Multivariate Analysis Of The Diabetes Dataset

Rayyan Kazim, Safi Khan, Xinyi Chen, Tony Xu, Zesen Chen

McMaster University

12/2/2024

Dataset, Exploratory Data Analysis (EDA), Data Preparation

- Using the ‘R‘ programming language, we study the diabetes dataset obtained from Kaggle which has 768 rows and 9 columns. Each row corresponds to an unique patient record, and each of the columns represent the unique variables.

- There is one binary response variable that denotes presence of diabetes, the others are predictor variables. Each of the predictors are integers except for "BMI" and "DiabetesPedigreeFunction" which are categorized as float variables.

- The full dataset will be utilized and scaled. 75% of the data will be used for training, and the rest will be used for testing.

EDA Continued

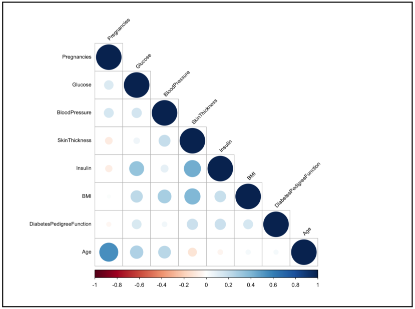

Figure 1: Correlation of Variables

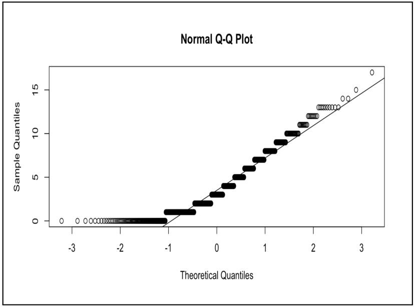

Figure 2: Normal QQ-Plot

- Figure 1 showcases that "SkinThickness" is well correlated with "BMI" and "Insulin," while "Age" is well correlated with "Pregnancies." Additionally, "Glucose" is reasonably correlated with "insulin," "BMI," and "Age."
- Figure 2 indicates that our dataset is not normally distributed.

Methodologies - Supervised Learning Analysis

- Used the following Supervised Learning Analysis methods: k-Nearest Neighbours (k-NN), Random Forest, and Boosting to study our dataset.

- k-NN is a non-parametric, supervised learning classifier that uses proximity to make classifications about the grouping of a dataset.

- Random Forest is a bootstrapping sampling method that combines the results of multiple decision trees to draw on a conclusion.

- Boosting is similar to Random Forest, however it is not a bootstrapping sampling method. Boosting also uses the entire dataset, or some subsample thereof, to generate the ensemble.

Methodologies - Binary Logistic Regression

- We performed Binary Logistic Regression because the response variable of our dataset, "Outcome," is binary.
- The initial model incorporates all eight predictor variables.
- The objective of using this technique is to identify the most significant predictors by using backwards elimination.
- This process ensures the final model will have only the most significant variables.

Discussion - Binary Logistic Regression

- Table 1 illustrates that "BloodPressure," "SkinThickness," "Insulin," and "Age," are statistically insignificant because their $p-$values are greater than $\alpha = 0.05$ level of significance, indicating that they should be removed in the final model.

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 1: Binary Logistic Regression Full Model Output

Discussion - Binary Logistic Regression Continued

- The $p-$values from Table 2 imply we have sufficient statistical evidence to conclude the variables present in the model have a significant influence in predicting "Outcome."
- Table 2 highlights Odds Ratio values of each predictor, showcasing the highers odds each variable has on predicting a positive or non-positive outcome.

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 2: Binary Logistic Regression Reduced Model Output with Odds Ratio

- We obtain that the MCR for binary logistic regression is 0.23958.

Discussion - k-Nearest Neighbours

- Figure 3 highlights that tuning with 5-fold cross validation suggests that the best value for k is 3.

- Running the k-NN algorithm with $k = 3$, we conclude that the MCR $= 0.2916667$.

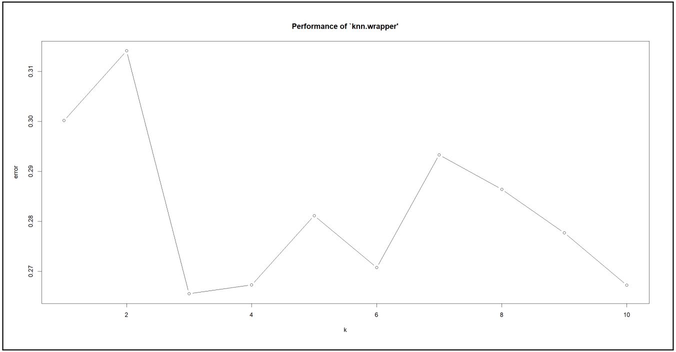

Figure 3: Output Plot From k-NN Classification

Discussion - Random Forest

- Tuning with 5-fold cross validation gives the best value for $\mathcal{M}$ and M as 4 and 200 respectively.
- Running Random Forest with $\mathcal{M} = 4$ and $M = 200$, we conclude that the MCR $= 0.28125$.
- Figure 4 showcases that "Glucose" and "BMI" are the two most important variables in the model.

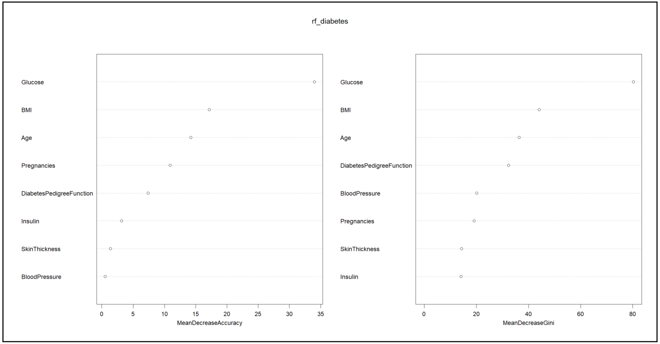

Figure 4: Random Forest Variable Importance

Discussion - Boosting

- Boosting achieved an accuracy of 73.44% (an MCR of 0.2656), and an AUC of 0.802.
- Figure 6 identifies "Glucose," "BMI," and "Age" as the most most predictors for "Outcome" in our Boosting model.

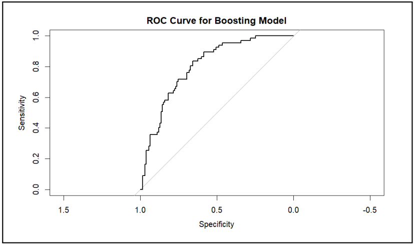

Figure 5: ROC Curve for Boosting Model

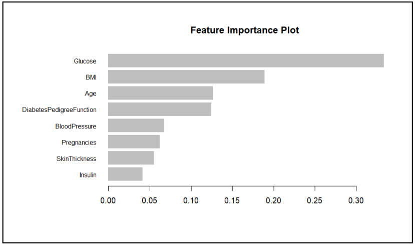

Figure 6: Age Feature importance plot

Conclusion

- The MCR values of each method indicate that Binary Logistic Regression was the most accurate out of all methods used.
- We effectively showcased the performance accuracy of the k-Nearest Neighbours algorithm, Binary Logistic Regression algorithm, and the ensemble methods: Random Forest and Boosting.
- The conclusions determined by each of the various methods align with the general scientific consensus on predictors for diabetes.