Cardiovascular Disease Prediction

Vinit Kishor Dhande (20202078)

Abstract

CVD is becoming the leading cause of mortality in the globe. There are, however, several strategies to lower your chance of acquiring these illnesses. If they do develop, there are several therapeutic options available. Often, there are no symptoms of the underlying disease of the blood vessels. A heart attack or stroke may be the first sign of underlying disease. In this project, we will see how different features affect the health and causes the cardiovascular disease. We are using the dataset from Kaggle which includes the objective features like age, height, weight, gender and some other features like systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake and activity of the patients. We have first checked if we have any false observations which may affect the prediction performance of the models. Then, we have removed such observations including some extreme values in the columns. We have visualized these variables to get the patterns from the data and to check how the target variables changes with respect to these variables. With Logistic regression, deep neural networks, support vector machines, random forests, and decision trees, we compared the outcomes of various modeling approaches. After hyperparameter tuning, we categorized the patients as having cardiovascular illness and obtained an optimal accuracy of 74.61 percent using the K-Nearest Neighbors. We also discovered that systolic and diastolic blood pressure have the greatest impact on the forecasts. As a result of putting these predictions into practice, we can diagnose individuals who are at risk of cardiovascular disease and provide appropriate counseling and medicines.

Introduction

The phrase "cardiovascular disease" refers to any illness that affects the heart or blood vessels. It's generally linked to fatty deposits in the arteries (atherosclerosis) and an elevated risk of blood clots. It has also been linked to artery damage in organs like the brain, heart, kidneys, and eyes. CVD is one of the primary causes of mortality and disability in the United Kingdom, however it may frequently be avoided by maintaining a healthy lifestyle. Up to 90% of cardiovascular disease is thought to be avoidable. CVD is prevented by modifying risk factors such as good diet, exercising, avoiding cigarette smoking, and limiting alcohol use. It is also useful to treat risk factors such as high blood pressure, blood lipids, and diabetes. Antibiotics can reduce the incidence of rheumatic heart disease in patients who have strep throat. The benefits of taking aspirin in persons who are otherwise healthy are debatable. Except in Africa, cardiovascular illnesses are the main cause of mortality. In 2015, 17.9 million people died from cardiovascular disease (CVD), up from 12.3 million (25.8%) in 1990. CVD deaths are more prevalent and have been growing in most of the developing world, whereas rates in much of the industrialized world have been declining since the 1970s. Coronary artery disease and stroke are responsible for 80% of CVD fatalities in men and 75% of CVD deaths in women. Most of the cardiovascular disease affects people in their fifties and sixties. In the United States, 11% of individuals between the ages of 20 and 40 have CVD, whereas 37% of people between the ages of 40 and 60, 71% of people between the ages of 60 and 80, and 85% of those over the age of 80 have CVD.

Age, sex, tobacco use, physical inactivity, excessive alcohol consumption, unhealthy diet, obesity, genetic predisposition and family history of cardiovascular disease, raised blood pressure (hypertension), raised blood sugar (diabetes mellitus), raised blood cholesterol (hyperlipidemia), undiagnosed celiac disease, psychosocial factors, poverty, and genetic predisposition and family history of cardiovascular disease are all risk factors for heart disease. While each risk factor's specific contribution differs by community or ethnic group, the total impact of these risk factors is relatively constant.

Dataset

We are using the open cardiovascular disease dataset available on Kaggle which includes the different variables affecting the health and cause the cardiovascular disease. We will see the different risk factors in the dataset below.

1. Age:

The most major risk factor for developing cardiovascular or heart illnesses is age, with the risk almost doubling every decade of life. In adolescence, fatty streaks in the coronary arteries might develop. 82 percent of those who die of coronary heart disease are 65 or older, according to estimates. At the same time, at the age of 55, the risk of stroke doubles every decade. Multiple explanations are proposed to explain why age increases the risk of cardiovascular/heart diseases. One of these has to do with cholesterol levels in the blood. The amount of blood total cholesterol rises with aging in most populations. This rise in men's testosterone levels peaks around the age of 45 to 50. The rise in women lasts until they are 60 to 65 years old.

2. Gender:

Men have a higher risk of heart disease than women before menopause. It has been claimed that once a woman has through menopause, her risk is equivalent to that of a male, however more recent evidence from the WHO and UN contradicts this. A female with diabetes is more likely than a guy with diabetes to get heart disease. Men in their forties and fifties are 2 to 5 times more likely than women to have coronary heart disease. According to a World Health Organization research, sex accounts for around 40% of the variance in sex ratios of coronary heart disease mortality.

3. Alcohol:

The link between alcohol intake and cardiovascular disease is complicated, and it may vary depending on how much alcohol is drunk. There is a direct link between excessive alcohol use and cardiovascular disease. Although there is evidence that correlations between moderate alcohol intake and protection against stroke are non-causal, drinking at low levels without bouts of severe drinking may be linked with a decreased risk of cardiovascular disease. The health hazards of consuming alcohol outweigh any possible advantages at the population level.

4. Smoke:

The most common type of smoked tobacco is cigarettes. Tobacco smoking poses health risks not just via direct use, but also through secondhand smoke exposure. Smoking is responsible for around 10% of

cardiovascular disease; nevertheless, those who quit smoking before the age of 30 have about the same risk of mortality as never smokers.

5. Cholesterol:

Cholesterol is a kind of lipid present in the bloodstream. High cholesterol can cause your blood arteries to constrict, increasing your chances of getting a blood clot. The reality is that high levels of poor cholesterol, low-density lipoprotein (LDL), are a significant cause of heart disease. LDL causes fatty deposits to form in your arteries, restricting or obstructing the flow of blood and oxygen that your heart need. Chest discomfort and a heart attack might result because of this.

6. Activity:

Because it improves the blood flow to the heart, regular physical exercise can assist alleviate angina symptoms. It also improves your exercise ability, which can contribute to a decrease in the frequency and severity of angina attacks. It may also help prevent the progression of your coronary heart disease.

