BlackCat Capstone Project

What type of patients are at a higher risk of getting a Cardiovascular disease

RUCHITA MANUJA

4/JUNE/2019

Table of Contents

Introduction

According to a study conducted by Heart Foundation, Cardiovascular disease is a major cause of death in Australia, with 43,477 deaths attributed to CVD in Australia in 2017. Cardiovascular disease kills one Australian every 12 minutes. [1]

This study summarizes the results of the analysis conducted on a dataset of patient's medical examination and lifestyle to find the factors and what type of patient's are at a higher risk of getting affected by a Cardiovascular disease. Based on the analysis four classification models (KNN, Logistic Regression, Naïve Bayes and Random Forest) have been created which be used as a preliminary test for predicting whether a patient is at risk of getting affected by a disease or not.

Dataset Description

Cardiovascular disease dataset has been retrieved from the Kaggle (https://www.kaggle.com/sulianova/cardiovascular-disease-dataset_

The data is mainly in raw form (not scaled) and contains binary information (0 or 1) for the qualitative independent variables.

A summary of the dataset and its associated task is provided below:

Title	Cardiovascular Disease dataset
Dataset Associated Task	Classification
Number of observations	70,000
Number of Explanatory features	12
Missing Values	Null
Target	Presence or absence of cardiovascular disease

Table 1 - Data Summary Statistics

The Features information is displayed in Table 1 below:

Name	**Data Type**	**Measurement**	**Description**
Age	quantitative	Days	Age in Days
Height	quantitative	Centimetres	Height in cm
Weight	quantitative	Kg	Weight in kg
ap_hi	quantitative	mmHg	Systolic Blood Pressure
ap_lo	quantitative	mmHg	Diastolic blood pressure
Cholesterol	qualitative	1: normal, 2: above normal, 3: well above normal	Level of Cholesterol
Glucose	qualitative	1: normal, 2: above normal, 3: well above normal	Glucose Level

Smoking	Binary	1 = yes; 0 = no	Patient smokes or not
Alcohol intake	Binary	1 = yes; 0 = no	Patient consumes alcohol or not
Physical activity	Binary	1 = yes; 0 = no	Patient is involved in physical activity or not

Table 2- Dataset Features

Data Exploration

The dataset is mostly clean i.e contains no missing or noisy data. Each data point is associated with one of the binary classes of whether a patient is suffering from a cardiovascular disease or not.

Before analysing the data further the following actions were performed to assist with the analysis:

- 'Id' column was dropped as the id's in this column did not appear sequential and there were missing ids
- The Categorical columns were converted to Category datatype to assist us in creating dummies later.

Figure 1 displays the summary statistics of the raw data.

	age	height	weight	ap_hi	ap_lo	age_in_years
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000
mean	19468.865814	164.359229	74.205690	128.817286	96.630414	53.338686
std	2467.251667	8.210126	14.395757	154.011419	188.472530	6.765294
min	10798.000000	55.000000	10.000000	-150.000000	-70.000000	30.000000
25%	17664.000000	159.000000	65.000000	120.000000	80.000000	48.000000
50%	19703.000000	165.000000	72.000000	120.000000	80.000000	54.000000
75%	21327.000000	170.000000	82.000000	140.000000	90.000000	58.000000
max	23713.000000	250.000000	200.000000	16020.000000	11000.000000	65.000000

Figure 1 - Data Summary Statistics

The following observations can be made from the above summary statistics:

- From the height and weight columns, it can be noticed that minimum height is 55 cm and minimum weight is 10 kg. This appears to be an error, since minimum age is 10798 days/30years.
- Also the maximum height is 250 cm and the maximum weight is 200 kg which might be irrelevant for generalizing our data.
- We also see that our minimum value is -150 i.e a negative value. Negative value for Blood Pressure is also incorrect.
- The maximum value for ap_hi and ap_li is also extremely large. According to http://www.bloodpressureuk.org/microsites/u40/Home/facts/Whatisnormal the Blood pressure usually ranges between 90 to 250 for the top or maximum number (systolic) and 60 to 140 for the bottom or minimum number (diastolic)

Data Cleaning

Although the data does not have any missing value as seen in Figure 2, it does have some errors as described in the section above.

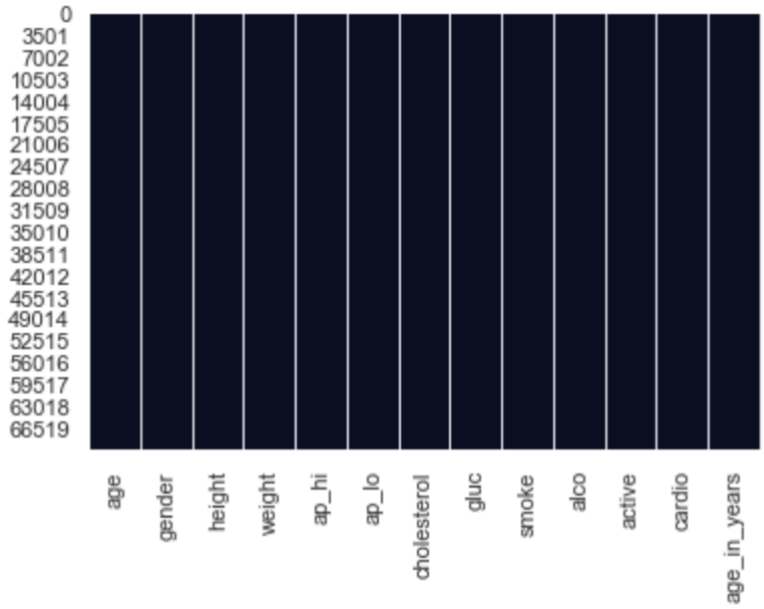


Figure 2 - Missing Values

The errors were investigated individually as described below:

1. **Error 1 – Negative Blood Pressure**
   The blood pressure cannot be negative. Both Systolic (ap_hi) and Diastolic (ap_lo) Blood Pressure were individually investigated. The data consisted of 7 negative values for Systolic Blood Pressure and 1 negative value for Diastolic blood pressure.

