness" is well correlated with "BMI" and "Insulin". "Glucose" is reasonably correlated with "Insulin", "BMI" and also "Age". "Age" is well correlated with "Pregnancies".

4. The dataset will be split into 2, where 1 will be the response variable, "Outcome". The other will consist of all the other variables, which are considered the predictor variables. All predictors should be used as they all are numeric/integer variables who are well/reasonably correlated with each other. 75 percent of the data will be used for training, whereas the rest will be used for testing.

5. For dimensionality reduction, we cannot use factor analysis since the dataset is not normally distributed. This can be confirmed by the shapiro test and normal QQ plot. For dimensionality reduction, we would use principal component analysis. We will use 3 principal components.

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.

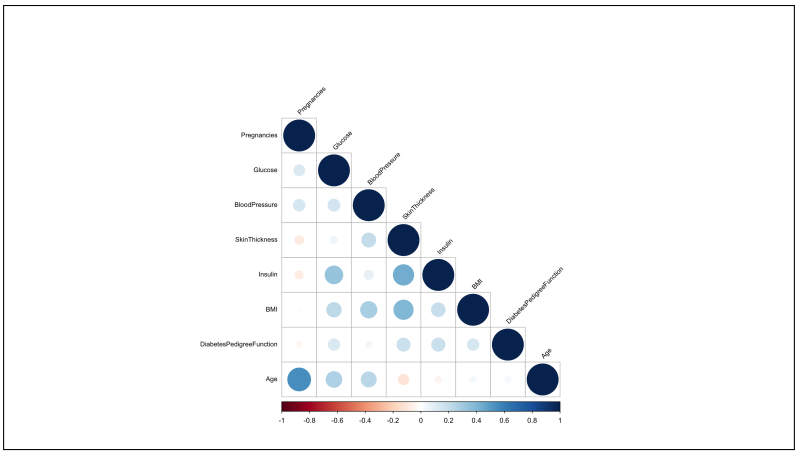

Figure 1: Correlation Plot

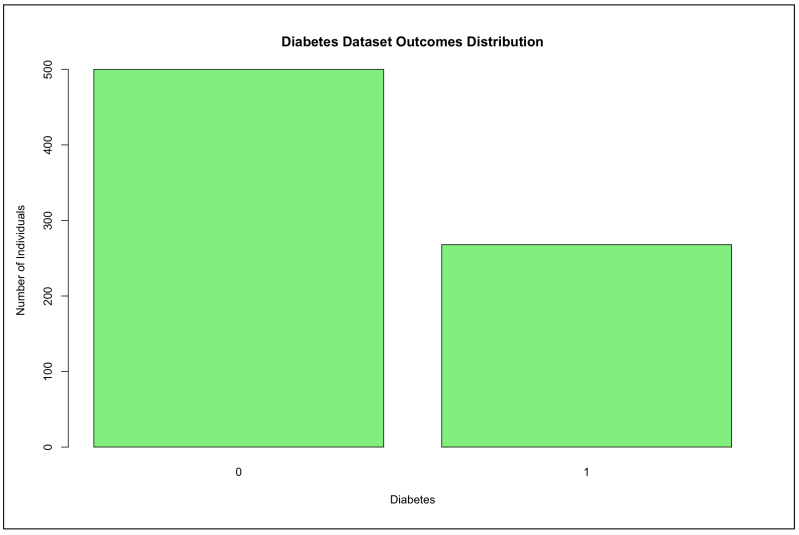

Figure 2: Outcomes Plot

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)

