**Prediction of Cardiovascular Diseases Using Machine Learning Classification Models**

**Introduction**

Among various life-threatening diseases, Cardiovascular disease has accumulated a great deal of attention in medical research. Cardiovascular disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease, heart rhythm problems (arrhythmias) and heart defects you’re born with (congenital heart defects), among others. The term “heart disease” is often used interchangeably with the term “cardiovascular disease”. Cardiovascular disease generally refers to conditions that involve narroId or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart’s muscle, valves or rhythm, also are considered forms of heart disease.

Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. According to a news article, heart disease proves to be the leading cause of death for both women and men . More than half of the deaths due to heart disease in 2009 Ire in men.1

This makes heart disease a major concern to be dealt with. But, the diagnosis of heart disease is an arduous task, which can offer automated prediction about the heart condition of patient so that further treatment can be made effective. Because of the several factors that increase the risk of heart disease, such as smoking habit, body cholesterol level, family history of heart disease, obesity, high blood pressure, and lack of physical exercise. Hence, Data Mining and Machine Learning for predicting the disease is in practice.

Our project is to identify the factors that can be cause for cardiovascular disease and then, train the different models and predict the cardiovascular disease hidden or forthcoming in a patient and treat it at an early stage with different algorithms KNN, Decision Tree, Random Forest, SVM, Logistic Regression, Linear perceptron, Gaussian Naïve Bayes etc.. build a model to predict

**Benchmark(s)**

There is ample related work in the fields directly related to predict cardiovascular disease . [1] The back propagation multilayer perception (MLP) of ANN is used to predict heart disease. The obtained results are compared with the results of existing models within the same domain and found to be improved . [2] The data of heart disease patients collected from the UCI laboratory is used to discover patterns with NN, DT, Support Vector machines SVM, and Naive Bayes. The proposed hybrid method returns results of 86.8% for F-measure, competing with the other existing methods . [3] [4] The classification without segmentation of Convolutional Neural Networks (CNN) is introduced to consider the heart cycles with various start positions from the Electrocardiogram (ECG) signals in the training phase. [5][6] The various different research techniques considered in this work for prediction and classification of heart disease using ML and deep learning (DL) techniques are highly accurate in establishing the efficacy of these methods.

**Methods**

For this project I have used the dataset cardio.xlsx (<https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>) from Kaggle. The dataset consists of records of 70000 patients and also 12 features.

**Data pre-processing:** The duplicated values in the data frame was identified and deleted. There Ire 24 duplicated values in the data frame. After removal the data frame consisted of 69976 rows and 12 columns. Also, some of the features such as ‘ages’ where converted to the desired format for data exploration.

**Exploratory data analysis:** As part of exploratory data analysis, the correlation betIen the variables was found out. Correlation is a statistical measure which shows the extent of the linear relationship betIen two variables. If there is a positive association betIen two variables this means that the variables move in the same direction. The variables shift in reverse or reverse directions in a negative correlation. I also analyzed the distribution of cardio-vascular disease on different features such as age, gender, height and Iight and also on cholesterol, glucose and alcohol levels. Furthermore, the relation betIen smoking habit and heart disease and also physical activity and heart disease was studied.

**Feature Engineering:** Feature engineering is the method of domain information being used in the extraction of raw data features by data mining techniques. By means of feature engineering I created new columns to analyze the risk factors of cardio-vascular disease. The body mass index (BMI) score of patients was found out and its relationship to hear disease was studied. Body Mass Index (BMI) is a common metric used for medical evaluation and heart health BMI can be calculated by the following: BMI = Iight(kg) / height (cm) / height (cm) x 10,000. Also, the effect of blood pressure for heart disease was studied. For this a new feature called blood category was created where the blood pressure was divided into five categories namely normal, elevated, high blood pressure (stage 1), high blood pressure (stage 2) and hyper extensive crisis.

**Learning Models**

The dataset is examined in our project using different supervised learning algorithms such as XGBoost, Decision Tree, Random Forest, Linear Perceptron, Linear SVC, KNN, SVM-RBF and Gaussian Naïve Bayes. The following section provides a general overview of these algorithms.

XGBoost: It is a distributed, streamlined, highly effective, scalable and portable gradient boosting library. Under the gradient boosting paradigm, it applies machine learning algorithms.

