Modeling Cardiovascular Disease

2021

Project Description

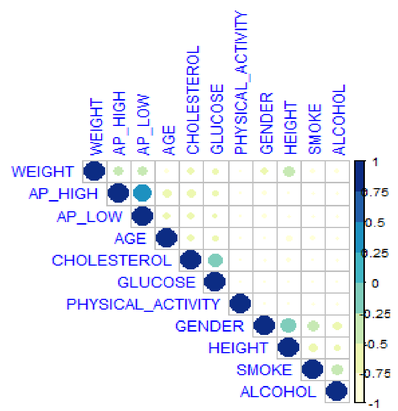
- Cardiovascular Disease is the number one cause of death in the United States resulting in 1 death every 36 second.
- Current diagnosis for Cardiovascular Disease requires many tests that can be expensive and time consuming.
- Supervised Machine Learning algorithms create predictive models to determine outcomes with high accuracy.

Data Source

- Developing a strong predictive model allows for major contributors to be identified.

Data Description

- Raise public awareness to decrease the number of deaths caused by the major contributors of the disease.
- Develop applications or websites that allows users to self-diagnose Cardiovascular Disease and seek medical attention.



Correlation Plot of the dataset predictors.

Method	Accuracy
Logistic Regression	72.89%
Support vector machine	72.43%
Random forest	74.29%
K nearest neighbor	72.21%

Methods Table comparing the accuracy percentages.

Team Members:

Joseph Barr
Linh Ta
Sophia Wang

Methodology

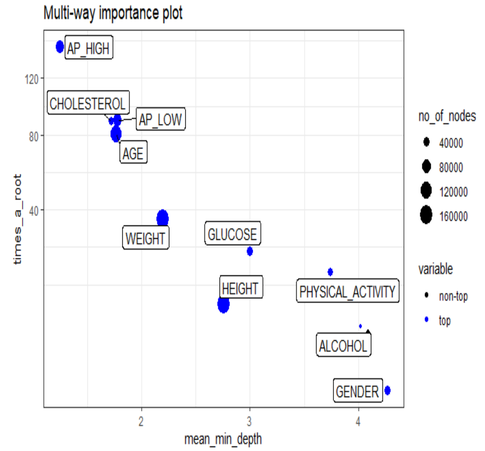
1. Initial investigation of the data showed that some predictors had a stronger correlation than others
2. Four separate models were built that classified the Cardiovascular disease.
3. The prediction accuracy of the methods were compared to find which provided the highest overall prediction rate.
4. The method with the highest prediction accuracy was then further investigated to determine the most important contributors to predicting Cardiovascular Disease in that method.

Results and Conclusions

1. The initial investigation showed that blood pressure was the most important predictor.
2. The model types used were logistic regression, support vector machine, random forest, and K nearest neighbor.
3. The random forest model proved to have the highest overall prediction accuracy at 74.29%.
4. The multi-way importance plot shows that the superiority of high blood pressure is transparent.
5. In conclusion, since the random forest searches for the best feature among random subsets of features, therefore, high blood pressure is the most important and dangerous symptom of Cardiovascular Disease.

Implementation (Tuning, R Functions, Algorithm)

Technical term 1: An output based on inputs.
Technical term 2: Models binary dependent variables.
Technical term 3: Hyperplane that divides dimensions into groups.
Technical term 4: Generates decision trees
Technical term 5: Finds the closest neighbor to each point



Multi-way importance contributors to Cardiovascular disease.

References

- "Heart Disease Facts." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 8 Sept. 2020, www.cdc.gov/heartdisease/facts.htm.
- James, Gareth, et al. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2021.
- V, David. "Cardiovascular Diseases Dataset (Clean)." *Kaggle*, 14 Mar. 2020, www.kaggle.com/aiaiaidavid/cardio-data-dv13032020/code.
- "What Is Cardiovascular Disease?" Www.heart.org, www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease.

Acknowledgments

This project was mentored by Helen Zhang, whose help is acknowledged with great appreciation.