**Artificial Intelligence Project Report**

**Cardiovascular Heart Disease Prediction**

******

**Submitted by : Ali Qureshi 20k-0149**

**Ahmed memon 19k-0214**

**Abdul Rehman 20k-0239**

**Submitted to: Ms. Sania Urooj**

**Mr Fahad Sherwani**

**Date: 3 May 2023**

**Index**

Abstract…………………………………………………………………………………………3

Introduction……………………………………………………………………………………3-4

Methodologies……………. ……………………………………………………………………4

Decision Tree Classifier…………………..…………………………………………………….4-5

Gradient Boosting classifier…………………………………………………………………….6-7

Logistic Regression…………………………………………………………………………7-8

Max Voting Classifier…………………………………………………………………………9-10

Conclusion…………………………………………………………………………………….. 10

**Abstract**

Cardiovascular disease (CVD) makes our heart and blood vessels dysfunctional and often leads to death or physical paralysis. Therefore, early and automatic detection of CVD can save many human lives. Multiple investigations have been carried out to achieve this objective, but there is still room for improvement in performance and reliability. This study is yet another step in this direction. The classification algorithms used are Logistic Regression, Decision Tree Classifier and Gradient Boosting algorithm, and Maximum Voting Classifier have been employed for CVD detection using publicly available kaggle dataset. The performances of the models are optimally increased by removing outliers and attributes having null values. We have used data visualization to understand the data. Experimental-based results demonstrate that a higher accuracy in detection of CVD is 71.47% and an area-under-the-curve value of 71.5% are obtained using the Max Voting Classifier model, higher than the Logistic Regression, Decision Tree Classifier and Gradient Boosting algorithm model. Therefore, the proposed Maximum Voting Classifier model was recommended for automatic CVD detection. The proposed methodology can also be employed in detecting other diseases. In addition, the performance of the proposed model can be assessed via other standard data sets.

**Introduction**

Health is a crucial part of everyone’s life. However, owing to multiple reasons like unhealthy lifestyles, work stress, psychological strain, and external factors such as pollution, hazardous work environment, and lack of proper health services, millions of people worldwide fall prey to chronic ailments like cardiovascular diseases (CVD), which affect both the heart and blood vessels, resulting in death or disability. In recent years, it was reported that the majority of human deaths were due to CVD . The associated conditions are hypertension, thromboembolism, hyperlipidemia, and coronary heart disease, which culminate in heart failure. Hypertension is the primary cause of CVD. In 2012, 7.4 million people were reported to have died from coronary heart disease, while ackg6.7 million people died from stroke. The World health Organization estimates that nearly 17 million people die every year from CVDs, which accounts for approximately 31% of global deaths. Early diagnosis of CVD can potentially cure patients and save innumerable lives. Diagnosis and treatment of patients at early stages by cardiologists remain a challenge. Every traditional CVD risk-assessment model implicitly assumes each risk factor related to CVD outcome in a linear fashion. Such models have a tendency to oversimplify complex relationships, including several risk factors with non-linear interactions. Multiple risk factors should be properly incorporated, and more correlated nuances between the risk factors and outcomes should be determined. To date, no large-scale study has used routine clinical data and machine learning (ML) in prognostic CVD assessment. The goal of this study is to determine if ML can enhance cardiovascular risk prediction accuracy in population primary care at large and find out which ML algorithm result had fairly high brevity. In recent years, multiple ML-based CVD detection models have been proposed. A review of previous studies is presented to identify the research problem and objective of each study. ML helps a cardiologist to predict diseases at an early stage and treat the patient accordingly. There are many ML techniques such as support vector machines, artificial neural networks, decision trees, and K-Nearest Neighbour (K-NN), each with its strengths and weaknesses. These methods have been applied in broader areas like in predicting liver, human heart (echocardiogram signals), and skin diseases. Results of each technique differ owing to several constraints. Observations from related studies reveal that there is further scope for the development of automated CVD detection using other ML models that provide improved performance. This study contains an in-depth statistical analysis of input data sets to understand the effects of data range on the CVD predictions. It includes a correlation study of categorical and continuous features of patients. In addition, data visualization and scatter plots for pairs of important features were obtained to understand the significance of the correlation between important features. These are discussed and analyzed in the results section.