7. Height:

The shorter you are, the greater your chance of developing coronary heart disease. That's one of the major findings of a recent study headed by the University of Leicester, which found that every 2.5 inches of height difference increased your risk of coronary heart disease by 13.5 percent.

8. Weight:

Being overweight has a significant link to an increased risk of coronary artery disease (CAD). Being overweight raises your likelihood of developing CAD risk factors. Obesity over at least two decades, according to longitudinal research, is likely to be an independent risk factor for coronary artery disease. A ten-kilogram increase in body weight raises the risk of coronary heart disease by 12%.

9. Blood pressure:

The Whitehall research used data on the 10-year mortality outcome of 18 403 male government workers aged 40-64 to examine systolic and diastolic blood pressures as indicators of death owing to coronary heart disease. After adjusting for age, men in the top quintile of systolic pressure (higher than 151 mm Hg) were shown to be 5% more likely to die from coronary heart disease than men in the top diastolic quintile (greater than 95 mm Hg). The findings showed that systolic levels should be used as a criteria for diagnostic and treatment decisions by doctors.

10. Glucose:

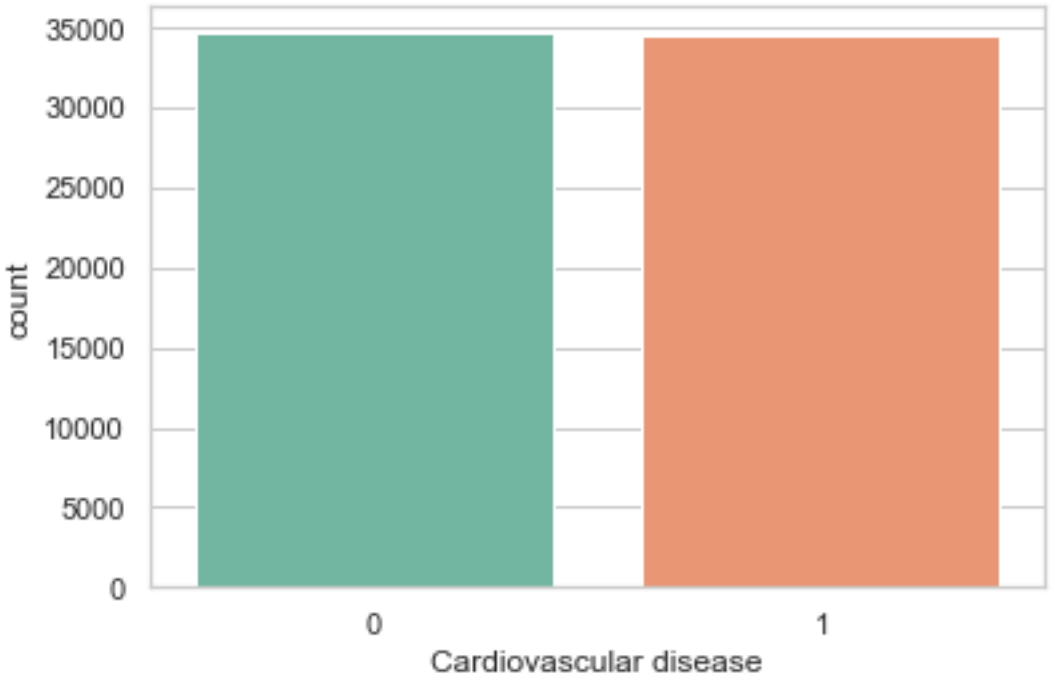
According to a research, there was a continuous positive relationship between normal fasting glucose and the risk of CVD down to 4.9 mmol/l. Overall, each 1 mmol/l reduction in normal fasting glucose was linked to a 21% (95 percent CI 18–24%) reduction in total stroke and a 23% (19–27%) reduction in total IHD risk. Across age ranges, the correlations were comparable in men and women.

11. Cardio:

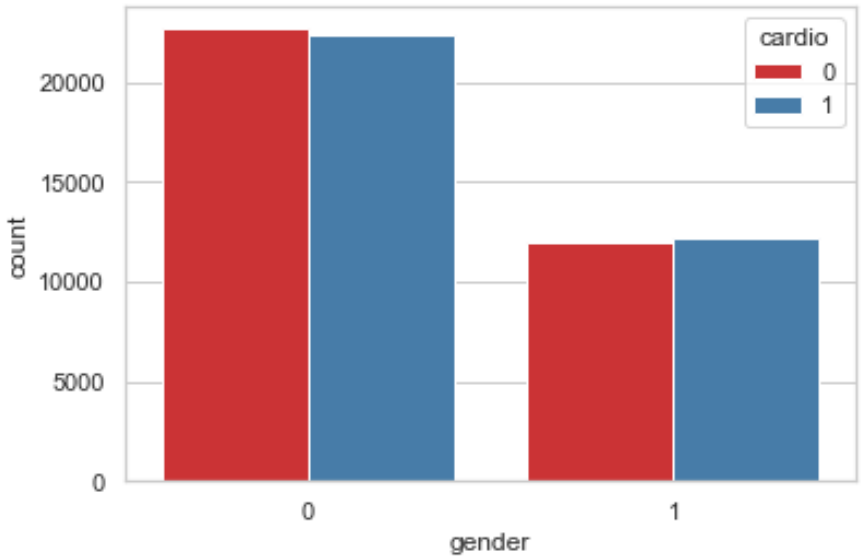
This is the target variable which tells us that the patient is diagnosed with CVD or not.