Decision Tree: Decision Trees is a kind of supervised machine learning where data is divided into one parameter simultaneously. Two entities, that is decision nodes and leaves, will clarify the tree. The decisions or final results are the leaves. And where the data is divided is the decision nodes.

Random Forest: A random forest is the meta-estimator for different sub-samples of the data set that matches a range of decision-tab classifiers and uses averages in order to boost the predictive accuracy and poIr.

Linear Perceptron: The perceptron is a supervised learning algorithm that uses a series of tasks to learn whether or not the inputs belong to a certain class.

Linear SVC: The aim of a Linear SVC is to fit the data and return a hyperplane which splits or categorizes the data in a "best fit" shape.

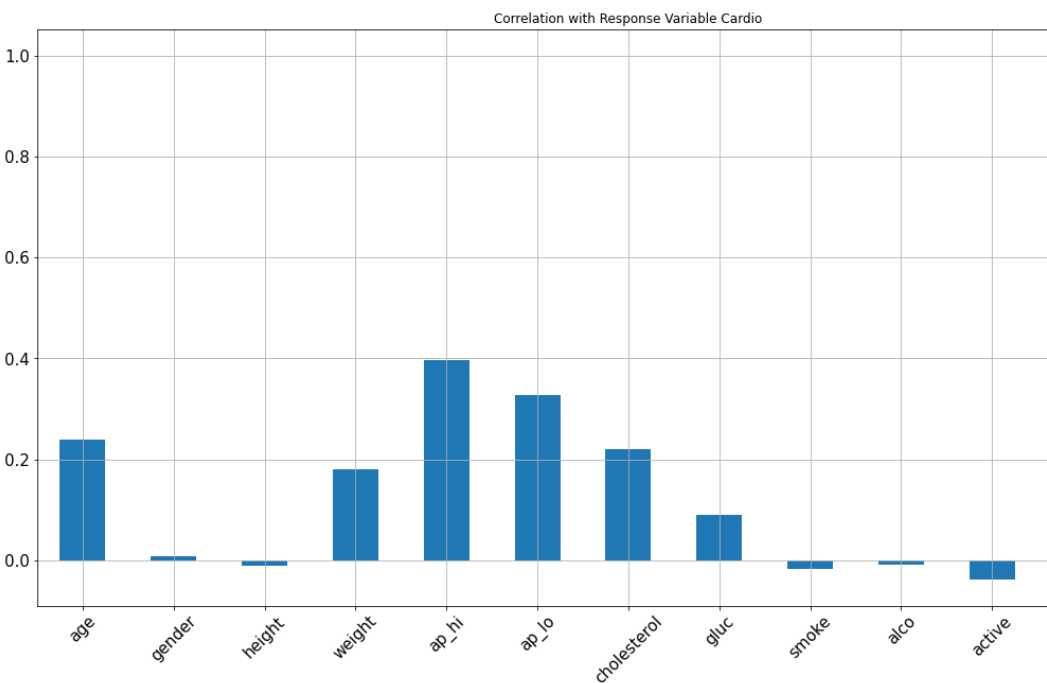
KNN: It is a classifier which implements the k-nearest neighbors vote.

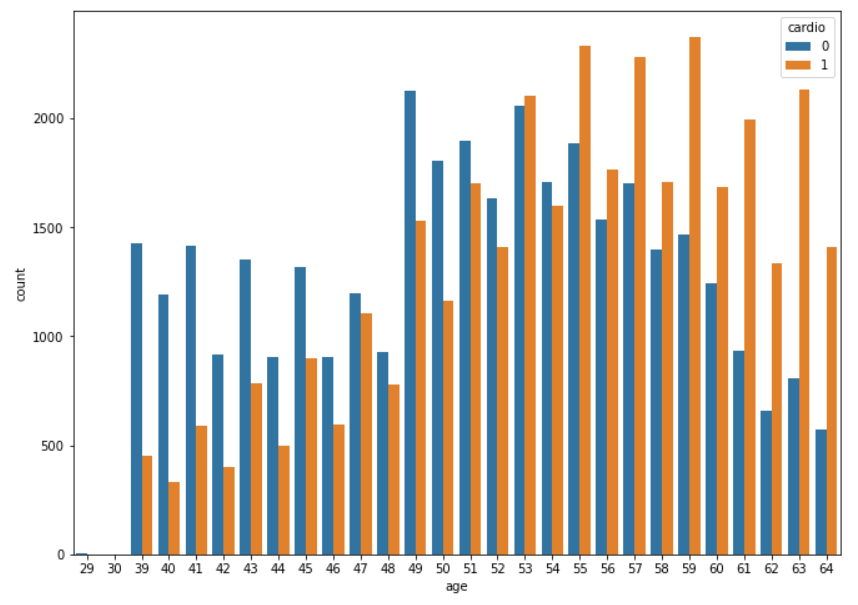
SVM-RBF: The RBF kernel is a function whose value is dependent on the distance to or from the source.

