Heart Attack Prediction with Classification Algorithms

Prajjwal Pandey
Computer Science
Arizona State University
Tempe Arizona United States
ppande27@asu.edu

ABSTRACT

One of the most complicated human diseases in the world has long been thought to be heart disease and it is the leading cause of death worldwide. Heart disease patients need an immediate diagnosis, prompt treatment, and ongoing monitoring. Heart disease diagnosis and prediction have historically employed a variety of data mining techniques but the techniques that are used in early stages to identify heart disease were complicated, and its complexity is one of the major reasons that affect the standard of life. Machine learning prediction models are created in order to address these issues with the complexity of the earlier methods of diagnosing a heart attack. This project is aimed to test the accuracy of several algorithms against each other and determining which algorithm is performing the best. Ultimately, an ensemble learning technique, Stacking is used to combine multiple classification models via a meta classifier and improve the accuracy. The goal is also to find the features which are major symptoms of heart attack and finding the chances of heart attack within an age-group. Ensembling technique increased the accuracy of the model as compared to other models, like, Logistic Regression, Naive Bayes.

In computing the accuracy with significant predictors in heart disease prediction, this work was able to make a major contribution. The features which are major symptom of heart attack and age-group more prone to heart attack, based on the dataset are also found.

1 Introduction

Heart disease has been the leading cause of death worldwide for the last few decades, even in developed countries like the United States where the rate is 1 death in every 34 seconds from cardiovascular disease. A heart attack occurs when blood flow to the heart is greatly reduced or blocked. Symptoms of a heart attack might vary. Mild symptoms are present in some people while Others display severe symptoms and some individuals show no symptoms. These symptoms can be because of several risk factors or abnormal values of different features like age, cholesterol, exercise induced angina.

The field of data mining and machine learning are quite broad and diversified, and these are becoming more widely used. Different classifiers are included in machine learning. Using a given data file, supervised, unsupervised, and ensemble learning techniques are used to predict and look for accuracy. This information will be helpful to many individuals; therefore, it can be used in this Heart Attack Prediction project.

In this study, a heart disease prediction system was presented to use different data mining techniques and algorithms to identify approaching heart disease and an ensemble method with increased accuracy. Six data mining techniques—Logistic Regression, Naive Bayes, Random Forest Classifier, K-Nearest Neighbor, Decision Tree, and Support Vector Machine are used in this study and compared to the ensemble technique of stacking. The other goal of this project is that people often don't know which factor is most responsible for causing heart disease. So, the feature which are major symptoms of heart attack are also predicted.

This project's goal is to improve the accuracy from the existing methods in predicting, depending on the patient's medical characteristics—such as age, gender, cholesterol, fasting blood sugar level, etc.—whether they are likely to be diagnosed with a heart attack. A healthcare dataset on heart attack possibility, containing the characteristics and medical background of the patients, is chosen from Kaggle. The prediction about the patient's potential for heart disease is done by using this dataset and categorize the patients according to age group of which age group is likely to be prone to heart disease. Out of the algorithms-Logistic Regression, Naive Bayes, Random Forest Classifier, K-Nearest Neighbor, Decision Tree, and Support Vector Machine used, Ensemble with Stacking increased the accuracy to 90%.

2 Related Work

Several data mining techniques have been used in the past to help address the issue of heart attack prediction. One data mining technique used to find the connections between characteristics and generate mining rules that produce specific predictions is weighted association rule mining (WARM). Here different weights are assigned to features based on their importance [1]. In one of the data mining projects of predicting heart disease using machine learning algorithms. Various Algorithms like Logistic Regression, KNN, and Random Forest Classifier were used where the accuracy resulted in 87.5%. Dataset used in this paper was from UCI repository. After data collection, extraction and preprocessing, the accuracy of three different classifiers were tested against each other and one classifier is used to get the desired results [4].