Exploratory Data Analysis and Data Preprocessing:

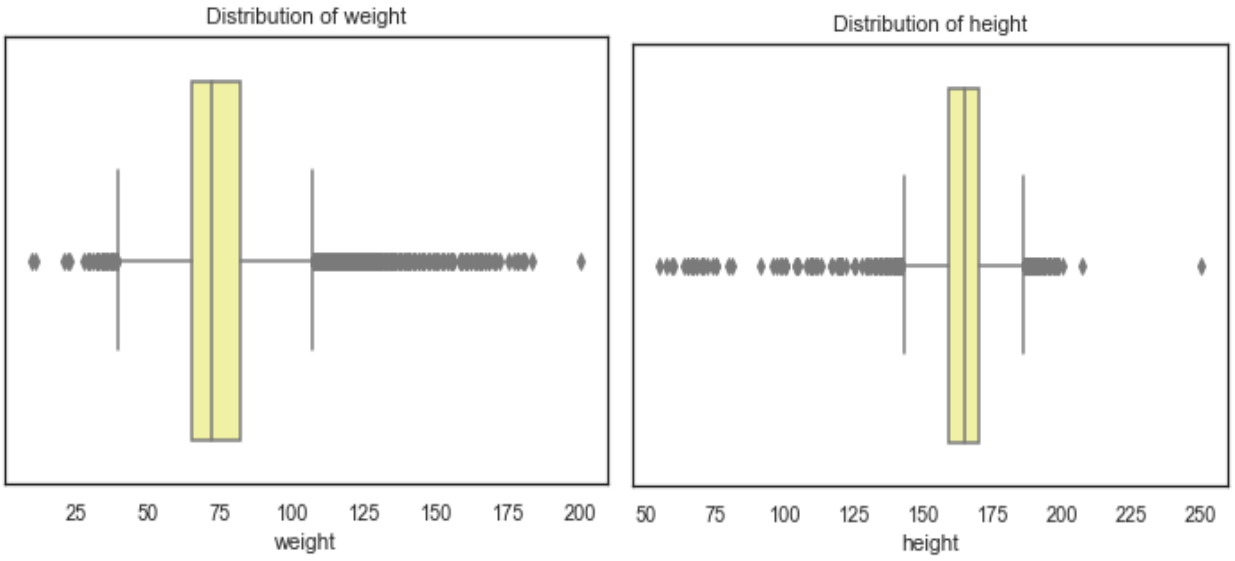
First, we will check if the data is balanced or not using the bar chart. The below chart illustrates that the data is balanced with respect to the target variable. There are around 68000 observations in the data, and we have approximately the same count of observations with respect to cardio column.



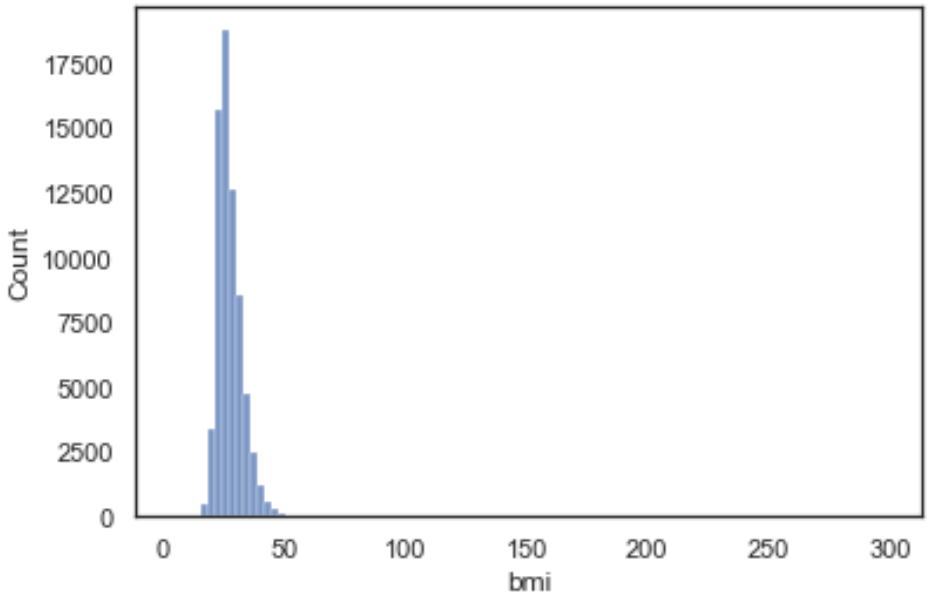
With 10000 individuals diagnosed with cardiovascular disease in each sex, we may conclude that the probability of men and women suffering from the condition are comparable. Overall, around 44% of patients observed are diagnosed with the CVD.



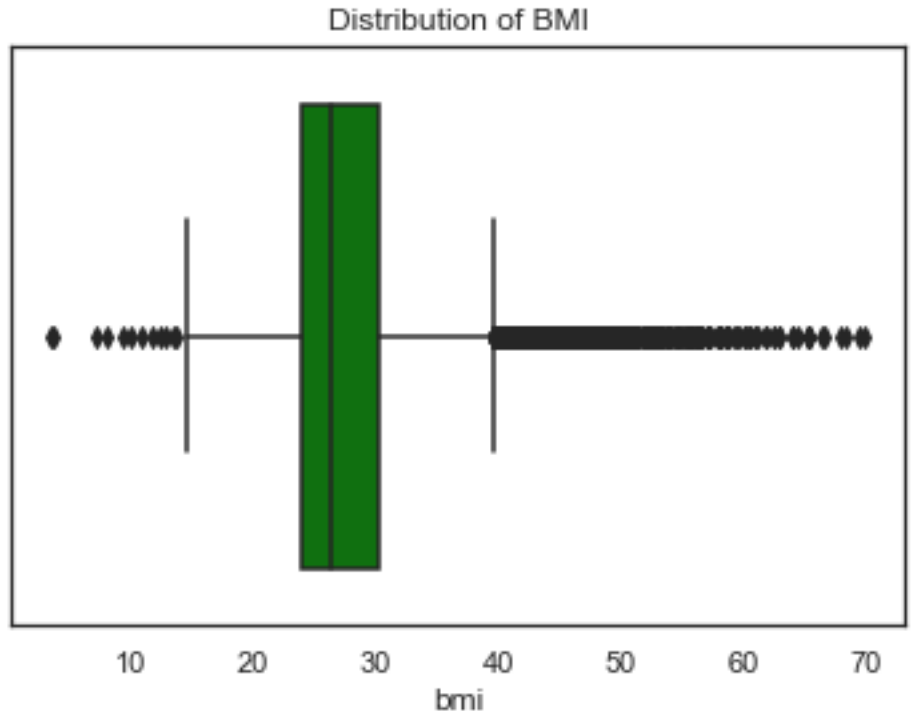
From the below plots of weight and height, we can depict that there are some extreme values in the data. In height column, we have an observation where height is more than 250cm which is not possible and may affect the prediction of the unknown data. Hence, we will remove such entries to avoid the conflicts in the modeling. But, in weight column there are extreme entries like 200kg but it is possible for a patient to have a weight more than 200kg.