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	age_in_years
4607	15281	1	165	78.0	-100	80	2	1	0	0	1	0	42.0
16021	22108	2	161	90.0	-115	70	1	1	0	0	1	0	61.0
20536	15581	1	153	54.0	-100	70	1	1	0	0	1	0	43.0
23988	18301	1	162	74.0	-140	90	1	1	0	0	1	1	50.0
25240	14711	2	168	50.0	-120	80	2	1	0	0	0	1	40.0
35040	23325	2	168	59.0	-150	80	1	1	0	0	1	1	64.0
46627	23646	2	160	59.0	-120	80	1	1	0	0	0	0	65.0

Figure 3 - Negative Systolic Blood Pressure Values

Figure 3 displays the negative systolic blood pressure values under the ap_hi colum. Looking at this values it could be an error during measurement where a '-' symbol may have been added in the front as the absolute value of this satisfies the rule of higher value of Systolic to Diastolic. Therefore, these negative values were converted to positive values by taking their absolute values.

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	age_in_years
60106	22571	1	167	74.0	15	-70	1	1	0	0	1	1	62.0

Figure 4 - Negative Diastolic Blood Pressure Values

Figure 4 displays the negative value for diastolic blood pressure under the ap_lo column. As there is only one value as such this was also converted to a positive value similar to the systolic blood pressure. However, this led to another discovery where the value of Diastolic Blood Pressure became higher than Systolic Blood Pressure.

Blood pressure values should fall within the following range [2].

- **Systolic:** 90 to 250 mmHg
- **Diastolic:** 60 to 140 mmHg

Any value outside of this range could be an error. Therefore any values outside the range above were dropped. This reduced our Number of observations to 68523.

Another consideration with Blood Pressure values is that the Systolic Blood Pressure is always higher than Diastolic Blood Pressure. The dataset has 24 cases where this was actually reverse and all these 24 rows were dropped , thus reducing the number of observations to 68499.

2. **Error 2 – Height & Weight**

To assist our investigation for Height & Weight a new feature for Body Mass Index (BMI) was created. Figure 5 shows the distribution of BMI and we can observe. The BMI range starts at a very low value of 3.47 and maximum is 298. These values are way to low or high for a BMI, because of the very low or very high weight and height values in our dataset.

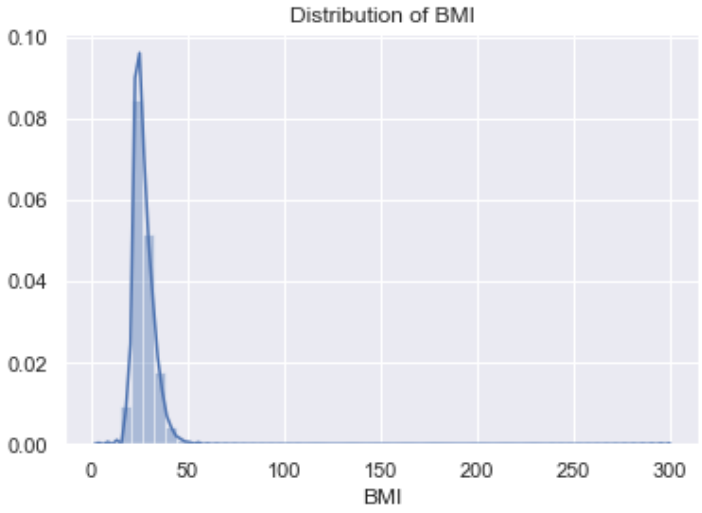


Figure 5 - BMI Distribution before cleaning

To have a better distribution for BMI, Height and Weight observations the first and last data quartiles were dropped. This reduced our number of observations to 65081, however the BMI distribution is well distributed as shown in Figure 6.

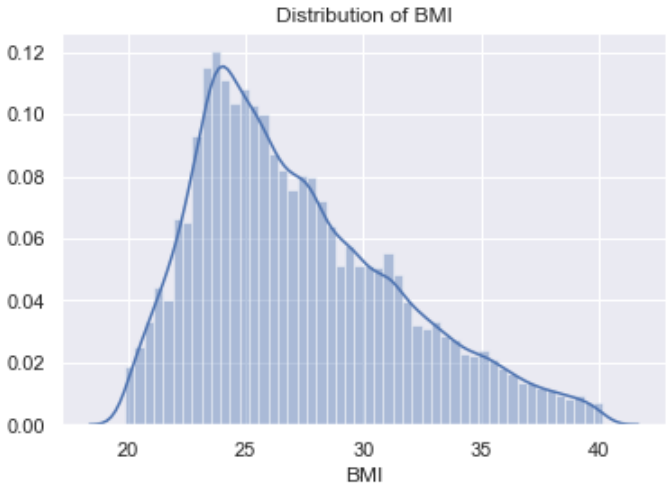


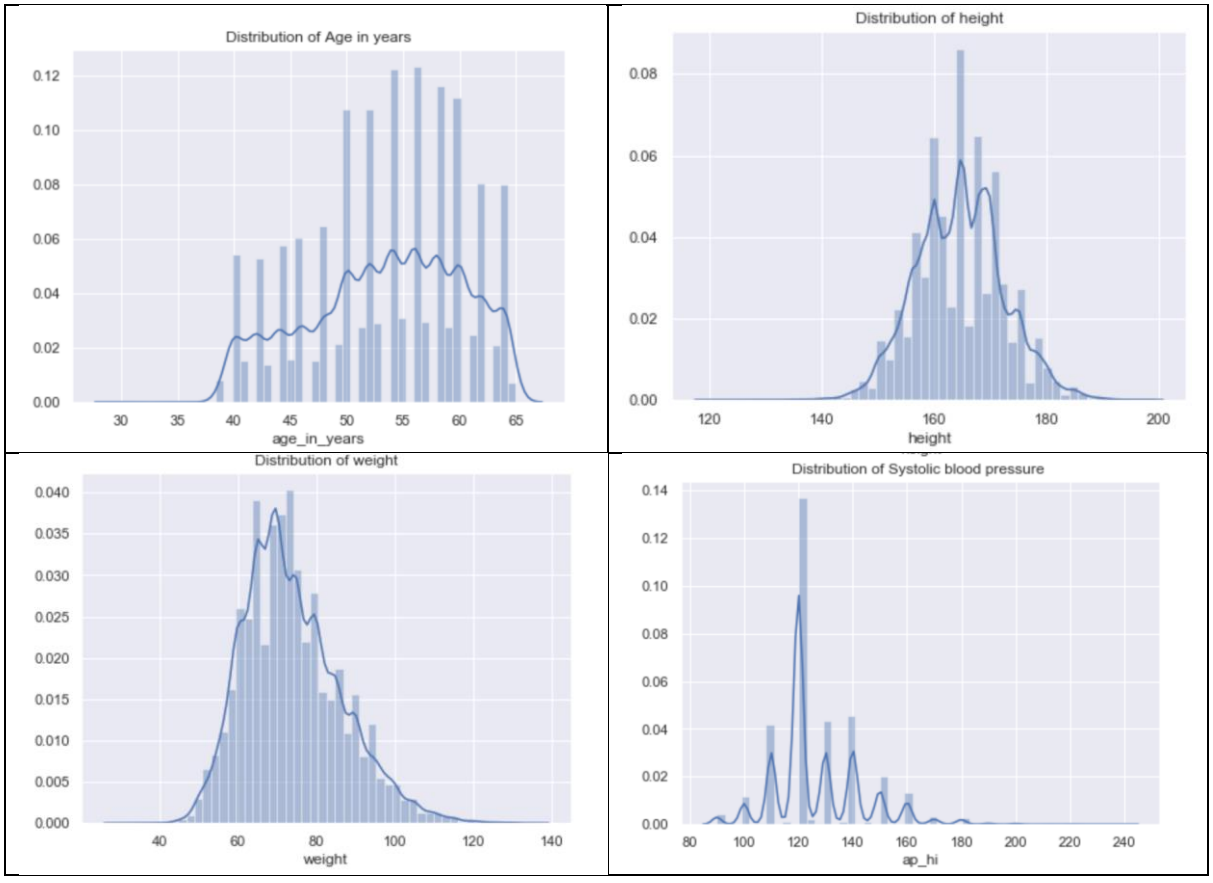
Figure 6 - BMI Distribution after dropping quartiles