Ensemble method based on distance for a KNN method is also used to predict the heart attack for patients. Two implementations were made in this method. The first uses three distances while the

other uses five distances. Additionally, based on the average accuracy each distance offers when used to the KNN algorithm a weighted version was added to them. The ensemble produced an average accuracy of close to 85% [3]. Machine Learning along with IOT techniques have also been used to predict the heart attack of a person. Sensors are used to record various features of patient like blood pressure, pulse rate and temperature. The patient's real-time data is collected from sensors and analyzed by a machine learning model to foretell the likelihood that the patient would develop cardiac disease. If the patients have a high likelihood, they should speak with their doctor right once to begin therapy without wasting much time [2]. Another study introduces a number of machine learning techniques for heart disease prediction that make use of patient data on key health indicators. In order to construct the prediction models, the study exhibited four classification techniques: Multilayer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), and Nave Bayes (NB). Before creating the models, data processing and feature selection were done. On the basis of accuracy, precision, recall, and F1-score, the models were assessed [5]. Neural networks have been used to diagnose and predict heart disease, high blood pressure, and other features [7]. To ensure an accurate result of having heart disease if we use the model for Test Dataset, a deep neural network was constructed incorporating the provided disease-related attributes. This network was able to produce an output carried out by the output perceptron and almost included 120 hidden layers.

3 Data Source

The dataset used consists of 76 attributes out of which 14 attributes are selected, the “target” field refers to the presence of heart disease in a patient. It is 0 when there is no or less chance of heart attack and 1 when there is more chance of heart attack. This data collection on heart disease was obtained from the UCI repository. The dataset contains medical history of 303 different patients of different age groups. The medical parameters of the patient, such as age, sex, chest pain type, resting blood pressure, fasting blood sugar level, maximum heart rate achieved etc., provided by this dataset allow us identify patients who have been diagnosed with heart disease or not. The dataset is split into train and test set where 80% goes into training data and 20% goes into test data. Each entry in this dataset, which has 303 rows and 14 columns, represents a single record. Table 1 contains a list of all features [4].

S. No	Observation	Description	Values
1.	Age	Age in Years	Continuous
2.	Sex	Sex of Subject	Male/Female
3.	CP	Chest Pain	Four Types
4.	Trestbps	Resting Blood Pressure	Continuous
5.	Chol	Serum Cholesterol	Continuous
6.	FBS	Fasting Blood Sugar	< ,or > 120 mg/dl
7.	Restecg	Resting Electrocardiograph	Five Values
8.	Thalach	Maximum Heart Rate Achieved	Continuous
9.	Exang	Exercise Induced Angina	Yes/No
10.	Oldpeak	ST Depression when Workout compared to the Amount of Rest Taken	Continuous
11.	Slope	Slope of Peak Exercise ST segment	up/ Flat /Down
12.	Ca	Gives the number of Major Vessels Coloured by Fluoroscopy	0-3
13.	Thal	Defect Type	Reversible/Fixed/Normal
14.	Num(Disorder)	Heart Disease	Not Present /Present in the Four Major types.

Table 1: All features used in the dataset

4 Methodology

The project examines number of machine learning algorithms and the comparison of them against each other. The algorithms used are Logistic Regression, Naive Bayes, Random Forest Classifier, K-Nearest Neighbor, Decision Tree, and Support Vector Machine. Some of these algorithms are used in previous papers to predict the heart attack. This project also improves the accuracy obtained from these algorithms using ensemble technique which uses stacking. The proposed method (Figure 1.) consists of several steps where first is to load the dataset and then the profile report of data set is created followed by pre-processing of data set which includes standardizing, missing (%), finding duplicate values, distinct count, mean and unique (%). After this, the data is split into training and test set where 80% is training data and 20% is test set. When creating an ideal prediction model, ensemble techniques combine several learning algorithms or models. When compared to the base learners alone, the model developed performs better. The selection of key features, data fusion, and other uses of ensemble learning are also possible. The three main categories of ensemble techniques are bagging, boosting, and stacking. In this project Ensemble using stacking is used. In stacking, heterogeneous weak learners are taken into account, learned in parallel, and combined by training a meta-learner to output a prediction based on the predictions of the various weak learners. A meta learner attempts to learn the optimal way to combine the input predictions to produce a better output prediction by using the predictions as input features and the ground truth values in the data D as per the figure 2. In stacking, an algorithm learns how to optimally combine the input predictions to get a better output prediction by using the outputs of sub-models as input.