**Methodologies**

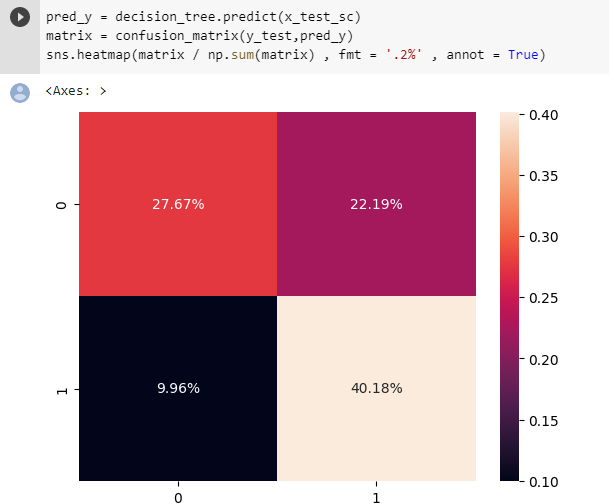
We have used libraries such as Pandas, numpy, matplotlib, scipy, seaborn, sklearn, and plotly for graphs. We have extracted our database from Kaggle. In which it has 70000 rows and 14 columns. The goal of this study was to determine whether or not a patient would develop CVD if a set of clinical information is available. The confusion matrix of each technique was obtained, and out of 70000 occurrences in the data set, 75% were used to train the ML models. To test the trained models, 25% of dataset were fed to know the class. This study intends to predict the likelihood of developing CVD via a computerized prediction route that can be useful to health professionals. The materials required for CVD detection are the test data of patients from publicly available standard CVD data from the kaggle repository. The classification algorithms used are Logistic Regression, Decision Tree Classifier and Gradient Boosting algorithm. Generally, the method comprises training of the proposed model via respective learning algorithms using relevant input test data of patients and then validating these models based on test data of patients. Finally, performance measurements are evaluated and compared.

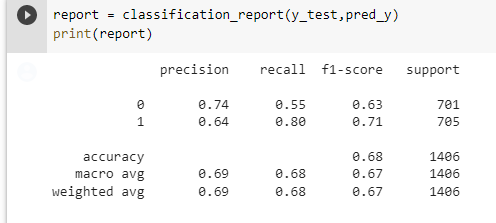
**Decision Tree Classifier**

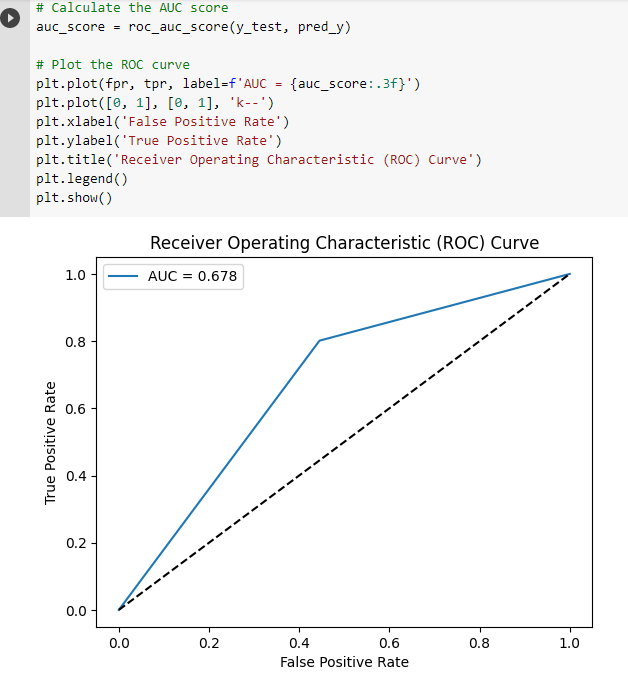
Decision tree is a flow chart that is shaped like a tree structure where each internal node states the test of an attribute, each branch states the output of the test and the leaf node states the classes or class distributions. The top node is called the root node or the root node. A root node will have several exit edges but no entry edge, internal nodes will have one entry edge and several exit edges, while a leaf node will only have one entry edge without having an exit edge. Decision tree is used to classify a sample of data that is not yet known class into classes that already exist. The data testing path is first through the root node and the last is through the leaf node which will conclude the class prediction for the data.