Data Analysis

Class Distribution

Our class distribution is well balanced with 50.39% records for patients with no Cardiovascular disease and 49.61% records with patients who have a Cardiovascular disease.

Distribution of Continuous Features



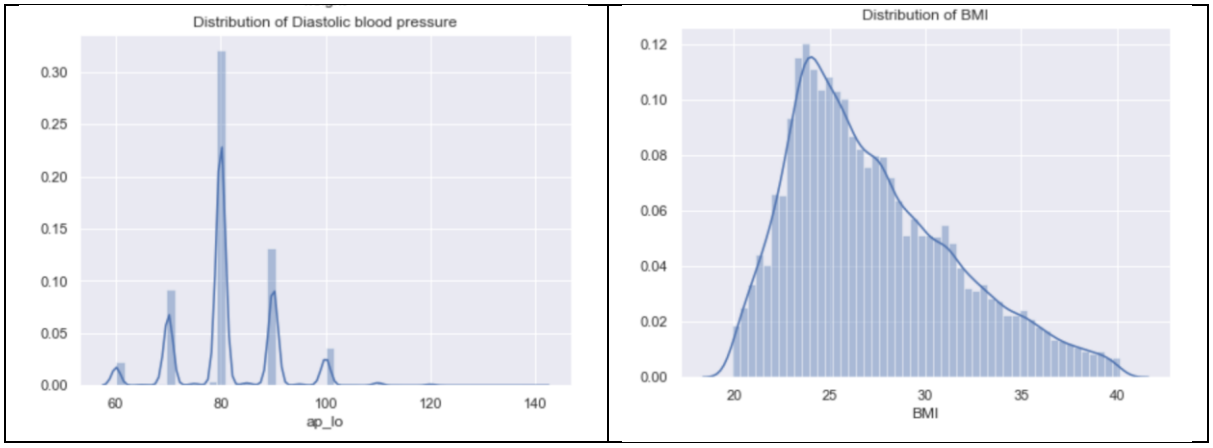
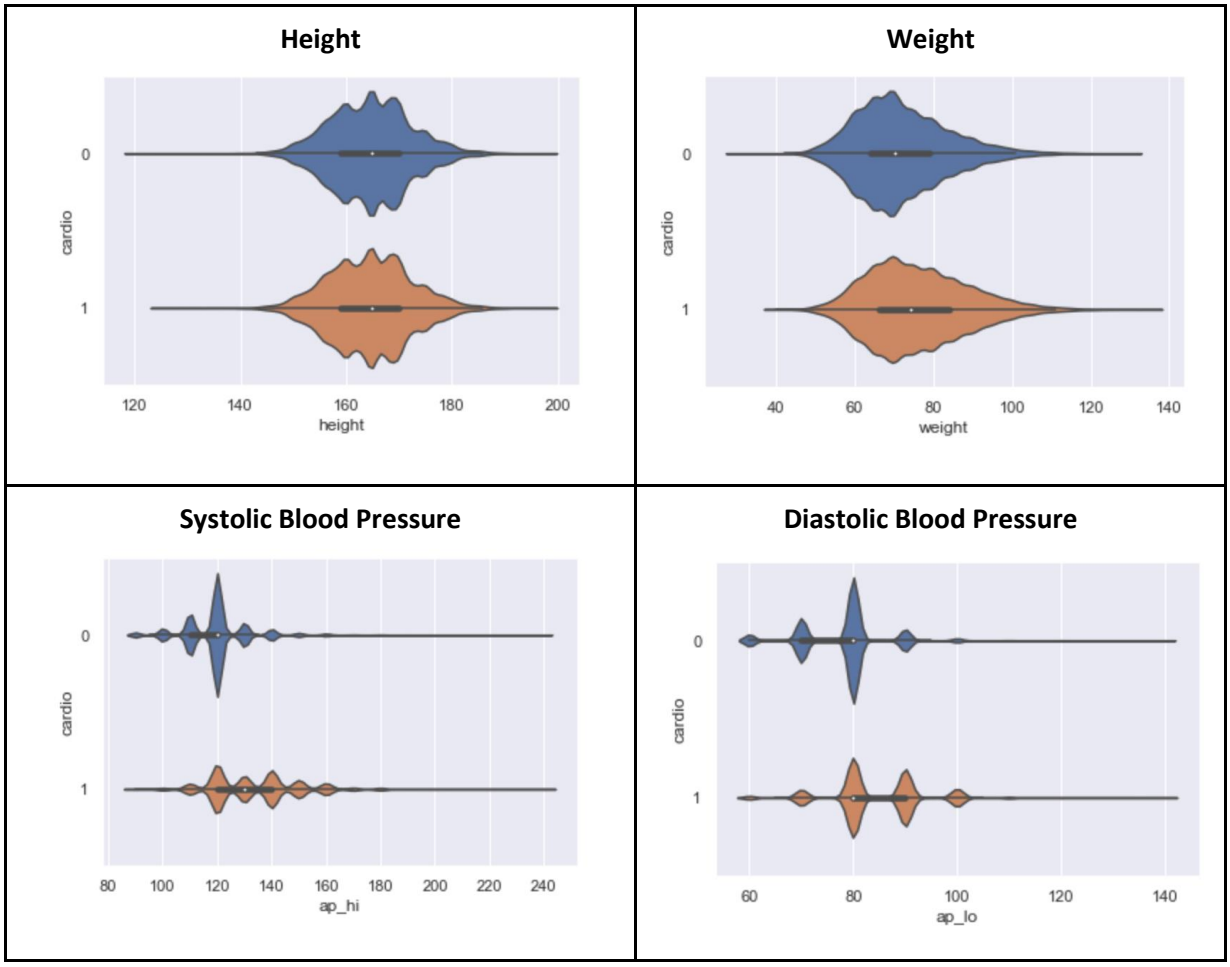