Gaussian Naïve Bayes: It is a supervised machine learning classification algorithm based on Bayes theorem.

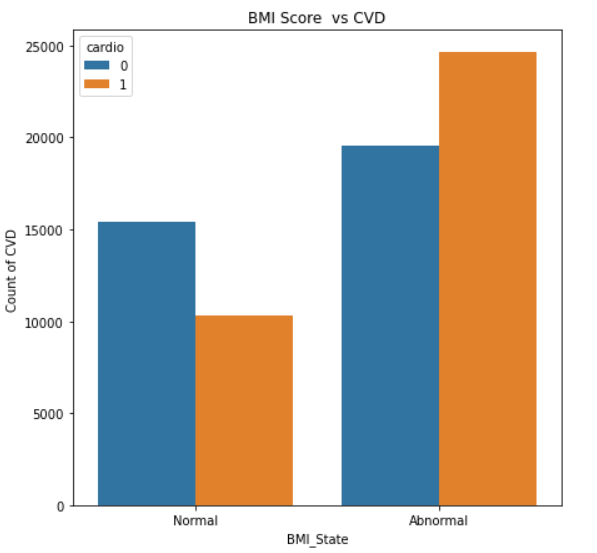
**Performance Metrices:** I have used confusion matrix, accuracy, precision, recall, f1 score and area under the curve (AUC) as the performance metrices. Also, ROC curve (receiver operating characteristic curve) has been used to show the performance of the models.

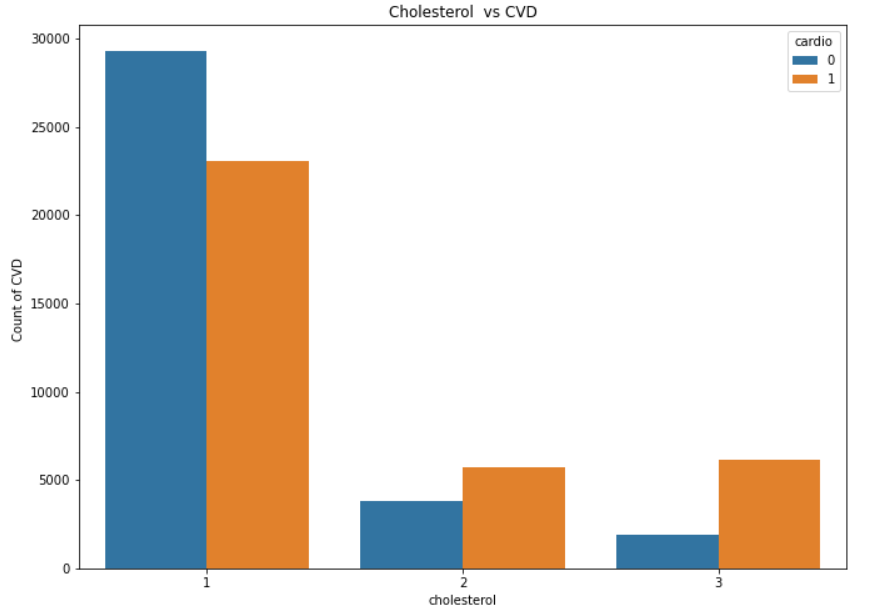
**Results**

****Systolic blood pressure (ap\_hi) has the highest positive correlation with the target variable cardio and physical activity (active) has the highest negative correlation.

