

# Multivariate analysis of the diabetes dataset

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

McMaster University

12/2/2024

# Dataset, EDA and data preparation

- ▶ Using the 'R' programming language, we will study the diabetes dataset obtained from Kaggle which has 768 rows and 9 columns where each row corresponds to an unique patient record.
- ▶ All variables are integers except for "BMI" and "DiabetesPedigreeFunction"
- ▶ The dataset will be split into 2, with the response variable being "outcome" and all other variables being predictor variables.
- ▶ 75 percent of the data will be used for training, and the rest will be used for testing.

# Exploratory Data Analysis (EDA) and data preparation

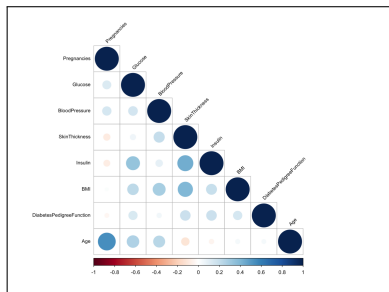


Figure: Correlation of Variables

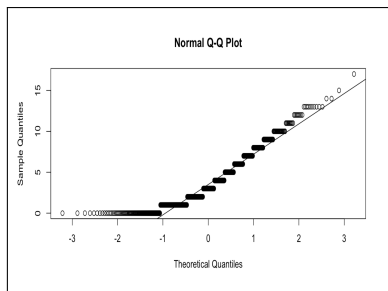


Figure: Normal QQ-Plot

- ▶ The normal QQ-plot suggests that our dataset is not normal
- ▶ "SkinThickness" is well correlated with "BMI" and "Insulin", "Age" is well correlated with "Pregnancies".
- ▶ "Glucose" is reasonably correlated with "insulin", "BMI" and "Age".

# Methodologies - Supervised learning analysis

- ▶ We used the methods: k-nearest neighbours ,random forest classifiers and boosting for our supervised learning analysis.
- ▶ K-nearest neighbours is a non-parametric, supervised learning classifiers that uses proximity to make classifications about the grouping of a dataset.
- ▶ RandomForest classifiers 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 - Logistic regression

- ▶ We will perform a binary logistic regression since our response variable "outcome" is binary.
- ▶ The initial model incorporates all eight predictor variables.
- ▶ The objective of this technique is to identify the most significant predictor using backwards elimination.
- ▶ We want the final model to have only the most significant variables. Removed predictors with p-values  $> 0.05$ .

# Discussions - Logistic regression

- Table 1 suggests that we should remove variables "BloodPressure", "SkinThickness", "insulin", and "Age" from the model, since their p-values were greater than 0.05.

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:** Binary Logistic Regression Full Model Output

## Discussions - Logistic regression

- ▶ Table 2 implies we have sufficient evidence to suggest those four variables have a significant influence over "outcome".
- ▶ Odds Ratio show the influence of various factors on the likelihood of developing diabetes

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:** Binary Logistic Regression Reduced Model Output with Odds Ratio

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

# Discussions - k-nearest neighbours

- ▶ 5-fold cross validation suggests that the best value for  $k$  is 3.
- ▶ Executing the `knn()` function with  $k = 3$ , we obtain that the MCR of the k-nearest neighbours is 0.2916667

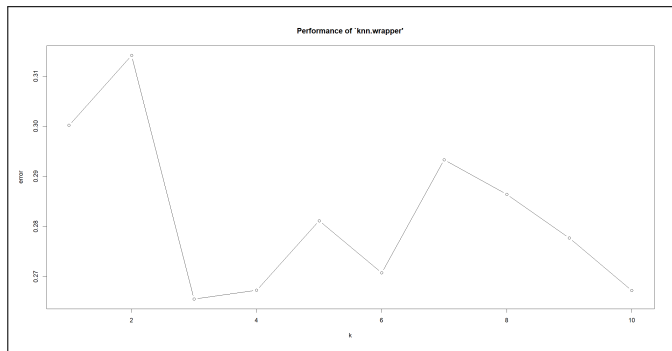
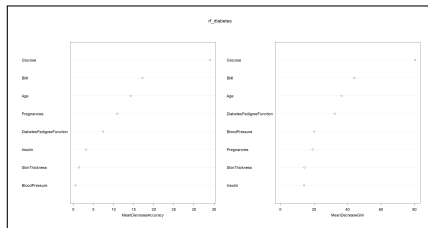


Figure: Output Plot From k-NN Classification



## Discussions - Random forests

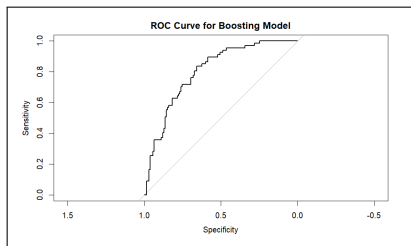
- ▶ 5-fold cross validation suggests that the best value for mtry is 4 and the best value for ntree is 200.
- ▶ Executing the RandomForest() function with mtry = 4 and ntree = 200, we obtain that the MCR of the RandomForest is 0.28125.
- ▶ We observe that Glucose and BMI are the two most important variables.



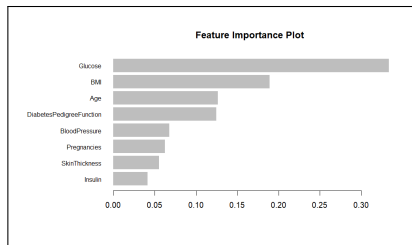
**Figure:** MeanDecreaseAccuracy & MeanDecreaseGini Plot from Random Forest

# Discussions - Boosting

- ▶ Boosting achieved an accuracy of 73.44 percent (a MCR of 0.2656), with an AUC of 0.802.
- ▶ 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.



**Figure:** ROC Curve for Boosting Model



**Figure:** Age Feature importance plot

# Conclusion

- ▶ The MCRs indicates that Binary logistic regression was the best method 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.