We have used Kfold to break data into into smaller samples and have extract out insights of data; for e.g. we have used the prediction model to predict CVD, we have used confusion matrix, and classification report to display results and have the accuracy of 93.97 % for training and 67.85%

We have used AUC curve for visualization of results. In which in x-axis; it is False positive rate and in y-axis, it is True Positive Rate with AUC= 0.678.



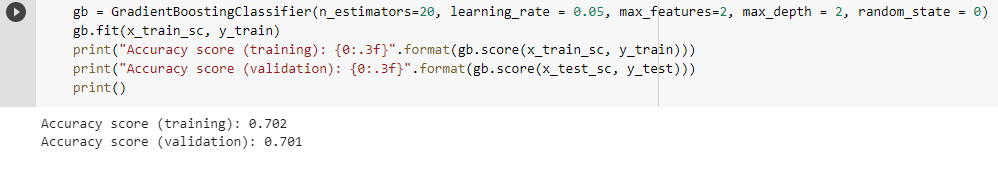


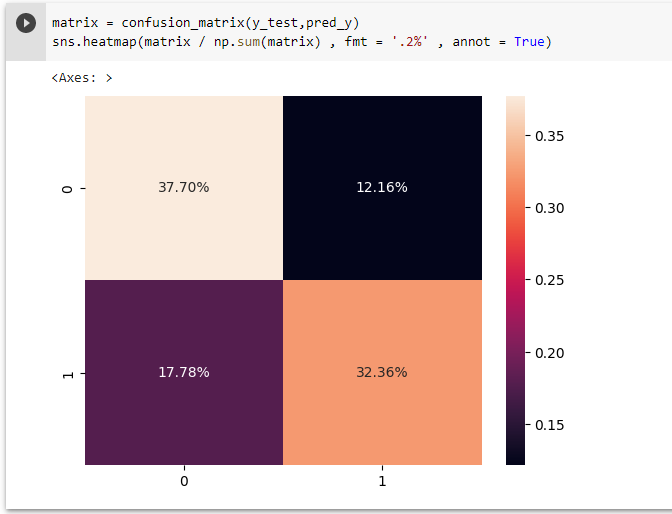


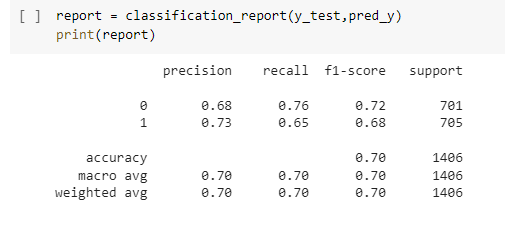
**Gradient Boosting Classifier**

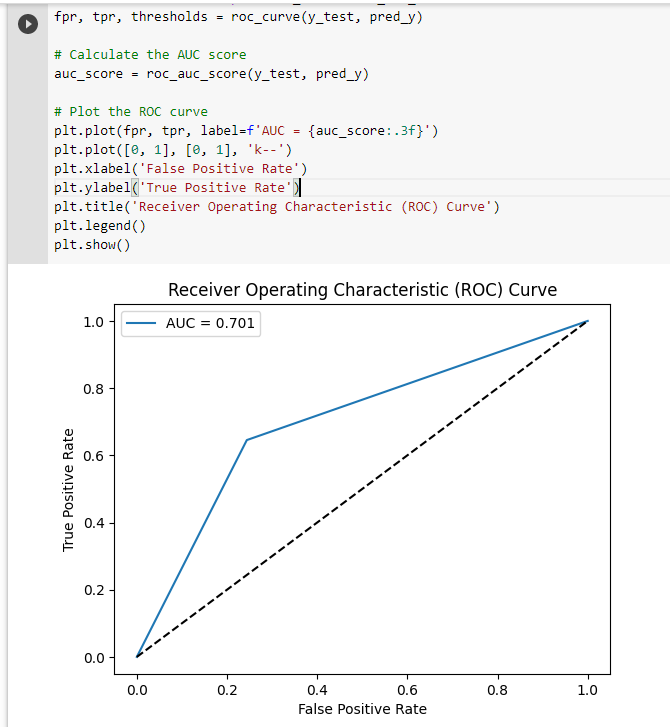
The results of this proposed model with gradient boosting algorithm achieves accuracy (70.2 %). Further, with validation of (70.1%), the proposed

We have used Kfold to break data into into smaller samples and have extract out insights of data; for e.g. we have used the prediction model to predict CVD, we have used confusion matrix, and classification report to display results.