Table 3 - Distribution of Continuous Features

The above distribution charts in Table 3 show that most of our continuous variables are mostly normally distributed across the mean. We do some peaks in the age and blood pressure values. However, it is good to know that we some have values on the extreme ends values as it will be good to determine whether blood pressure or age affects the chance of getting a Cardiovascular disease.

Relationship between continuous features and presence of Cardiovascular disease



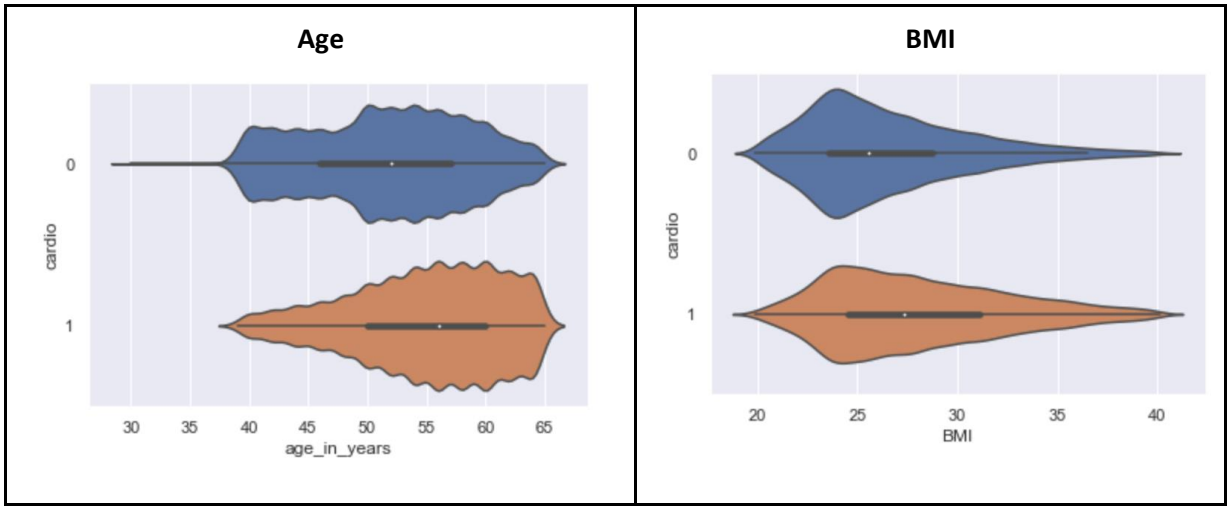


Table 4 - Relationship between continuous features and presence of Cardiovascular disease

The charts in Table 4 show the comparison between the distributions of various features with respect to disease. These could be useful features in predicting the presence of a cardio vascular as the distributions are quite distinct for each type.

Correlation

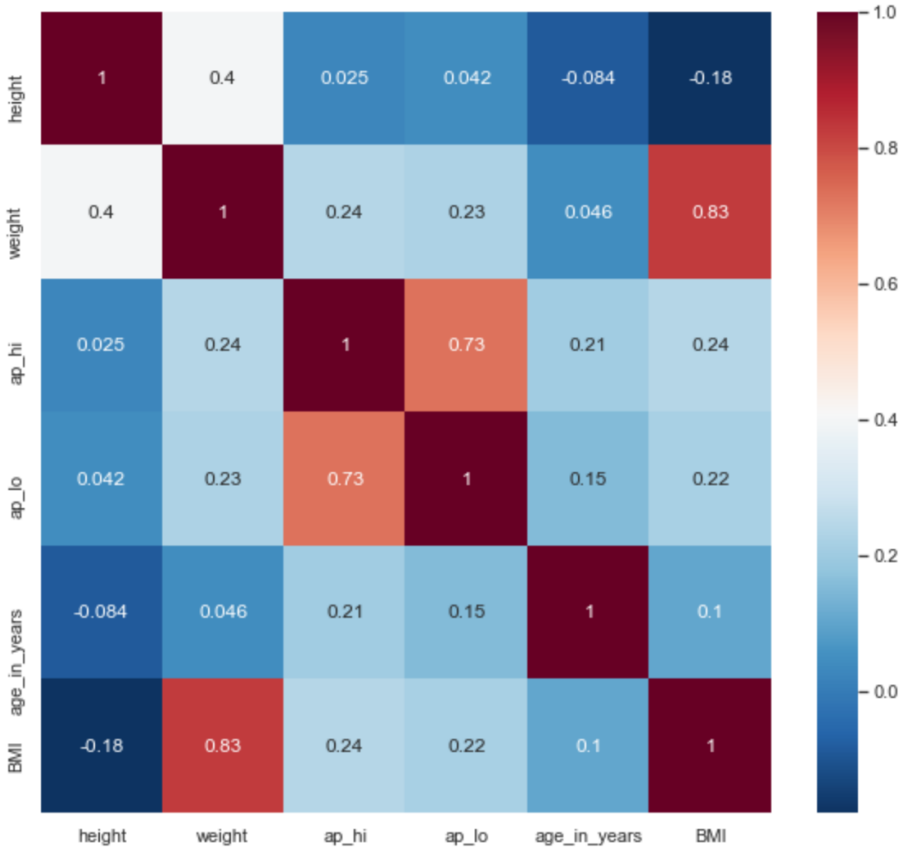


Figure 7 - Correlation Plot for Continuous Features

Figure 7 shows the Correlation matrix for the continuous features. From the Correlation matrix we don't see a very strong correlation among our features. Among the correlations the strongest positive correlation is observed for BMI & weight (0.83) and Systolic & Diastolic Blood Pressure (0.73)

