# **Michael Gonzalez DSC 690 Winter 2020**

**Professor Williams**

***Can you avoid developing cardiovascular disease?***

# Abstract

This project will look at anonymous health insurance data from a Kaggle competition. This data will be used in developing a predictive model to see if it possible to find features that can be used as early signs of developing cardiovascular disease. The goal of this project is to use these features to detect cardiovascular disease at an early stage before, it become unpreventable. It is a better cost-effective way to prevent someone from developing a disease than trying to treat the person. Soon, doctors will be able to check for this type of disease and hopefully with help from a predictive model be able to prevent this disease from getting any worst.

# Background

The selected dataset comes from the Kaggle competition website that includes cardiovascular disease data that removes the identities of people who gave their consent on being included in this dataset. Below is a link to the dataset: <https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>

In relation to the first project, this dataset stays within the healthcare industry. From looking at this dataset, is it possible to classify if a person has a cardiovascular disease based on features found in this dataset? Is it possible to use biometrics and demographic data to help in detecting cardiovascular disease? Analyzing this dataset will give me another angle in how to properly develop prediction models. This type of predictive model could be useful in predicting early signs of someone developing a cardiovascular disease.

I experimented with three different types of predictive models with this dataset. At the core of this project, it is a classification of which person has a cardiovascular disease or not. I took advantage of using three distinct models that would be able to give me different insights. This is like listening to people with different point of views. If I were to only use one type of predictive model, there would be a higher chance of bias. I decided to use Logistic Regression, Supported Vector Machines (SVM) and Naïve Bayes Classifier. Since these predictive models have different classification systems, it is unlikely that the results will have the same level of biases. On the other hand, if more than one model has similar results. It is likely that the predictive models are more in line with each other.

# The Walk Through

Starting with the dataset from the Kaggle website that was based on cardiovascular disease data, which was in a CSV format. I decided to use the Pandas package to read the dataset. I have also used the Sci-kit Learn package to use the different predictive models with this dataset. This dataset has 13 variables and 70,000 rows of data that includes Age, Gender, Height, Glucose levels, Smoker, and more. This dataset is huge, and it will be great to use for developing a model that will be able to predict cardiovascular disease. The codebook for this dataset will be in the appendix page of this report. The appendix page is located towards the end of this report.

Data Attributes:

The following code was used to get the information about the data frames in terms of the different variables present in this dataset.

Graphical user interface, text, application, email

Description automatically generated

Data Frame Check of the first five rows:

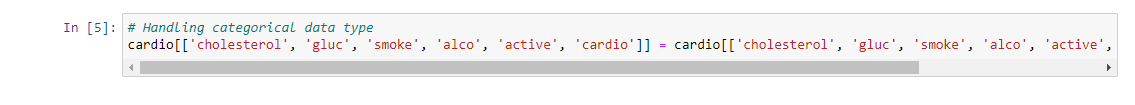
When I was checking the dataset, I had noticed that some of the variables needs to be encoded as categorical data types.

A picture containing text

Description automatically generated

Changing some variables to categorical data types:

The following code was used to change some variables to categorical data types to make it easier to with on, in future sections of this project.



Checking again of the first five rows:

I wanted to check the dataset again, after I encoded some variables as categorical data types.

A picture containing text

Description automatically generated

Checking the summary statistics from this dataset :

I wanted to check the summary statistics, to see if I can see any trends.

Graphical user interface

Description automatically generated with low confidence

# Model Fitting

Given this will be a classification problem, all the chosen models are going to different from each other. The first model is logistic regression. The next model well be support vector machine to see if I will be able to see any patterns that did not show up in the first model. The third model is naïve bayes classifier.

Splitting the data for the three models:

Graphical user interface, text, application, email

Description automatically generated

Using Standard Scaler on the data for the three models:

A picture containing application

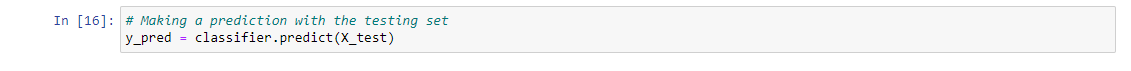
Description automatically generated

# Model Fitting (Logistic Regression)

Fitting data to Logistic Regression and the accuracy rate:

Graphical user interface, text

Description automatically generated



Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

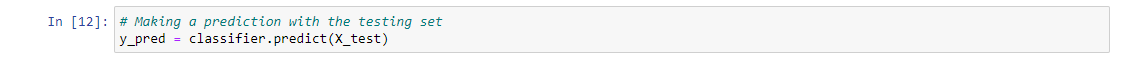
As you can see that the logistic regression model has an accuracy rate of 72%. This is confirmed with a classification report that was taken from the testing and predictions of the dataset.

# Model Fitting (Support Vector Machine)

Fitting data to SVM and the accuracy rate:

Application

Description automatically generated with low confidence



Graphical user interface, text, application, email

Description automatically generated

Wow, I could not believe that the logistic regression model and support vector machine model resulted with the same accuracy rate. This is really the best part of the report, when two models have the same outcome. After seeing this, I believe that any of these two models would have better accuracy rates with more time. Now, it is time to see what kind of results were given from naïve bayes classifier.