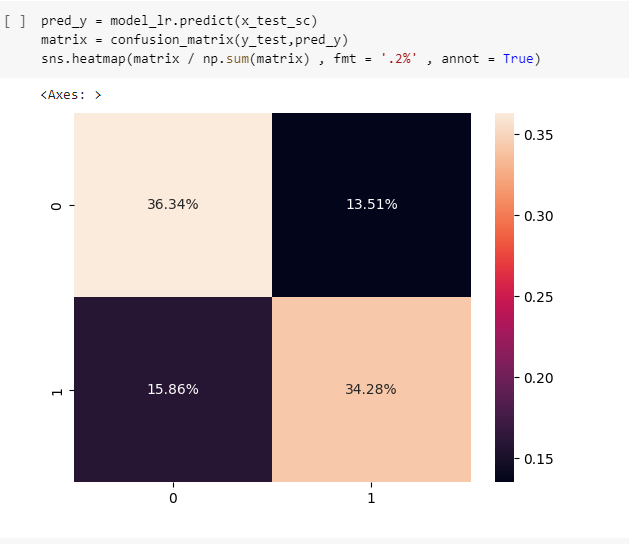


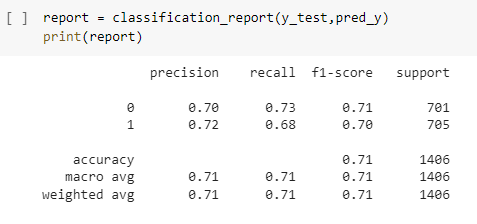


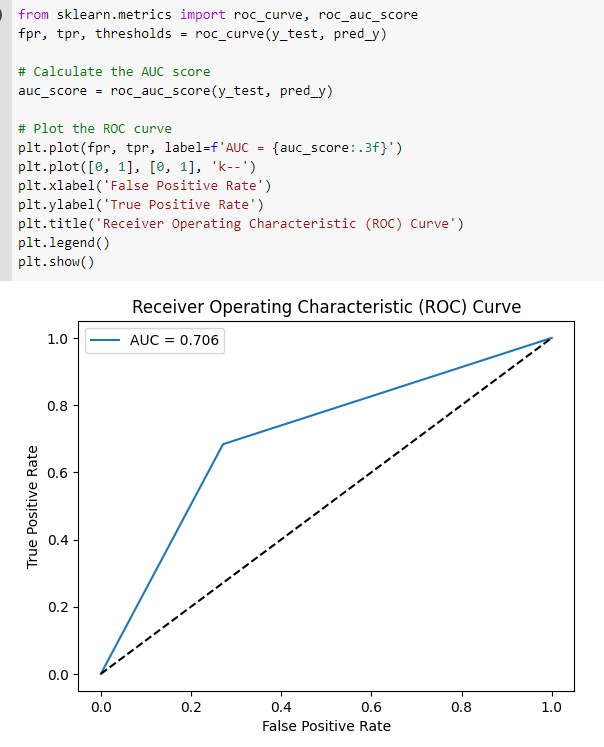
**Logistic Regression**

The Logistic Regression algorithm is one of the simple algorithms that have good classification capabilities as proven by previous studies. The proposed methodology have the accuracy of 69% for training and 70% for validation.

We have used Kfold to break data into into smaller samples and have extract out insights of data; for e.g. we have used the prediction model to predict CVD, we have used confusion matrix, and classification report to display results.