BMI Analysis

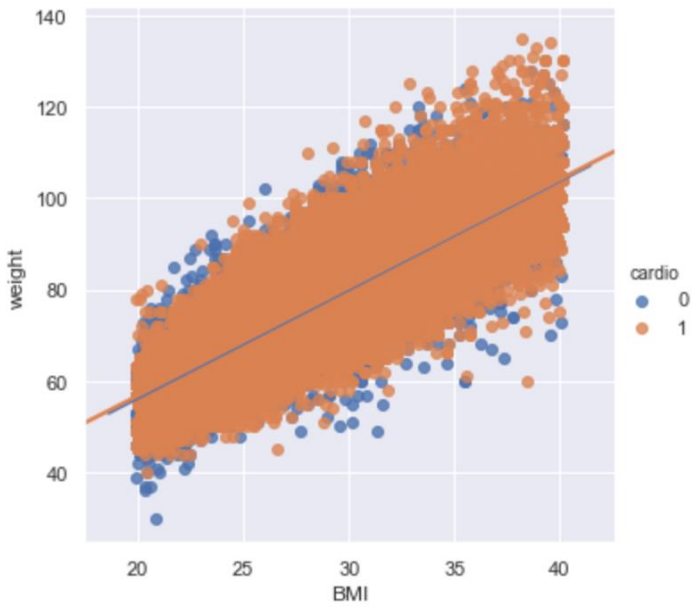


Figure 8 - BMI vs Weight (Scatter Plot)

Figure 8 displays the relationship between BMI and Weight. As expected the BMI increases with increase in weight, however we there is no difference whether a person is prone to get a Cardio Vascular disease or not as that appears to be equally distributed.

Gender Analysis

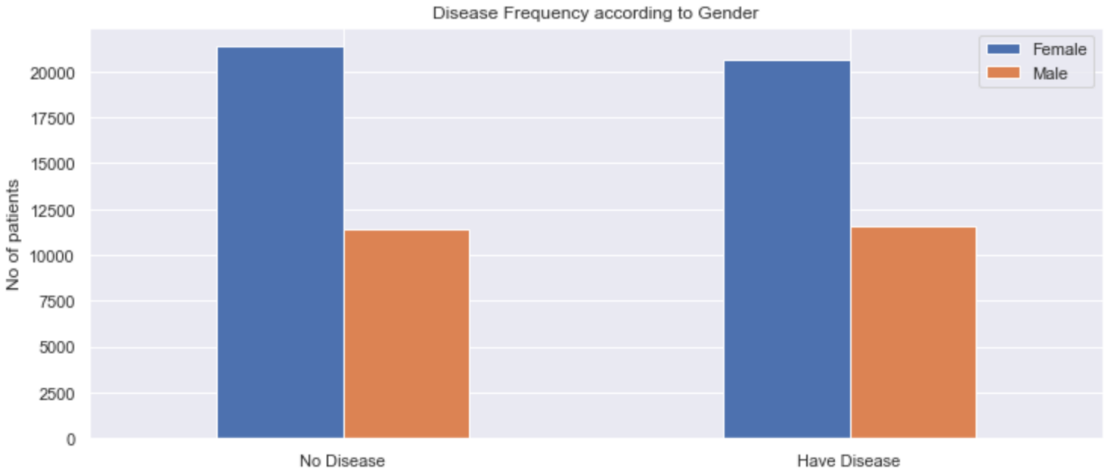


Figure 9 - Disease Frequency as per Gender

Figure 9 displays the frequency of disease presence for each Gender which when converted to percentages is as below:

- **Male** with **no** disease: 49.65%
- **Female** with **no** disease: 50.79%
- **Male** with **disease** 50.35%
- **Female** with **disease** : 49.21%

It appears that Males have a slightly higher risk of getting a heart disease than Females. But this is again very minimal i.e 1%.

Age Analysis

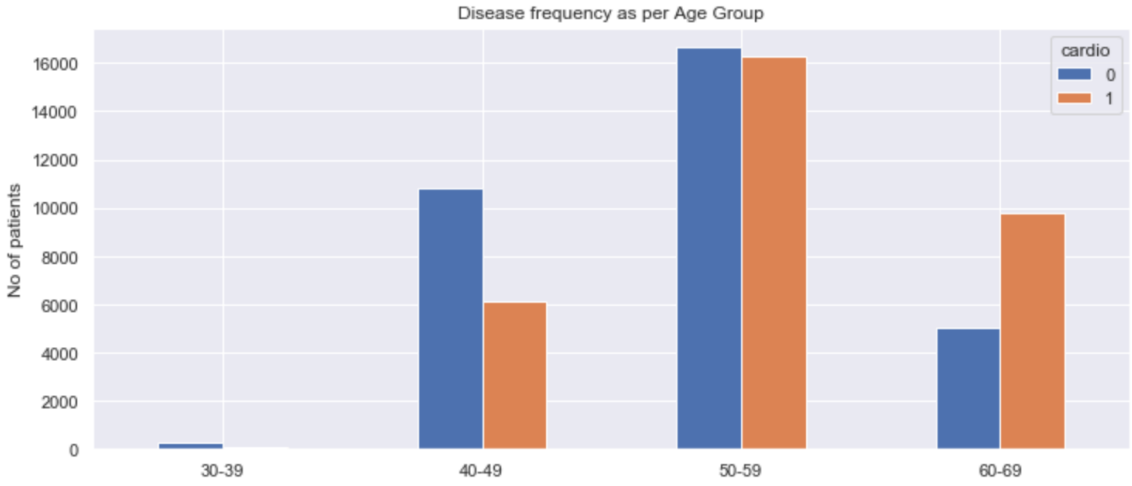


Figure 10 - Disease frequency as per age group

Figure 10 displays the frequency of disease presence for each Gender. We can see that number of patients with disease increases in comparison with the number of patients with no disease with increasing age.