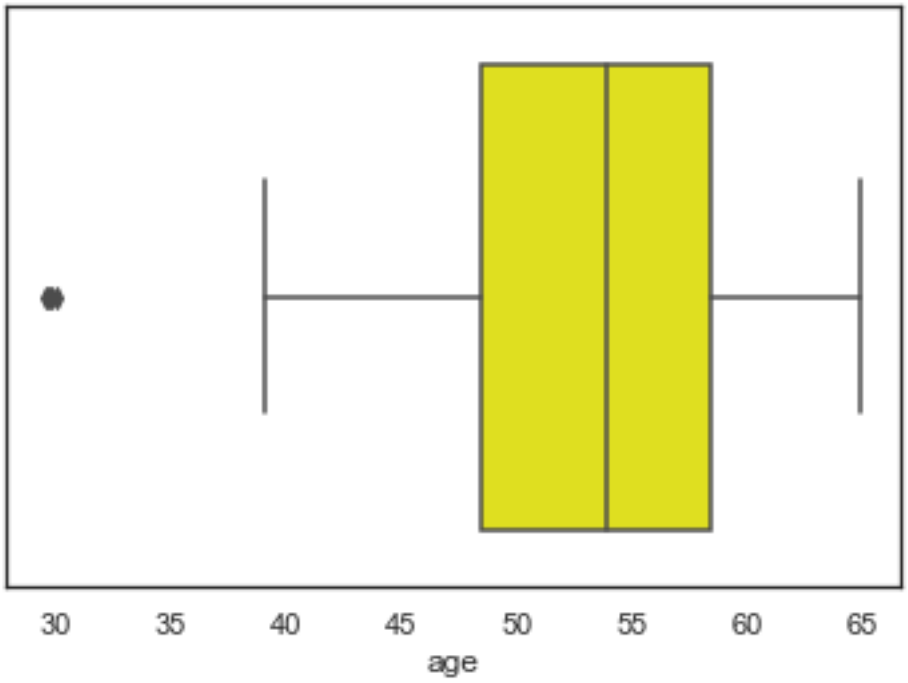
We have also created a new variable called BMI using the height and weight column. We can see that there are very large values ranging from 20-300 which is quite impossible to have. Hence, we have checked the data for these BMI values, and we can see that there are observations where height is 80cm but the weight is more than 150kg. Thus, it is important to remove such observations from the dataset. Thus, we have only considered the patients with less than 70 BMI index. We have successfully removed 35 entries with this approach.



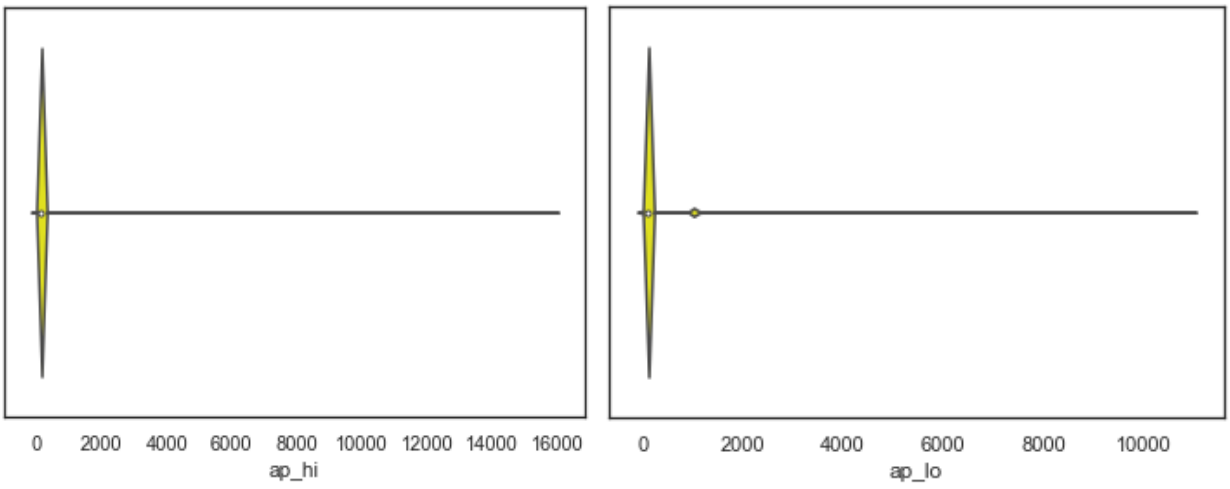
After removing entries:



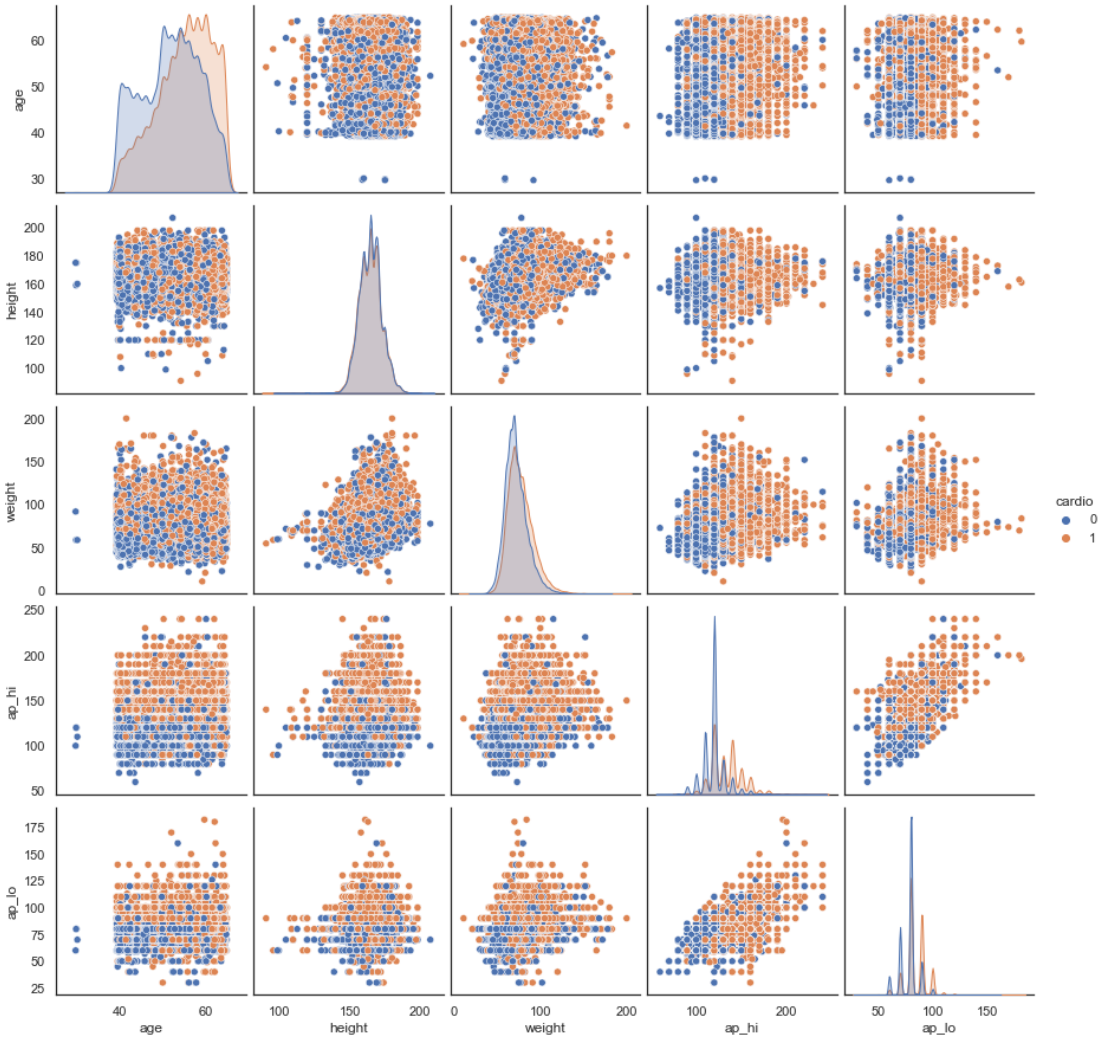
The age column distribution below suggest that there is significant evidence that there are no extreme values and we can go ahead with this dataset and continue the further data exploration.



From the below plots of systolic and diastolic blood pressure, we can depict that there are quite higher values which are clearly the sign of mis calculated or falsely entered while collecting the data.



Now, we will analyze the below pairs plot in which we can see that there is high probability that the patients with higher age will be more prone to the cardiovascular disease. Also, the people with relatively lower systolic and diastolic pressure will have less probability to have the CVD. But, it might be reverse if both the blood pressures are high then it might be risky.



Different Modeling Techniques Used:

1. Logistic Regression:

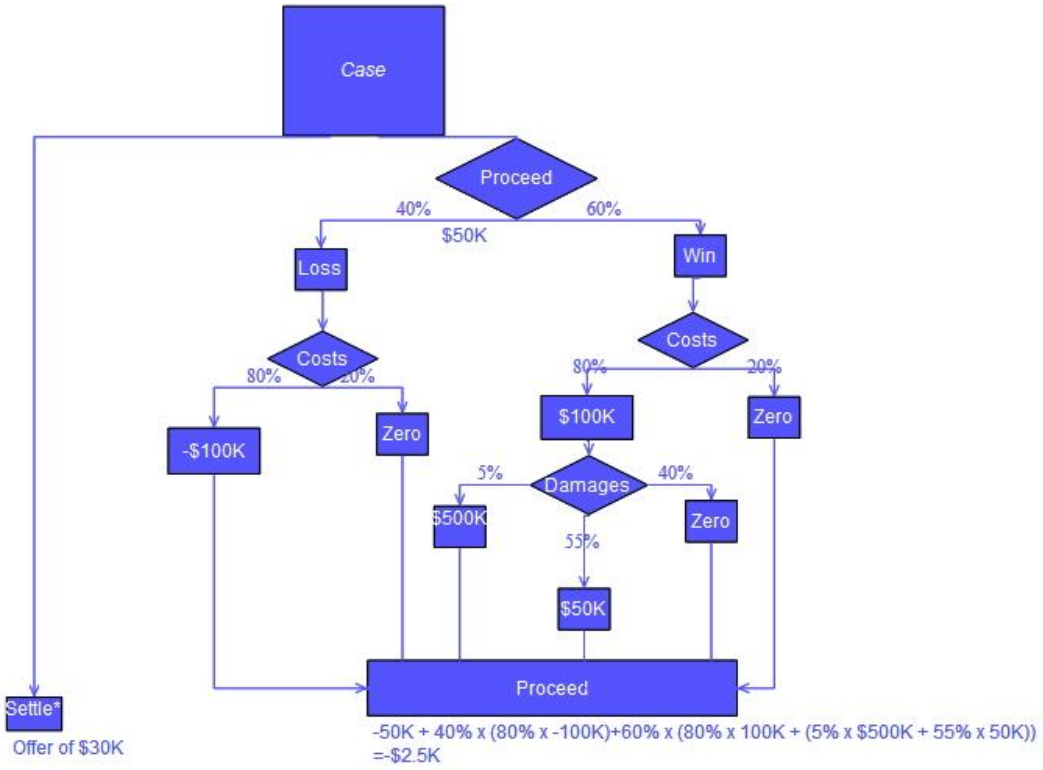
The logistic model (or logit model) is used in statistics to represent the likelihood of a specific class or event, such as pass/fail, win/lose, alive/dead, or healthy/sick, existing. This may be used to simulate a variety of occurrences, such as identifying whether a picture contains a cat, dog, lion, or other animal. Each identified object in the image would be assigned a probability ranging from 0 to 1, with a total of one.