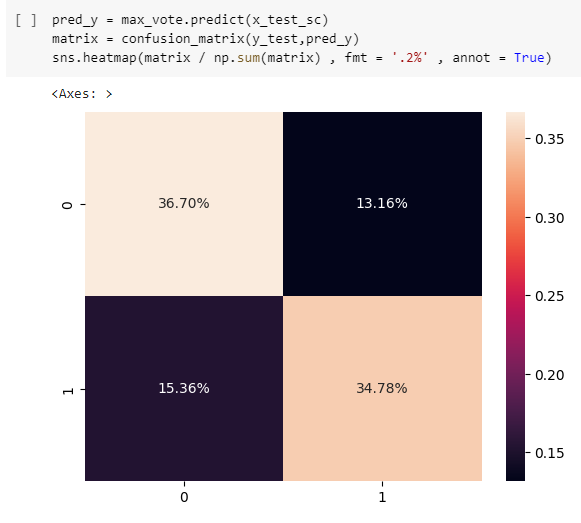


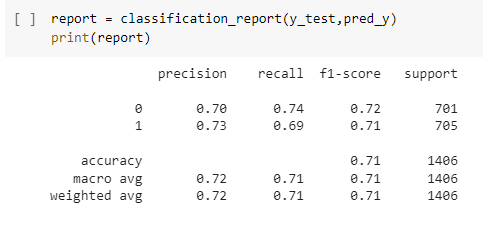


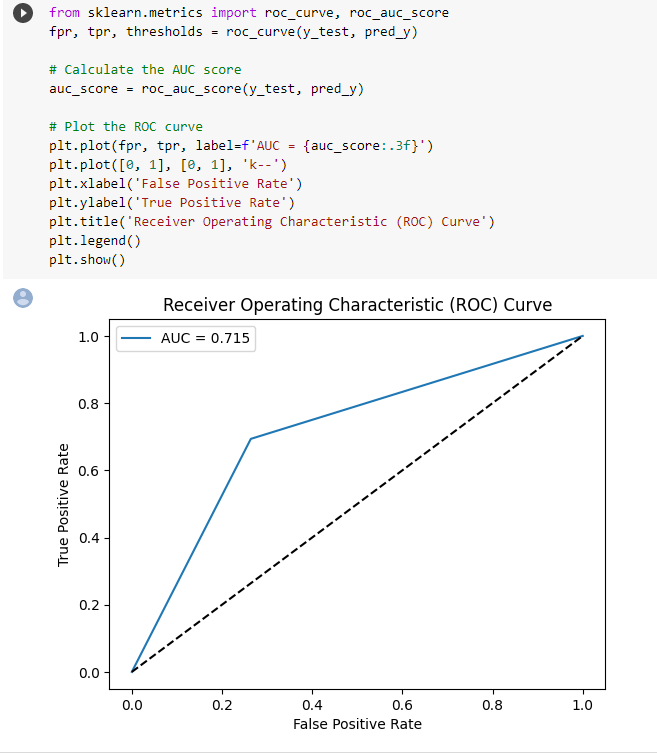
**Max Voting Classifier**

This is another classifier used for predicting cardiovascular heart disease. A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.

Voting classifier is a powerful method and can be a very good option when a single method shows bias towards a particular factor.







**Conclusion**

This study presents a comparison of four ML techniques for CVD prediction. The classification algorithms used are Logistic Regression, Decision Tree Classifier and Gradient Boosting algorithm, and Maximum Voting Classifier. Between these algorithms, Maximum Voting Classifier provides better accuracy (71.47%) than all other models. The proposed approach suggests that ML can be used as a clinical tool in the detection of CVD and will be particularly useful for physicians in the event of a misdiagnosis. The constructed MLP model offers consistent accuracy compared to other techniques mentioned and is also capable of predicting other diseases. In this study, the performance of the model was improved by removing attributes with null values using an explorative data analysis method and by increasing the number of hidden layer nodes. The proposed method is expected to assist in the further development of the healthcare field. The proposed method can also be used for the classification of other chronic diseases such as breast cancer, liver disease, diabetes mellitus, and thyroid. The developed models can be applied to large data sets to predict chronic diseases using IoT and cloud computing techniques. From the above analysis, the application of ML techniques will vastly aid in preventing fatalities and supplement the efforts of doctors in fighting CVD occurrence among all patient categories of different age groups, genders, and socio-economic backgrounds. If implemented, this would be a classic case of new-age technology application for the benefit of mankind.