The various algorithms mentioned in the figure including ensemble with stacking, where there are three classifiers in level 1 (Random Forest, K-Nearest Neighbor, Support Vector Machine) and meta-classifier (Support Vector Machine), are run and compared using bar plots and then the classification report is created and compared for all the algorithms. In Data Visualization, Male and Female with Heart Disease are visually

represented on a bar plot. In a similar way, people with heart disease are represented using a bar plot categorized by the age group. The important feature which is major symptom of heart attack is predicted using SelectKBest class and chi-squared statistical test, using feature importance of random forest model, and using correlation matrix to indicate how features are related to each other or to the target variable and the heart attack possibility is categorized according to the age group on the test set.

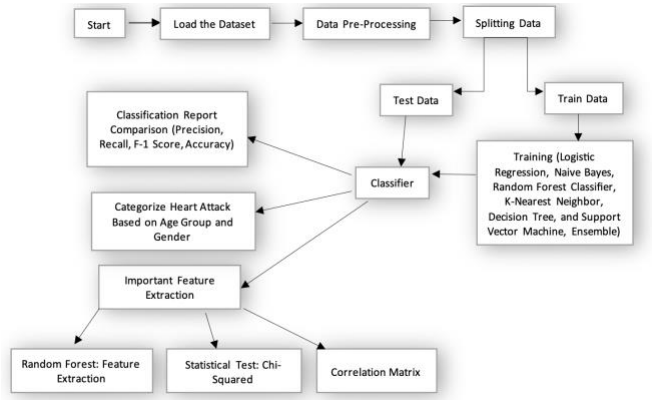

Figure 1: Proposed Method

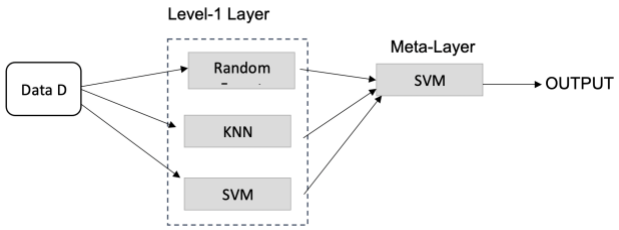

Figure 2: Ensemble Stacking

5 Results and Discussions

From the test conducted on different algorithms, Logistic Regression, Naive Bayes, Random Forest Classifier, K-Nearest Neighbor, Decision Tree, and Support Vector Machine. The accuracy of Logistic Regression was 85%, Naïve Bayes at 85%, Random Forest Classifier at 87%, Decision Tree Classifier at 82%, KNN at 89%, and SVM at 89%. The accuracy using the ensemble method with level 1 classifiers (Random Forest, K-Nearest Neighbor, Support Vector Machine) and meta-classifier (Support Vector Machine) where came out to be 90% (figure 3). The algorithms are faster and more precise than those employed by earlier studies. They also save a significant amount of money, making them very cost-effective. In Data Visualization, it seems that males are more susceptible to heart attacks than female (figure 4) and the disease is particularly common in seniors, defined as those who are 60 years of age and older, as well as in individuals who fall between the age range of 41 to 60. However, it's uncommon among people aged 19 to 40 and extremely uncommon among those aged 0 to 18 (figure 5). With Chi-Squared statistical test for selecting 10 best features, thalach (maximum heart rate achieved) came at the top, while with feature importance of random forest (figure 6) and correlation matrix (figure 7), cp (chest pain) was the most significant feature. The fundamental advantage of stacking ensembles is that they can protect a variety of effective model's abilities to address classification and regression issues. Additionally, it helps to develop a superior model with predictions that outperform all individual models and in this project ensemble method increased the accuracy than rest of the methods.