Categorical Variables Analysis

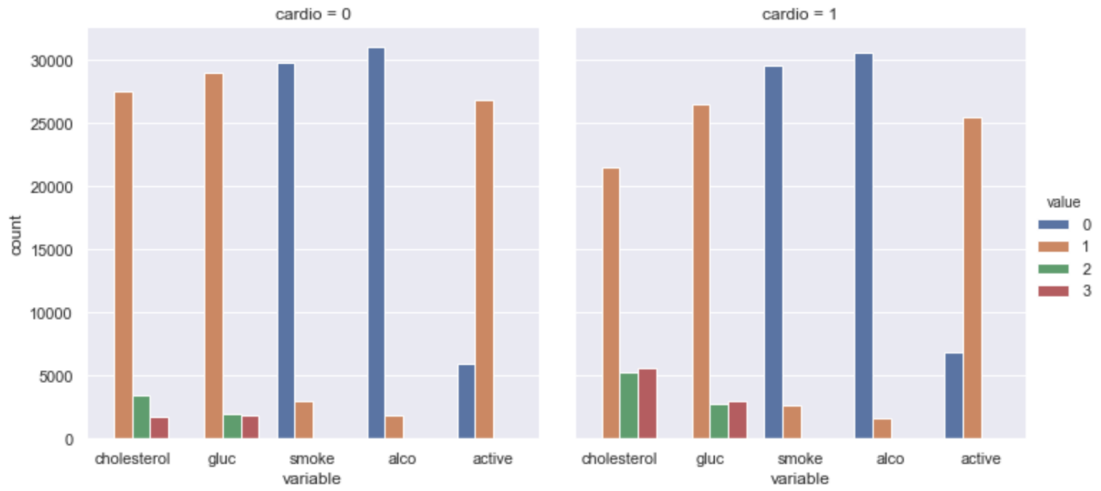


Figure 11 - Categorical Variables Analysis

To understand the graph in Figure 11 the following conventions are to be noted:

- Cholesterol - 1: normal, 2: above normal, 3: well above normal
- Glucose - 1: normal, 2: above normal, 3: well above normal
- Smoking - (1 = yes; 0 = no)
- Alcohol intake - (1 = yes; 0 = no)
- Physical activity - (1 = yes; 0 = no)

The following observations were made from the bar plots in Figure 10:

- Patients with Cardiovascular disease have higher cholesterol & blood glucose level.
- Patients with Cardiovascular disease are also less active compared to patients who do not have a cardiovascular disease.
- Smoking and Alcohol don't seem to contribute much to whether a patient has a disease or not.

Data Pre-processing

Feature Engineering

Using the height and weight a new feature was created for the Body Mass Index during analysis. As BMI includes the height and weight , the height and weight columns were not used in our models.

One-hot Encoding & Standardization

The original dataframe was initially split into two separate dataframes. One for categorical variables and another for Continuous variables. One-hot encoding was applied to the Categorical variables to represent them as binary values. For the continuous variables standardization was applied as our values are on different scales.

After applying the individual operations on each individual dataframe, they were joined to form one dataframe.

Splitting the dataset

The dataset was split with 70% for training purposes and 30% for testing purpose.

Modelling

The experiments were conducted in Google Colab using a GPU runtime to speed up the performance.

Model 1 – K Nearest Neighbours

GridSearchCV operation was performed to find the best possible parameters to use. The best parameters were **{'metric': 'euclidean', 'n_neighbors': 49, 'weights': 'uniform'}.**

The model was created as per the above parameters with a 10-fold cross validation.

The results of KNN are as below:

Accuracy: 72% **, Runtime:** 4.08 sec

	precision	recall	f1-score	support
0	0.70	0.78	0.74	9838
1	0.75	0.66	0.70	9687
accuracy			0.72	19525
macro avg	0.72	0.72	0.72	19525
weighted avg	0.72	0.72	0.72	19525

KNN ROC Curve:

The ROC Curve for KNN classifier is displayed in figure 12 below. The AUC is 0.720 , therefore the model is predicting 72% of the values correctly.

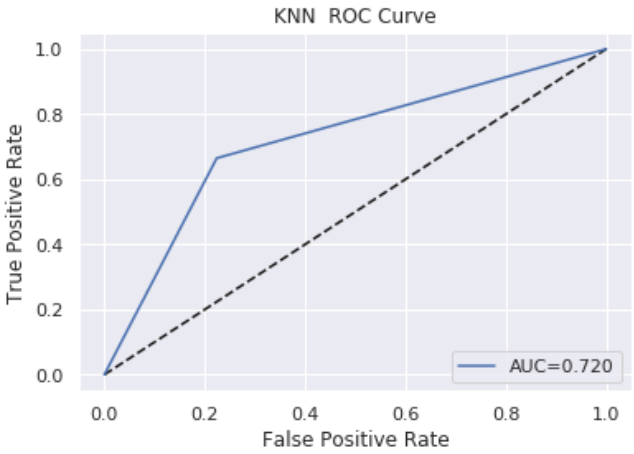


Figure 12 - KNN ROC Curve

Model 2 – Logistic Regression

GridSearchCV operation was performed to find the best possible parameters to use. The best parameters were **{'C': 1.0, 'solver': 'liblinear'}.** The penalty was set to L2.

The model was created as per the above parameters with a 10-fold cross validation.

The results of Logistic Regression are as below:

Accuracy: 71% **, Runtime:** 0.84 sec

	precision	recall	f1-score	support
0	0.69	0.77	0.73	9838
1	0.74	0.65	0.69	9687
accuracy			0.71	19525
macro avg	0.72	0.71	0.71	19525
weighted avg	0.72	0.71	0.71	19525

Logistic Regression ROC Curve:

The ROC Curve for Logistic Regression classifier is displayed in figure 13 below. The AUC is 0.778 , therefore the model is predicting 77.8% of the values correctly.

Figure 13 - Logistic Regression ROC Curve

Model 3 – Gaussian Naïve Bayes

The Naïve Bayes model was created with **Priors = None** with a 10-fold cross validation.

