Multivariate analysis of the diabetes dataset

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

McMaster University

15/11/2024

Dataset

- We will study the dataset diabetes obtained from kaggles Rahman (2024)
- The dataset contains 768 rows and 9 columns.
- Each row corresponds to an unique patient record.
- We will be using the R programming language.

Exploratory Data Analysis (EDA) and data preparation

- All variables are integers except for "BMI" and "DiabetesPedigreeFunction"
- "SkinThickness" is well correlated with "BMI" and "Insulin", "Age" is well correlated with "Pregnancies".
- "Glucose" is reasonably correlated with "insulin", "BMI" and "Age".
- 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.

Methodologies - Supervised learning analysis

- We used the methods: k-nearest neighbours Peterson (2009),random forest classfiers Zhou (2012) and boosting Chen (2015) for our supervised learning analysis.
- K-nearest neighbours Peterson (2009) is a non-parametric, supervised learning classifiers that uses proximity to make classifications about the grouping of a dataset.
- RandomForest classifiers Zhou (2012) is a bootstrapping sampling method that combines the results of multiple decision trees to draw on a conclusion.
- Boosting Chen (2015) 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 Faraway (2016) 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.

Discussions - Logistic regression

- Removed variables "SkinThickness" and "insulin" from the model, since their p-values were greater than 0.05.
- We have evidence to suggest that "Pregnancies", "Glucose", "Blood Pressure", "BMI", "DiabetesPedigreeFunction" and "Age" have a significant influence over "outcome".

Discussions - k-nearest neighbours

- 5-fold cross validation Alamer (2024) suggests that the best value for k is $k = 3$
- Executing the knn() Alamer (2024) function with $k = 3$, we obtain that the MCR of the k-nearest neighbours is 0.2916667

Discussions - Random forests

- 5-fold cross validation Alamer (2024) suggests that the best value for mtry is 4 and the best value for ntree is 200.
- Executing the RandomForest() Alamer (2024) 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.

Conclusion

This section is under construction.

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.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2):1883.

Rahman, M. H. (2024). Kaggles predicting diabetes onset based on diagnostic measures.

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