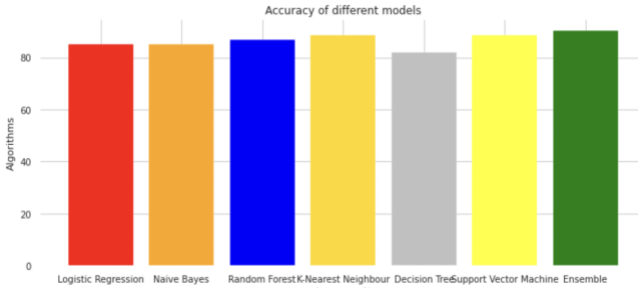

Figure 3: Accuracy percentage of different models

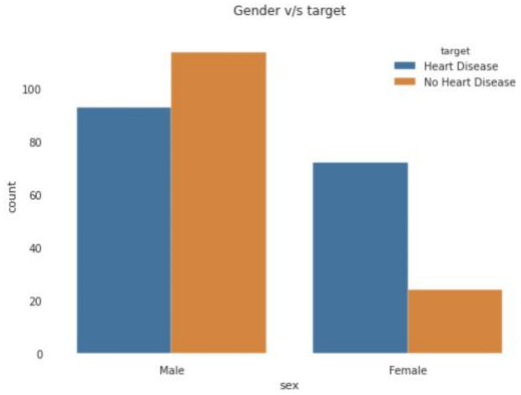

Figure 4: Gender v/s target

Figure 5: Age v/s Count of Heart Diseased Patients

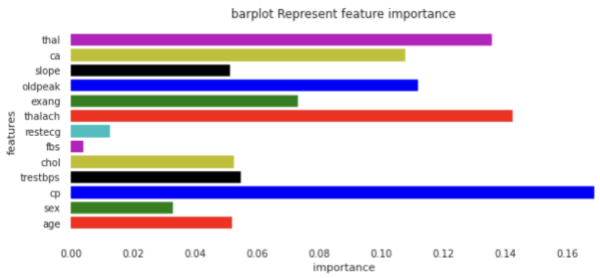

Figure 6: Random Forest Feature Importance

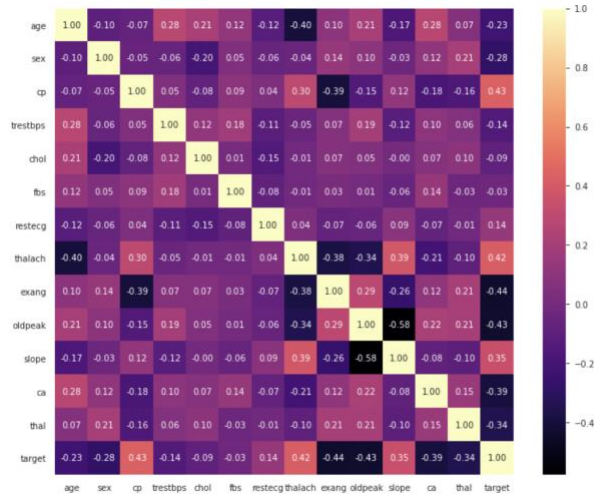

Figure 7: Correlation matrix of features

6 Conclusion and Future Work

Utilizing ML classification modeling techniques, a cardiovascular disease detection model has been created. By extracting the patient medical history that results in a fatal heart illness from a dataset that contains patients' medical history such as chest pain, sugar level, blood pressure, etc., this method predicts persons with cardiovascular disease. The accuracy of the ensemble stacking model is 90% which exceeds the accuracy of individual models- Logistic Regression, Naive Bayes, Random Forest Classifier, K-Nearest Neighbor, Decision Tree, and Support Vector Machine. The true positive and false positive rates of different models are also compared which are displayed in Figure 8 where true positive rate of ensemble is better than the rest of the algorithms. The confusion matrix of ensemble methods denotes significant improvement where true positives were 23 and true negatives were 32 and false positives were 3 and negatives were 2 (figure 9). The advantage of stacking is that it uses predictions from various machine learning models to create an even better predictive model is known as an ensemble of models. The likelihood that the model will correctly identify whether a specific person has heart disease or not increases with the use of more