The results of Naïve Bayes are as below:

Accuracy: 69% , **Runtime:** 0.15 sec

	precision	recall	f1-score	support
0	0.66	0.80	0.72	9838
1	0.74	0.58	0.65	9687
accuracy			0.69	19525
macro avg	0.70	0.69	0.69	19525
weighted avg	0.70	0.69	0.69	19525

Naïve Bayes ROC Curve:

The ROC Curve for Logistic Regression classifier is displayed in figure 14 below. The AUC is 0.692 , therefore the model is predicting 69.2% of the values correctly.

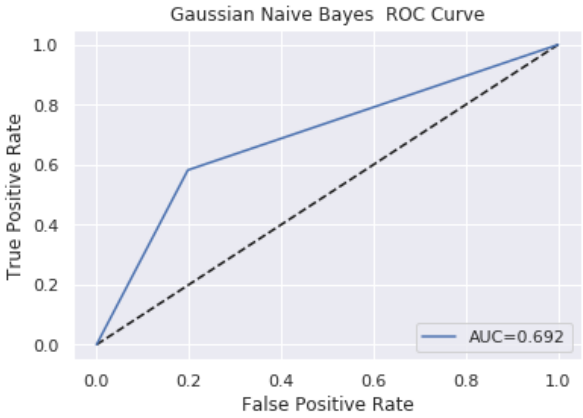


Figure 14 - Naive Bayes ROC Curve

The Naïve Bayes is performing worst than our KNN and Logistic Regression so it is not a model that we can choose our prediction.

Model 4 – Random Forest

For Random Forest, initially the model was created using the default parameters with a 10-fold cross validation.

The results of the base model are as below:

Accuracy: 67.6% , **Runtime:** 2.73 sec

	precision	recall	f1-score	support
0	0.66	0.69	0.68	9838
1	0.67	0.64	0.66	9687
accuracy			0.67	19525
macro avg	0.67	0.67	0.67	19525
weighted avg	0.67	0.67	0.67	19525

Random Forest ROC Curve for default parameters:

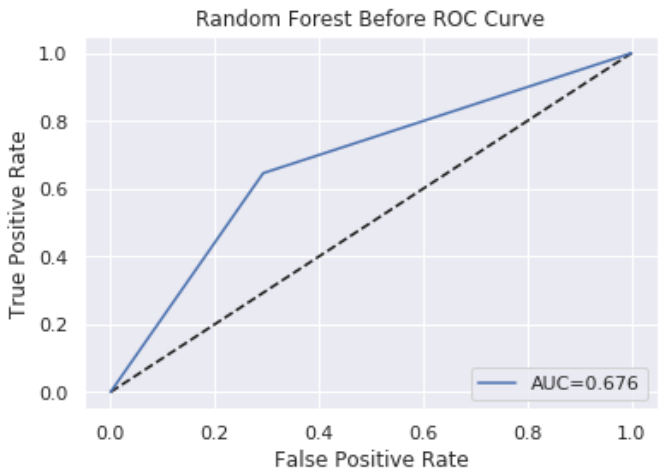


Figure 15 - Random Forest ROC Curve (Default Parameters)

Improving the Performance of Random Forest

To improve the performance of random forest functions were created to find the right parameter for the following parameters by plotting the AUC Score for the training and testing set against each of these parameters.

- Number of Estimators (n_estimators)
- Max Depth
- Max Features

Number of Estimators

n_estimators represents the number of trees in the forest. As per Figure 16, the testing performance is not changing after 10 estimators. Therefore, the minimum number of estimators to be selected was set to 10.

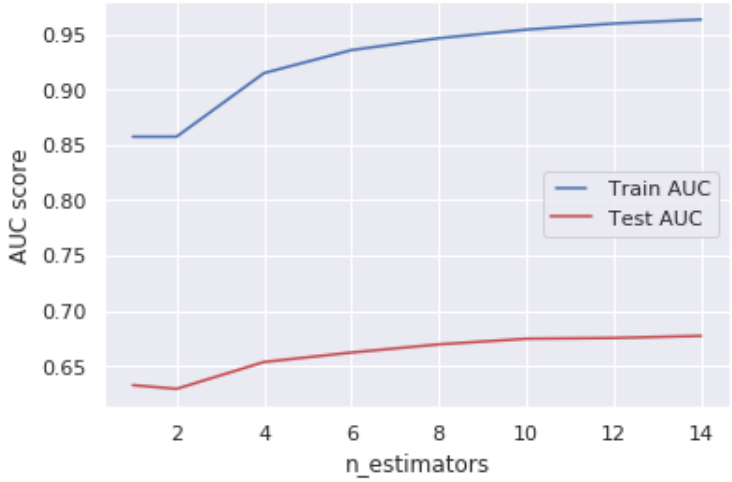


Figure 16 - Number of Estimators (Random Forest)

Max Depth

max_depth represents the depth of each tree in the forest. The deeper the tree, the more splits it has and it captures more information about the data. As per Figure 17, we see that the Model is overfitting for large depth values. A depth value of 11 would be ideal as the test performance is reducing after that.

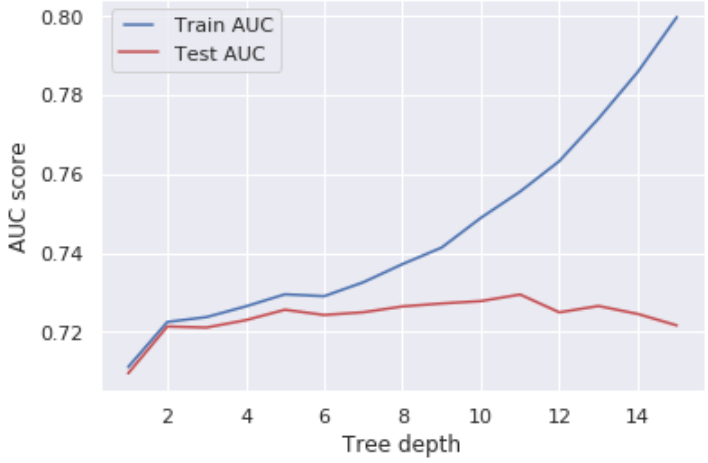


Figure 17 - Maximum Depth of tree (Random Forest)

Max Features

max_features represents the number of features to consider when looking for the best split. As per Figure 18, we see that the Model performance is declining after 8 features. Therefore, the maximum number of features was set to 8.