# Model Fitting (Naïve Bayes Classifier)

Fitting data to naïve bayes classifier and the accuracy rate:

A picture containing text

Description automatically generated



Graphical user interface, text, application, email

Description automatically generated

From the look of the accuracy rate of the naïve bayes classifier. The rate is not on the same level as the other two models. To be honest, 58% accuracy rate is not bad, but when comparing to the other models. It makes no sense to use the naïve bayes classifier on this dataset. Since the logistic regression model and the support vector machine model had the same results. I can say that using any of these two models would be great for any future analysis.

# Conclusion

The focus of this project was to see if it is possible to predict if a person would have a cardiovascular disease based on features. Developing and testing three predictive models, the accuracy of two models was fantastic. The logistic regression model and the support vector machine model were the best models that had a good accuracy rate. The accuracy rate from these models was 72% and with more time the accuracy rate could be higher.

# References

[1] Kirchner, T. How machine learning is revolutionizing insurance pricing at AXA. November 15, 2018. <https://digital.hbs.edu/platform-rctom/submission/predicting-your-casualties-how-machine-learning-is-revolutionizing-insurance-pricing-at-axa/> (Information on predicting insurance rates)

[2] Lawler, R. How K Health is using AI to change healthcare. November 20, 2019. <https://www.samsungnext.com/blog/how-k-health-is-using-ai-to-change-healthcare> (Information on using AI on healthcare data)

[3] Faggella, D. 7 Applications of Machine Learning in Pharma and Medicine. March 4, 2020. <https://emerj.com/ai-sector-overviews/machine-learning-in-pharma-medicine/> (Article on machine learning application in medicine)

[4] Kent, J. Using Big Data, Machine Learning to Reduce Chronic Disease Spending. September 18, 2018. <https://healthitanalytics.com/news/using-big-data-machine-learning-to-reduce-chronic-disease-spending> (Article on machine learning reducing insurance payment on chronic diseases)

[5] Nashif, S., Raihan, Md.R., Islam, Md.R. and Imam, M.H. (2018) Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. World Journal of Engineering and Technology, 6, 854-873. <https://www.scirp.org/journal/paperinformation.aspx?paperid=88650> (Article on machine learning prediction with cardiovascular diseases)

[6] Bhanot, K. Predicting presence of Heart Diseases using Machine Learning. February 12, 2019. <https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c> (Article on using machine learning to detect cardiovascular disease)

[7] Egerton, T., Sochima, E. & Deedam-Okuchaba, Fortune. (2019). A Model to Detect Heart Disease using Machine Learning Algorithm. INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING. 7. 1-5. 10.26438/ijcse/v7i11.15. <https://www.researchgate.net/publication/337657343_A_Model_to_Detect_Heart_Disease_using_Machine_Learning_Algorithm> (Article on machine learning used to find heart disease)

[8] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN COMPUT. SCI.* 1,345 (2020). <https://link.springer.com/article/10.1007/s42979-020-00365-y> (Article on heart disease prediction with machine learning)

[9] University of Southern California. Using AI to detect heart disease: Researchers apply machine learning to create a quick and easy method for measuring changes linked to cardiovascular disease. April 17, 2018. <https://www.sciencedaily.com/releases/2018/04/180417100547.htm> (Article on heart disease detection with machine learning)

[10] Baldie, J. Predicting Heart Disease Diagnoses with Machine Learning. January 27, 2018. <https://medium.com/@jonbaldie/predicting-heart-disease-diagnoses-with-machine-learning-2e1a8f5213f8> (Article on the use of machine learning for heart disease diagnoses)

**Appendix**

Dataset description:

There are 3 types of input features:

Objective: factual information;

Examination: results of medical examination;

Subjective: information given by the patient.

Features:

Age | Objective Feature | age | int (days)

Height | Objective Feature | height | int (cm) |

Weight | Objective Feature | weight | float (kg) |

Gender | Objective Feature | gender | categorical code |

Systolic blood pressure | Examination Feature | ap\_hi | int |

Diastolic blood pressure | Examination Feature | ap\_lo | int |

Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |

Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |

Smoking | Subjective Feature | smoke | binary |

Alcohol intake | Subjective Feature | alco | binary |

Physical activity | Subjective Feature | active | binary |

Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

All the dataset values were collected, at the time of the medical examination. This code book is the most up to date information based on the selected dataset.

**10 Questions**

1. What made you want to do a project on this topic?
2. Were there any issues finding any research materials?
3. Do you think machine learning could help in detecting cardiovascular disease?
4. Why are there many articles on cardiovascular disease and machine learning?
5. What can normal people benefit from this type of work?
6. Who do you think benefits more from this work, men, or women?
7. Did you think doctors would trust machine learning more than their own experience?
8. From the accuracy rate being ok, how will you make it higher?
9. What did the results say about being able to detect cardiovascular disease?
10. When the model did not provide useful insights, what did you do?