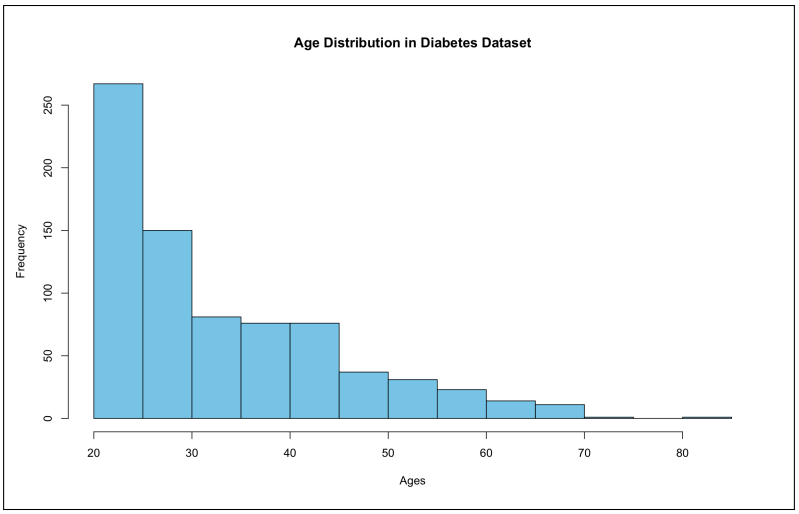

Figure 3: Ages Plot

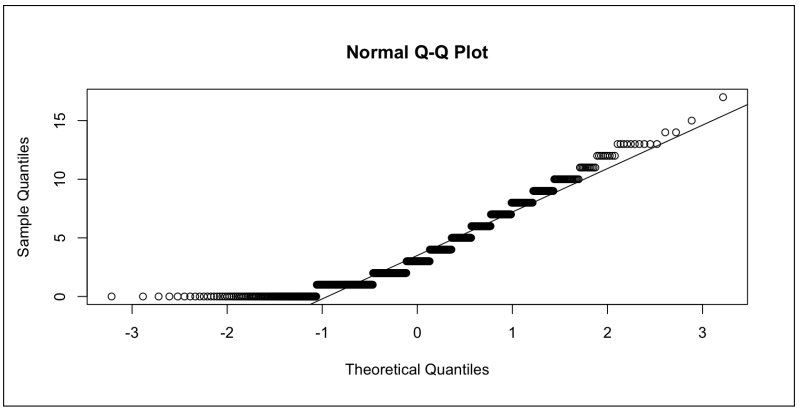

Figure 4: Normality Checking Plot

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 (2024a) is an extension of the bagging

algorithm Alamer (2024a) 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 a binary logistic regression analysis Faraway (2016), a supervised learning method, to explore the relationship between various predictor variables and the binary response variable "Outcome," indicating the presence or absence of diabetes. The initial model incorporated eight predictor variables: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. These variables were selected based on their potential relevance to diabetes risk, informed by existing domain knowledge. The objective of the analysis was to identify the most significant predictors using the backward elimination method, a stepwise regression technique. This approach began with a model including all predictors, subsequently removing variables in sequence based on the highest p-value exceeding 0.05. This systematic elimination of statistically insignificant predictors not only enhances the model's interpretability but also minimizes the risk of overfitting, ensuring that the final model retains only variables with significant contributions to predicting the outcome.

2.3 Boosting

Boosting Chen (2015) Friedman (2001) is an ensemble learning technique that 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 clinical measurements.

The model was trained on 70% of the data, leaving 30% 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)	-8.4398695	0.8176393	-10.322	< 2e-16
Pregnancies	0.1135092	0.0375672	3.021	0.00252
Glucose	0.0343876	0.0042392	8.112	4.99E-16
BloodPressure	-0.0134342	0.0059047	-2.275	0.0229
SkinThickness	0.0098866	0.0083722	1.181	0.23765
Insulin	-0.0015283	0.0009882	-1.546	0.122
BMI	0.0776662	0.0174657	4.447	8.72E-06
DiabetesPedigreeFunction	0.8088779	0.332009	2.436	0.01484
Age	0.0297298	0.0109345	2.719	0.00655

Table 2: Binary Logistic Regression Output

From (Table 2), the coefficient for "SkinThickness" had the highest p-value (0.23765 > 0.05), indicating it was not significantly associated with the outcome. At this stage, the misclassification rate was 0.22396. Consequently, "SkinThickness" was removed from the model. After removing "SkinThickness," the coefficient for "Insulin" had the highest p-value (0.24334 > 0.05), indicating it was also not significant. Then 'insulin' was excluded from the model.