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

2. Decision Tree:

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (for example, whether a coin flip will come up heads or tails), each branch reflects the test's conclusion,

and each leaf node represents a class label (decision taken after computing all attributes). The categorization criteria are represented by the routes from root to leaf. A decision tree and its closely related impact diagram are used as a visual and analytical decision support tool in decision analysis to determine the anticipated values (or expected utility) of competing options.



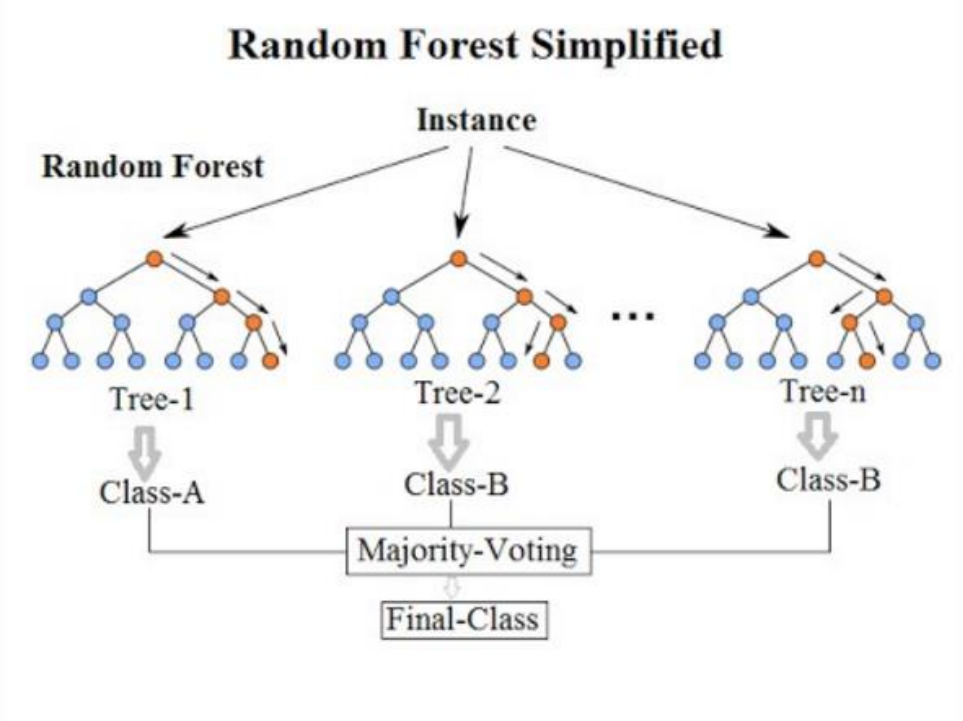
3. Support Vector Machine:

Support-vector machines are supervised learning models using learning algorithms that examine data for classification and regression analysis in machine learning. An SVM training algorithm creates a model that assigns new examples to one of two categories, making it a non-probabilistic binary linear classifier, given a collection of training examples, each marked as belonging to one of two categories. SVM translates training examples to points in space to widen the distance between the two categories as much as possible. New instances are then mapped into the same space and classified according to which side of the gap they land on. SVMs may do non-linear classification as well as linear classification by implicitly translating their inputs into high-dimensional feature spaces, which is known as the kernel trick.

4. Random Forest:

Random forests, also known as random decision forests, are an ensemble learning approach for classification, regression, and other problems that works by training a large number of decision trees. For classification tasks, the random forest's output is the class chosen by the majority of trees. The mean or average forecast of the individual trees is returned for regression tasks. Random decision forests address the problem of decision trees overfitting their training set. Random forests outperform decision trees in

most cases, but they are less accurate than gradient enhanced trees. Data features, on the other hand, might have an impact on their performance.



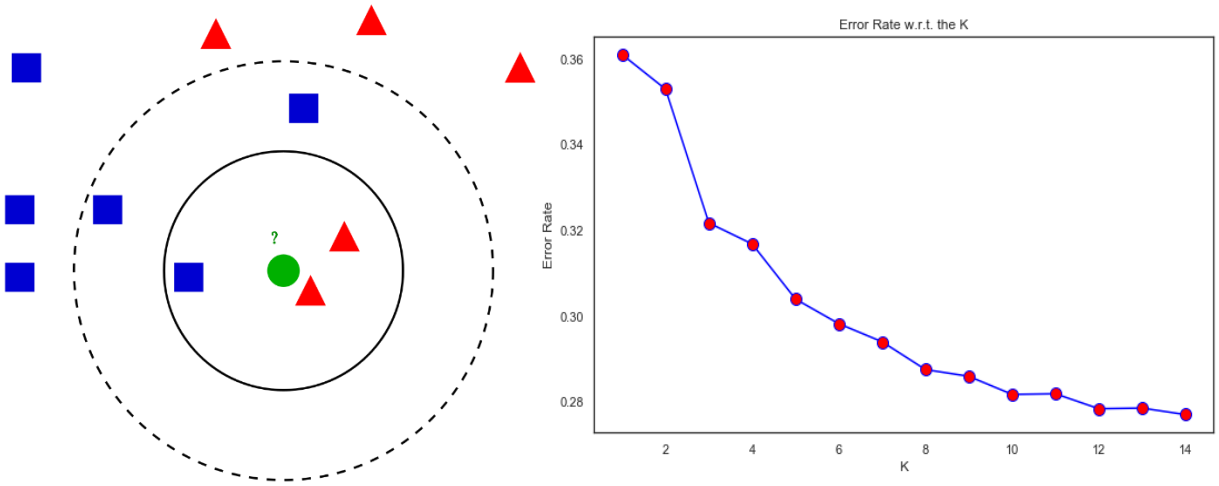
5. K-Nearest Neighbors Classification:

