|  |
| --- |
| A picture containing cup, coffee, indoor  Description automatically generated |
| Heart Disease  Machine Learning Project |
| |  |  |  | | --- | --- | --- | | Pamela Neh Ambe , Rula Abu Affar, Sarah Parsons-Lappin, Sirlene Andreis | 12/12/20 | HD in Science in Data Analytics for Business | |

Machine Learning Project – Heart Disease Prediction

Rula Abu Affar Student No: sba20361

Sirlene Andreis Student No: 2017195

Pamela Neh Ambe Student No: sba20263

Sarah Parsons - Lappin Student No: sba20219

Under supervision of **Dr. Muhammad Iqbal**

**Table of Contents**

[Introduction 3](#_Toc58681295)

[Problem 3](#_Toc58681296)

[Motivation/Challenges 4](#_Toc58681297)

[Data description and Data Preparation 4](#_Toc58681298)

[Machine Learning Models discussed 7](#_Toc58681299)

[First Model: Decision Tree Classifier 7](#_Toc58681300)

[Second Model: Random Forest 8](#_Toc58681301)

[Third Model: Logistic Regression 9](#_Toc58681302)

[Third Model Logistic Regression using Hyperparameter and Grid Search CV 10](#_Toc58681303)

[Fourth Model: Artificial Neural Network 11](#_Toc58681304)

[Results 12](#_Toc58681305)

[Conclusions 13](#_Toc58681306)

[References 14](#_Toc58681307)

[Appendix 15](#_Toc58681308)

[Group Contribution 15](#_Toc58681309)

# Introduction

From 1990 to 2017, heart disease has been, by far the number one cause of death worldwide.

In 2017, 17.79 million people died from this disease according to the research made by Ritchie and Roser (2018), the death cause is on the top of the list and represents the double that the numbers caused by cancer.

Deaths by cause

Chart

Description automatically generated

Source: Ritchie and Roser, (2018).

Heart disease is a consequence of accumulated events related to our health behaviour and the main risk factor to develop it are physical inactivity, unhealthy diet, tobacco and harmful use of alcohol. According to the WHO (2017), the risks are more likely to happen with individuals that already have high blood pressure, are overweight or in obesity conditions and the ones with high indices of glucose. To prevent cardiovascular diseases, WHO suggests that small interventions can be put in place by individuals, such as smoking cessation (which research shows can reduce by 75 percent recurrent vascular problems), prevention of development of hypertension, implementation of regular physical activities, and reducing alcohol consumption.

# Problem

The main objective is to predict what is the probability of developing heart disease if you are part of the group who consume alcohol, smoke and have high levels of glucose?

# Motivation/Challenges

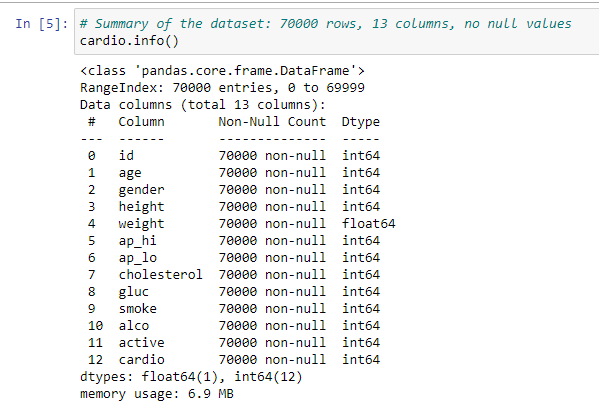
The motivation for us as a group is to explore the topic as we see the potential for further research.

We chose the Cardiovascular Disease dataset available at Kaggle to base our project on. The reason for the choice was the volume of data available and the various machine learning algorithms that could be applied to it.

The challenge was to find a dataset with the required information and in a format that we could use to base our models on.

# Data description and Data Preparation

This data set contains 13 columns and 70,000 rows. The features are:



Where cholesterol is a range (1: normal, 2: above normal, 3: well above normal), Gluc: glucose range between 1 to 3 levels. The variables Smoke, Alcohol, Active and Cardio are binary representations.

On our exploratory Data Analysis, we identified that there are no missing values in the data set but some features need to be feature engineered.

First, we drop the column Id as it does not have relevant value for future predictions.