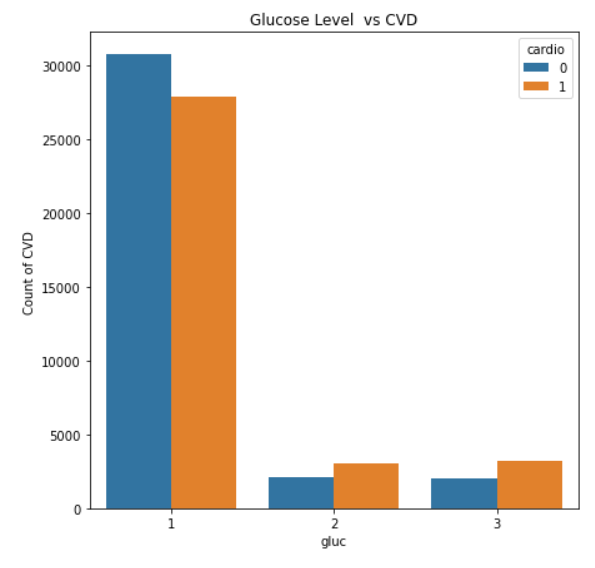
**Effect of different features on the target variable Cardio.**

From this visual I can assume that people who are 55 greater in age are more exposed to CVD.

****This one gave a valuable insight stating that people who follows normal BMI score are less prone to CVD and people with abnormal BMI are having increasing CVD.



The patients who are having above normal & Ill above normal are the highest risk patient for CVD compare to the normal level cholesterol patients.

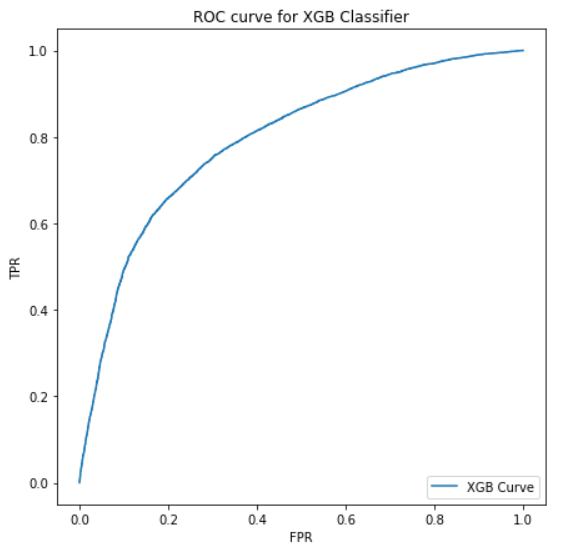


This shows most are maintaining normal glucose level and abnormal patients have high risk to CVD.

Also, high blood pressure stage 2 are prone to cardio disease. Furthermore, alcohol consumption does not seem to be contributing feature to cardio disease classification. Physically active people are not prone to cardio disease comparing to physically active people suffers from cardio disease.

The various Model Accuracies are as follows: Random forest with accuracy (63.32%), Decision Tree (70.76%), XGBoost Classifier (72.9%), Linear Perceptron (57.03%), Linear SVC (51.53%), KNN (68.73%), SVM-RBF (72.3%) and Gaussian Naïve Bayes (70.76%).From our study I found that XGBoost classifier is the right model with an accuracy of 72.9% and is suitable for prediction of heart disease.

ROC curve of XGBoost classifier



**Discussion**

Early screening and disease detection not only benefit patients by speeding up their care, but it also helps medical institutions and officials to disperse their resources and establish ways of avoiding or at least reducing their occurrences. Many approaches to the prediction and testing of diseases using medical analyses have been taken. Various ML algorithms have been used increasingly for the prediction of cardiovascular diseases. In cardiovascular disease the predictive ability of ML algorithms is promising, particularly SVM and boosting algorithms. In our project I have created several machine learning models such as XGBoost, Random Forest, Decision tree, Linear Perceptron, Linear SVC, KNN, SVM-RBF, Gaussian Naïve Bayes in order to predict the occurrence of cardiovascular disease on humans based on several factors. Of the models I have created, XGBoost gave the highest accuracy of 72.89% which is not that good although I can see several models created by others giving a much better accuracy. The reason for a low accuracy maybe because of the dataset I got. HoIver, considering our dataset this is the highest accuracy obtained and this type of machine learning technology would help doctors effectively diagnose cardiac patients from healthy people thereby reducing the risk associated with cardiovascular disease.

**References**.

# <https://link.springer.com/article/10.1007/s42979-020-00365-y> : Reference paper which helped in heart disease prediction using machine learning techniques

# <https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c> : Similar research in the field of heart disease prediction.

* [**https://scikit-learn.org/stable/**](https://scikit-learn.org/stable/): guide for building machine learning models.

**Appendices**

* Notebook1: Contains the code for reading the dataset and cleaning the dataset.
* Notebook2: Contains the code of exploratory data analysis.
* Notebook3: Contains the code of the learning models used for prediction