The result of k-NN classification is a class membership. When a majority of its neighbors classify an item, it is placed in that class. In the case of k = 1, the item is allocated to a class based on its nearest neighbor. It is a classification method in which the function is only estimated locally and full computation is delayed until the function evaluation is complete (K-NN). Since this method depends on distance for classification, normalizing the training data can enhance its performance substantially if the features represent distinct physical units or arrive in wildly different sizes.

Optimizing the model by hyperparameter tuning:

As a result of the "elbow" technique of determining the ideal number of clusters for K-means clustering, the K-Elbow Visualizer implements this. There are many unsupervised machine learning algorithms available, however K-means is a simple method that organizes data into a predetermined number of clusters (k) However, this technique is relatively naïve because it relies on the user to define the cluster size in advance. Elbow conducts k-means clustering on the dataset over a range of k value and then computes an average score for all clusters for each value of the k. This score is calculated by default as the sum of the square distances between each point and its allocated center.

We have implemented the elbow method to the given dataset, and we can see that there is no significant evidence about the optimum value of k, but we can get the maximum value of k=10 as optimum. Also, we can see that the graph bends at k=10.



6. Artificial Neural Network:

"An artificial neural network (ANN) is composed of artificial neurons, which loosely resemble the brain's neurons." Synapses in biological brains can send signals to other neurons through each link. Signals are received by artificial neurons, which then process them to send signals to other neurons that it is linked to. Every neuron's output is determined by some non-linear function of its inputs, and every connection's "signal" is a real number. As a result, edges are used to describe the relationships. This means that as you learn more, you'll be able to change the weight of neurons and edges. The weight affects the intensity of the signal at a link by increasing or decreasing it. As a result, neuronal signals may only be delivered when the aggregate signal reaches the threshold. Neurons are often grouped into layers. The inputs to various levels may be transformed differently. When a signal reaches the output layer, it may have traversed many layers before reaching the first one (the input layer).

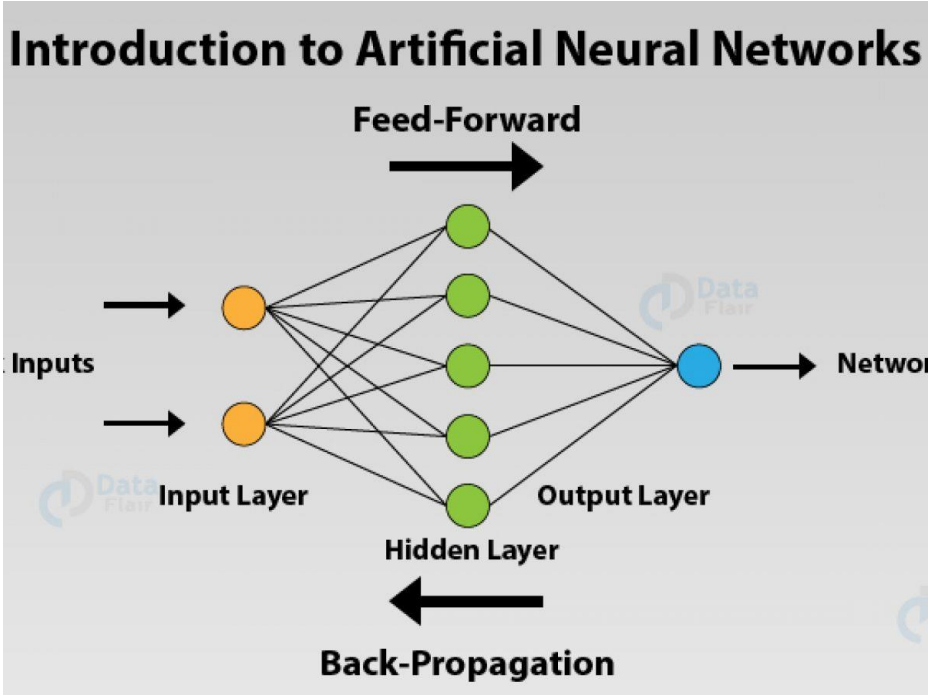
First Model:

In this project, we have implemented two models. In the first model, we can see that there are 13 input features. We have a four-layer network including the input layer. The activation parameter is helpful in applying the element-wise activation function in a dense layer and here we are using "**relu**" activation technique except in the fourth one we are using "**sigmoid**" activation. Optimizers are algorithms or methods used to change the attributes of the neural network such as weights and learning rate to reduce the losses. Here we are using the **"adam"** optimizer. To predict class 1, cross-entropy calculates an average difference between the actual and projected probability distributions. Cross-entropy is zero when the score has been reduced. In this model we are using loss function as "binary_crossentropy". The "monitor" allows you to specify the performance measure to monitor in order to end training. Adding a delay to the trigger in terms of epochs on which we would want to observe no improvement can accommodate for this. The "patience" argument can be used to achieve this.