Secondly, we feature engineer “age” that was originally in days, and we rounded it to years.

The “height” of 55cm and 250cm is not a normal measure, also with a “weight” of 10kg as that can only be considered for a child, however the minimum age found was 30 years which tells us there is an outlier present. To have a more accurate measure we drop the first and last 0.25 quantile for both measures, which eliminates the outliers.

Next, we engineered the height and weight features to form a new feature ‘BMI’.

The gender column was not clearly defined, we needed to look at the mean values for 1 and 2 to identify which were female. We did this by averaging the height of each class and considering the high value is allocated for men, so we end up having women = 1 and men = 2.

Then we eliminated negative min values for ap\_hi and ap\_lo and max values that do not make sense in medical records. By analysing the different values for Diastolic and Systolic blood pressure we decided to categorize by levels of incidence that range between 1 to 5 as described below, and add as a new feature called “BPC”, and drop the columns “ap\_hi” and “ap\_lo”.

According to the American Heart Association (AHA), Blood pressure range by:

* Category 1 = (ap\_hi) systolic pressure < 120 and (ap\_lo) diastolic pressure < 80 mm Hg.
* Category 2 = (ap\_hi) systolic pressure >= 120 & < 130 and (ap\_lo) diastolic pressure < 80 mm Hg.
* Category 3 = (ap\_hi) systolic pressure >= 130 & < 140 and (ap\_lo) diastolic pressure >= 80 & < 90 mm Hg.
* Category 4 = (ap\_hi) systolic pressure >= 140 & < 180 and (ap\_lo) diastolic pressure >= 90 & < 120 mm Hg.
* Category 5 = (ap\_hi) systolic pressure >= 180 and (ap\_lo) diastolic pressure >= 120 mm Hg.

The category 1 and 2 are considered normal blood pressure and do not require any intervention.

Category 3 is High blood pressure and considered stage 1 hypertension.

Category 4 is a more serious condition and considered stage 2 hypertension.

Category 5 is a serious health problem and requires urgent treatment.

After feature engineering we and up with 62772 entries and 12 columns that are:

Graphical user interface, text

Description automatically generated

During the data exploration phase, we tried to find out the feature which has the most significant impact over the prevalence of heart disease, by using the count method, we know that the difference of patients that do not have cardiovascular disease are nearly 2% higher than the ones that have cardiovascular disease.

Chart

Description automatically generated

We created a correlation chart that shows us none of the features has significant impact over the heart disease, as it is a combination of many features that causes heart disease. Since it is a balanced dataset, we can normalize it and begin to implement models that help us to discover the pattern of this combination.



# 

# Machine Learning Models discussed

In all the models commented below we had split 3 times the dataset, test respectively 10%, 20% and 30%.

## First Model: Decision Tree Classifier

We initially trained the model with a max depth = 3 achieving accuracy 0.7162 in the train and 0.7193. To build our model of Decision Tree Classifier we use the Gini criterion and define the max depth = 5.

The test condition in this model depends on attribute types binary, nominal, ordinal and continuous. Our model has these types of attributes, this is one of the reasons why it was chosen, in the build tree in first split 10% test and 90% training, the model predicts train accuracy equal to 0.7236 and test accuracy 0.7242, the tree display BPC and split between two groups age and cholesterol as both are continuous attributes.

The performance of a tree can be further increased by pruning. It involves removing the branches that make use of features having low importance. This way, we reduce the complexity of a tree, and thus increasing its predictive power by reducing overfitting.

The root of the tree is BPC where it shows high level of impurity GINI=0.5 which mean half of people have BPC more than 3.5 and the other half have BPC less than 3.5, people who has high BPC,

People who have BPC more than 3.5 and cholesterol less than 1.5 and alcohol more than 22.6 are more likely to develop heart disease, while people who have BPC less than 3.5 and they are less than 54.5 years old and have cholesterol less than 2.5 are less likely to develop heart disease.