(Table 3) demonstrates that all remaining variables exhibited p-values below 0.05, confirming their statistical significance. Additionally, the misclassification rate decrease to 0.21354,

Coefficients	Estimate	Std.Error	Z-Value	Pr(>|Z|)
(Intercept)	-8.304974	0.80035	-10.377	<2e-16
Pregnancies	0.115468	0.037168	3.107	0.00189
Glucose	0.03207	0.003896	8.232	<2e-16
BloodPressure	-0.012428	0.005724	-2.171	0.02992
BMI	0.082539	0.0163	5.064	4.11E-07
DiabetesPredigreeFunction	0.808528	0.329062	2.457	0.01401
Age	0.029419	0.010701	2.749	0.00598

Table 3: Binary Logistic Regression Output Post-Adjustment

signaling enhanced model performance. Consequently, the variables 'Pregnancies,' 'Glucose,' 'Blood Pressure,' 'BMI,' 'Diabetes Pedigree Function,' and 'Age' were identified as having 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 78.3% on the test set, with an AUC of 0.85. The AUC, derived from the ROC curve (Figure 5), demonstrates the model's strong ability to distinguish between diabetic and non-diabetic individuals.

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 5: ROC Curve for Boosting Model

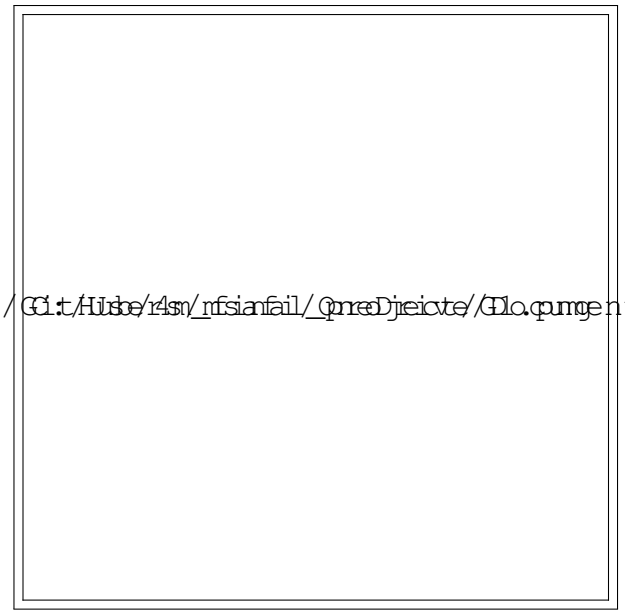

Figure 6: Feature Importance Plot

3.3 k-Nearest Neighbours

After executing the tune.knn() function in 5-fold cross validation. We found that the best value for k is 5. Then, after inputting $k = 3$ into the knn() function, we found that the MCR of the k-nearest neighbours is 0.2916667.

3.4 Random Forest Classifiers

After executing the tune.RandomForest() function in 5-fold cross validation. We found that the best value for mtry is 4 and the best value of ntree is 200. Then, after inputting mtry $= 4$ and ntree $= 200$ into the RandomForest() function, we found that the MCR of the random forests is 0.28125.

Finally, we also observe that Glucose and BMI are the two most important variables according to the variable importance plot.

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. (2024a). Lecture notes in multivariate analysis, lecture 16.

Alamer, E. (2024b). Lecture notes in multivariate analysis, lecture 2.

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.

Kurita, T. (2019). Principal component analysis (pca). *Computer vision: a reference guide*, pages 1–4.

Marden, J. I. (2004). Positions and qq plots. *Statistical Science*, pages 606–614.

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.