Second Model:

As for the second model, we've introduced random uniform initializer for kernel and bias, which produces uniformly distributed tensors. Weight and bias initialization for each layer can be set via kernel_initializer and bias_initializer keyword arguments respectively within dense layer. A variable's initialization method, contained in the Initializer object, can be pre-specified without knowing the shape or datatype of the variable. Also, we have added the fourth hidden layer in the model with the above settings. We have added a dropout which is a Simple Way to Prevent Neural Networks from Overfitting. A decent starting point is a dropout value of 20 - 50 percent of neurons. A probability that is too low has little impact, whereas a value that is too high results in the network under-learning.

From the final results of both the model, we have successfully improved the performance of the ANN model by introducing the initializers and dropout.



Comparison and Evaluation of Models:

There are four ways to check if the predictions are right or wrong:

TN / True Negative: the case was negative and predicted negative

TP / True Positive: the case was positive and predicted positive

FN / False Negative: the case was positive but predicted negative

FP / False Positive: the case was negative but predicted positive

Precision — What percent of your predictions were correct?

Precision = TP/(TP + FP)

Recall — What percent of the positive cases did you catch?

Recall = TP/(TP+FN)

F1 score — What percent of positive predictions were correct?

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

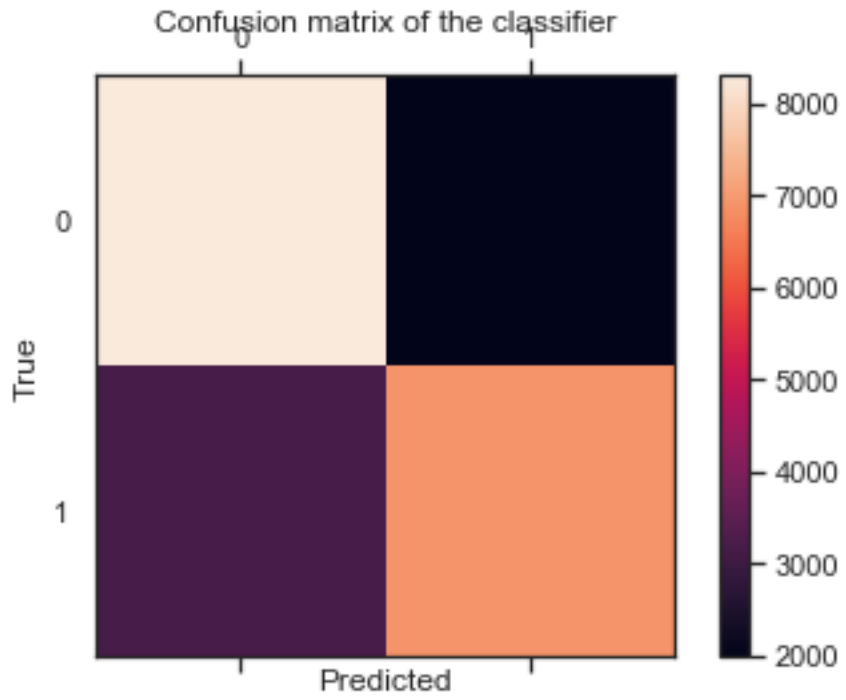
If we compare the below models with accuracy, then we can say that we have improved the performance of the K-Nearest Neighbors after tuning and getting the accuracy of 74.61%. Thought, we can see that the other models are also giving the results around 70-73% . We have also got the 73.17% accuracy after the improvement in the first ANN model.

Then we will evaluate the performance of the KNN classifier. With this model, we are able to recall the 81% of the positive results successfully. Also, we are predicting the new data precisely around 75% for both positive and negative target.

	Accuracy
K-nearest Neighbors After Tuning	74.610000
ANN-2	73.170000
Support Vector Machines	73.060000
ANN-1	73.060000
Logistic Regression	72.540000
Random Forest	71.040000
K-nearest Neighbors	69.590000
Decision Tree	62.670000

	precision	recall	f1-score	support
0	0.72	0.81	0.76	10291
1	0.78	0.69	0.73	10097
accuracy			0.75	20388
macro avg	0.75	0.75	0.74	20388
weighted avg	0.75	0.75	0.75	20388

```
[[8295 1996]
 [3180 6917]]
```



Conclusion:

Finally, we had a great time working on this project throughout all phases, and we consider it to be one of the most fascinating projects we've ever worked on for a number of reasons. As a consequence of the modules, we were able to use the bulk of our learning outcomes from the whole python/machine learning sections, and we developed skills and clarified confusions. We used the features provided to perform the visuals and investigation. In addition, we were able to successfully eliminate the data's questionable findings. After hyperparameter tweaking, we categorized the patients as having cardiovascular illness and obtained an optimal accuracy of 74.61 percent using the K-Nearest Neighbors. With Logistic regression, deep neural networks, support vector machines, random forests, and decision trees, we compared the outcomes of various modeling approaches. We also discovered that systolic and diastolic blood pressure have the greatest impact on the forecasts. As a result of putting these predictions into practice, we can diagnose individuals who are at risk of cardiovascular disease and provide appropriate counseling and medicines.

Reference:

1. Wikipedia: https://en.wikipedia.org/wiki/Cardiovascular_disease#Genetics
2. Cardiovascular diseases factsheet: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
3. KNN Elbow method: https://www.scikit-yb.org/en/latest/api/cluster/elbow.html
4. Binary loss function: https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/
5. Dense layers in ANN: https://machinelearningknowledge.ai/keras-dense-layer-explained-for-beginners/#1_Units
6. Dropout regularization in Deep learning: https://machinelearningmastery.com/dropout-regularization-deep-learning-models-keras/
7. Blood Glucose and Risk of Cardiovascular Disease in the Asia Pacific Region: https://care.diabetesjournals.org/content/27/12/2836
8. What to know about cardiovascular disease: https://www.medicalnewstoday.com/articles/257484
9. CARDIOVASCULAR DISEASE: https://irishheart.ie/heart-and-stroke-conditions-a-z/cardiovascular-disease/
10. Heart Disease: https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118