Graphical user interface, application

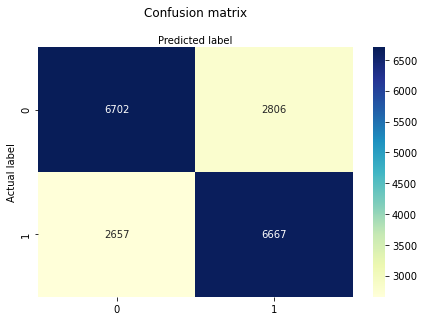
Description automatically generated

## Second Model: Random Forest

We applied Random forest trying to choose the best split among large number of individual decision trees (100-200) with a maximum feature of 5 and a maximum row of 32, Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction, we got accuracy of 72%.

## Third Model: Logistic Regression

Logistic regression is used when the dependent variable is a binary variable, what was the case of this study, cardio is represented as a binary variable.

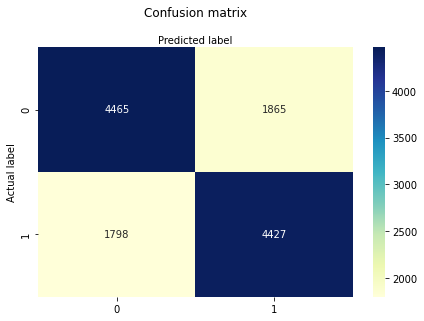


At 70% split the results were:

Accuracy: 0.7099086661002549

Precision: 0.7037897181463105

Recall: 0.715036465036465

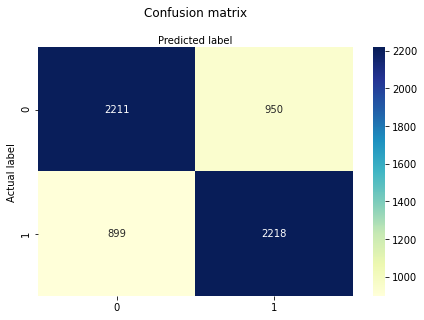


At 80% split the results were:

Accuracy: 0.7082437275985664

Precision: 0.7035918626827717

Recall: 0.7111646586345381



At 90%, we got the follow results:

Accuracy: 0.7054794520547946

Precision: 0.7001262626262627

Recall: 0.711581649021495

Precision=TP/(TP+TN)

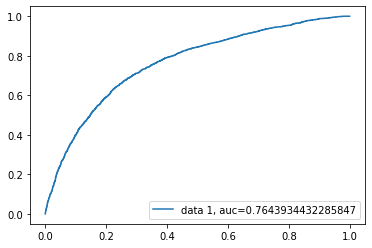
Recall= TP/(TP+FN)

Accuracy= (TP+TN)/Total

The metrics accuracy score had a precision of 0.7099 (71%), where the precision was 0.7037, which means how correct the prediction was. And the recall score identifies that 71.5% of the time patient has cardio condition. From observation there are no significant differences when split in different percentages, regarding accuracy, recall, and precession.

The AUC score was 0.76 in this case, which is not so bad, the perfect classifier should be 1 and 0.5 is considered worthless.

ROC curve (split 10/90%)



## Third Model Logistic Regression using Hyperparameter and Grid Search CV

After we applied our model, we start to train our model by testing it using different value of parameter C range [0.01-0.99] the best accuracy we get when C= 0.01

Average cross-validated in-sample F1 score 0.7054 {'C': 0.01}

Confusion matrix showing the actual numbers v predicted:

Chart, treemap chart

Description automatically generated

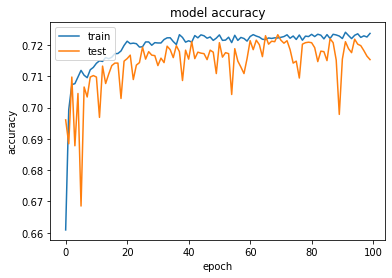
## Fourth Model: Artificial Neural Network

In our ANN model, we decided to go for the Sigmoid function to develop the mode activation in the three different levels of Dense, because this function squashes the values between 0 and 1. The number of units selected for each layer in Keras is extremely important as we increase the units consequently increasing the complexity of the model and more weights to be trained, our sequential units were 32, 8 and 1. In total we used 11 input dimensions as we have 11 features in our dataset.

The model compile used considered the loss as binary cross entropy, with optimizer Adam and the metrics look for return an accuracy number. We have tried different variations of the epoch and decided go for 100 with a batch size 16.