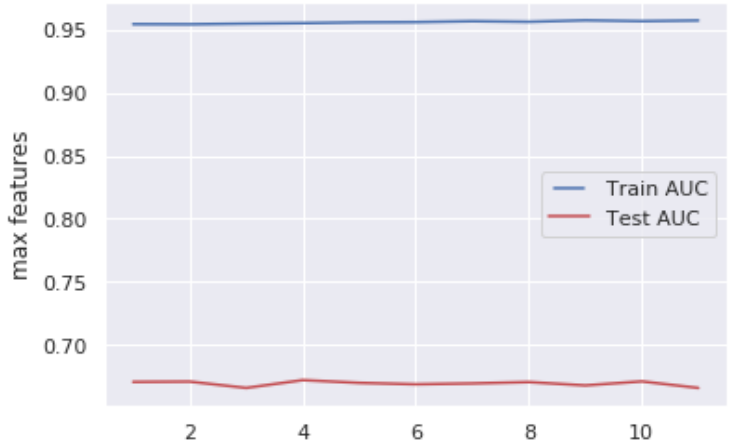


Figure 18 - Maximum Features (Random Forest)

A GridSearchCV was then performed to choose the best parameters for our model. The best parameters were set to **{'n_estimators': 10, 'max_features': 8,'max_depth': 11 ','criterion': gini, 'min_sample_split': ,9 'min_sample_leaf':10 }.**

The model's preformed improved by 6% as per the results below:

Accuracy: 72.6% , **Runtime:** 2.54 sec

	precision	recall	f1-score	support
0	0.71	0.78	0.74	9838
1	0.75	0.67	0.71	9687
accuracy			0.73	19525
macro avg	0.73	0.73	0.73	19525
weighted avg	0.73	0.73	0.73	19525

Random Forest ROC Curve for hyperparameter tuning:

Figure 19 displays the ROC Curve for Random Forest after the hyperparameters were tuned as above. This improved the Accuracy and AUC by 6% and performance by 0.19 sec. However it did not get better than Logistic Regression.

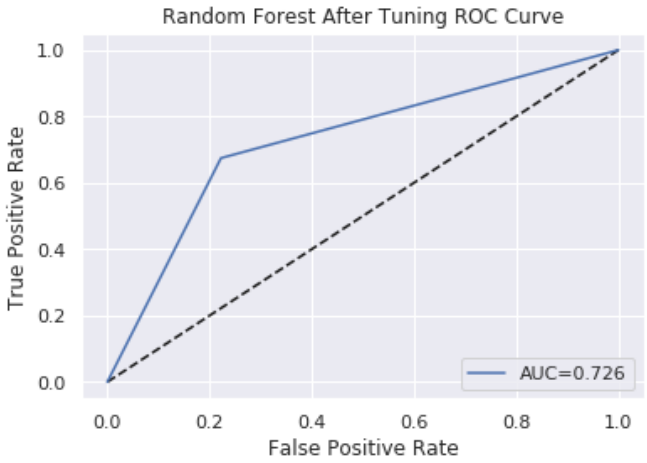


Figure 19 - Random Forest ROC Curve (Tuned Parameters)

Model Evaluation

For Evaluating the models the AUC was calculated and plotted for each model. Figure 20 shows the ROC curve for all the four classifiers above.

The AUC for Logistic Regression is the highest being 77.8% , therefore we can say that our Logistic Regression model is performing the best compared to the rest of the four models.

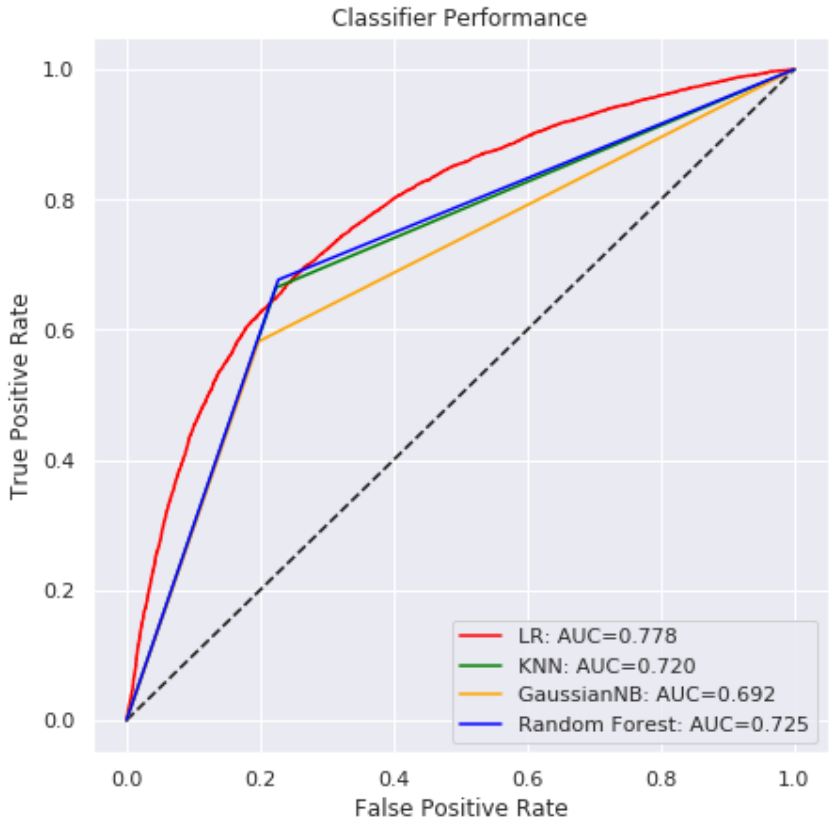


Figure 20 - Classifier Performance based on AUC

Discussion and Future Work

From the above the best AUC was received for Logistic Regression , however there is still room for improvement in the models. I was also unable to implement the SVM model as the performance was very slow. It worked only one time where the accuracy was 50%.

For Future tasks I would like to use my current models as the base models and try further parameters to improve the scores further. Another consideration would be to implement PCA, drop features based on Features Importance and try normalized data instead of standardized data and verify if the scores improve including the performance on SVM.

References

[1] “Heart Foundation,” [Online]. Available: https://www.heartfoundation.org.au/about-us/what-we-do/heart-disease-in-australia. [Accessed 1 06 2019].

[2] “Blood Pressure UK,” [Online]. Available: http://www.bloodpressureuk.org/microsites/u40/Home/facts/Whatisnormal. [Accessed 18 05 2019].