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 (Student ID)

First last (Student ID)

1 Introduction

1.1 Abstract

In this paper, we will perform a data analysis on the dataset: diabetes. We will use a number of methods that was learned in the course: STATS 4M03, Multivariate analysis. The purpose of this paper is to predict diabetes onset from detailed medical diagnostic measurements based on several health factors. In our paper, we found

.....

1.2 The Data

In this paper, we will study the dataset: diabetes. [1]

.....

1.2.1 Exploratory Data Analysis (EDA)

In this paper, we used several EDA methods to analyze our dataset. These methods include: PCA [2], and a normal QQ-plot [3]. some text [4]

1.2.1.1 PCA

From figure 1, we observe that we should take 3 principle components.

1.2.1.2 QQ-plot

From figure 2, we observe that the dataset is not normal

.....

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.

.....

2 Methodology

2.1 Supervised learning

In this paper, we performed supervised learning methods several supervised learning methods, these methods include: Classification trees, k-nearest neighbour bagging and random forest classfiers.[5]

2.1.1 Classification tree

2.1.2 k-nearest neighbours

2.1.3 bagging

2.1.4 random forest classfiers

2.2 Binary Logistic Regression

We performed a binary logistic regression analysis, 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.

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 1: Binary Logistic Regression

From Table 1, 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.

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
DiabetesPedigreeFunction	0.808528	0.329062	2.457	0.01401
Age	0.029419	0.010701	2.749	0.00598

Table 2: Binary Logistic Regression

Table 2 demonstrates that all remaining variables exhibited p-values below 0.05, confirming their statistical significance. Additionally, the misclassification rate decrease to 0.21354, 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.

4 Conclusion

.....

5 Bibliography

References

[1] Md. Hasibur Rahman. Kaggles predicting diabetes onset based on diagnostic measures, 2024.

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

[3] John I Marden. Positions and qq plots. *Statistical Science*, pages 606–614, 2004.

[4] Eman Alamer. Lecture notes in multivariate analysis, lecture 2, September 2024.

[5] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.