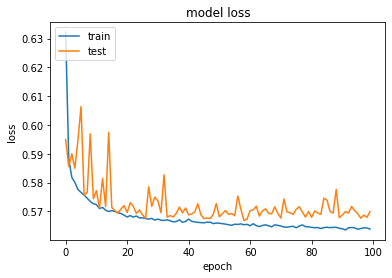
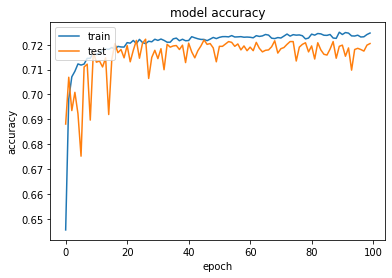
There was no significant difference when we applied 50, 100 or 200 epoch interactions, we have also tried increase the batch size for 32 instead of 16 and still got similar accuracy. When applied Relu activation instead of Sigmoid we got significantly worse results.

The two plots bellow shows the model accuracy and model loss for a train/test split 90/10 with 100 epoch interactions, in the blue line we can identify a constant training path after the 30 epochs, while the test has some scalation along the line.

A picture containing histogram

Description automatically generated

As we increase the train/test batch to 70/30, the train line start got inconstant pics and test accuracy got higher pics while the epoch number increase.



# Results

The results of the five models in machine learning we applied on our data set shows no significant difference in terms of accuracy. From the below table we could observe that Neural network model has the highest accuracy as the model is built on bases of algorithm and mathematical activities.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model name | Decision Tree | Random Forest | Neural Network | Logistic Regression | KNN |
| Accuracy  at 90% train | train 72.36%  test 72.42% | test 71.93% | 72.57% | train 70.86%  test 70.55% | 66.43% |
| Accuracy at  80% train | train 72.36%  test 72.06% | test 72.00% | 72.64% | train 70.86%  test 70.82% | 66.02% |
| Accuracy at  70% train | train 72.36%  test 72.19% | test 72.00% | 72.70% | train 70.81%  test 70.99% | 66.23% |

# 

# Conclusions

We uncovered some interesting facts about heart disease prediction during the course of the project.

That hypertension, BMI and age are the strongest correlated variables to heart disease. With surprisingly low correlation results for alcohol and smoking. We struggled to confirm these results as we didn’t have information as to how this data was collected. For this reason we checked with a domain expert (cardiac nurse) who confirmed that this is the case.

On real live data, we would have preferred to see a higher accuracy across the models, given that it is such a serious condition. However there are other factors during diagnosis such as genetics, medical history & stress that can’t be accounted for in our dataset.

# References

Ritchie, H. and Roser, M., 2018. Causes of Death [Journal] Our world in data. Available at: https://ourworldindata.org/causes-of-death, [Accessed 28th November 2020].

Ulianova, S., 2018. Cardiovascular Disease dataset[online] Kaggle. Available at: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset, [Accessed 14th November 2020].

WHO – World Health Organization, 2017. Cardiovascular diseases (CVDs). [online] Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=Key%20facts,to%20heart%20attack%20and%20stroke, [Accessed 28th November 2020].

# 

# A**ppendix**

## Group Contribution

During this project, we worked together as a group with our own individual roles & responsibilities.

|  |  |  |
| --- | --- | --- |
| Team Member | Role | Initial Responsibilities |
| Rula | Project Manager | Give direction to the team members, create schedule and project timeline.  Arrange meetings.  Review documentation before final submissions to & edit where necessary. |
| Sirlene | Researcher | Research data set for analysis possibilities in relation for business value.  Research of further data sets for use.  Research of analysis / modelling techniques |
| Pamela | Researcher | Research data set for analysis possibilities in relation for business value.  Research of further data sets for use.  Research of analysis / modelling techniques |
| Sarah | Communications  Co-ordinator | Documentation Editing  Set up & manage a shared folder for documentation |

Although we had our individual roles and responsibilities we all participated in each area of the project:

* Dataset Selection
* Feature Engineering
* Data Cleaning
* Model Selection and Implementation
* Visualisation
* Report Writing

We assigned each of the team with a model & collaborated in the training of these models to increase the accuracy.

We held regular Microsoft Team meetings.

We used WhatsApp as a means of communication & Dropbox to collaborate on documents.