training data [6]. The model also segregates the predicted heart attacks on various age groups of the test set (figure 10). This method can also be used to predict other diseases and the current accuracy of this model can be improved if we choose a different set of models in layer-1 and meta layer of the ensemble such as including extreme gradient boost whose accuracy is better than the other individual models. Example of ensemble averaging which integrates forecasts from other trained algorithms. The fact that each model contributes the same amount to the ensemble forecast, regardless of how well the model performs, is a drawback of this approach. A weighted average ensemble is an alternative strategy that weighs each ensemble member's input according to the confidence in their ability to produce the best predictions. The model average ensemble is outperformed by the weighted average ensemble.

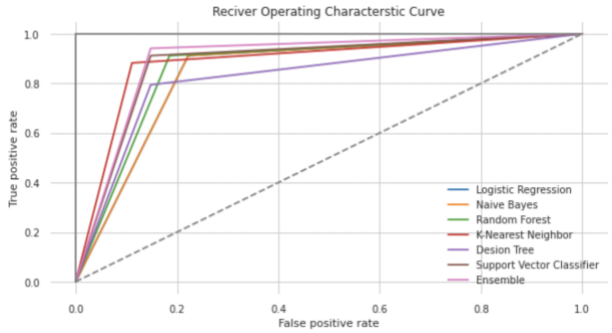

Figure 8: Comparison between true positive and false positive rates of various models

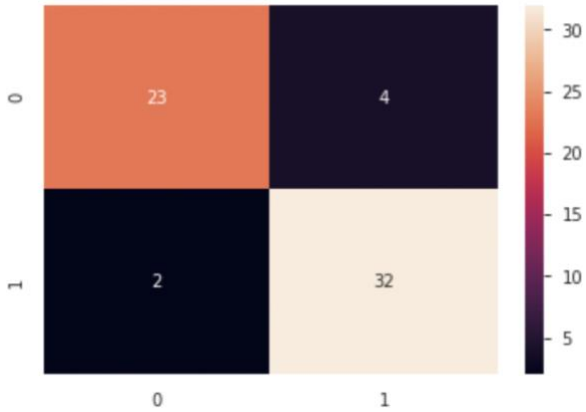

Figure 9: Confusion matrix of Ensemble

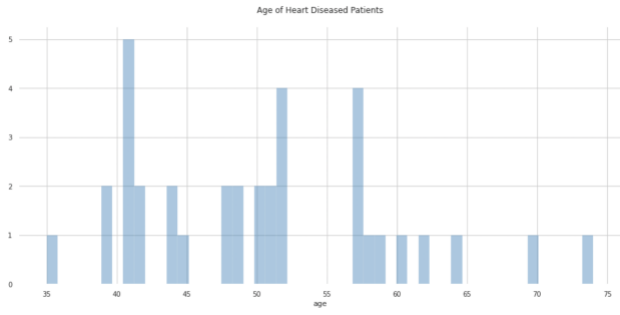

Figure 10: Heart disease segregation by age group

REFERENCES

[1] Yazdani, A., Varathan, K.D., Chiam, Y.K. et al. A novel approach for heart disease prediction using strength scores with significant predictors. BMC Med Inform Decis Mak 21, 194 (2021). DOI: https://doi.org/10.1186/s12911-021-01527-5

[2] Gangadhar Immadi, Akashansh Jain, Bhavya R, Kishlay Kumar, "Heart Disease Prediction Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6, Issue 3, pp.654-663, May-June-2020. DOI: https://doi.org/10.32628/CSEIT2063149

[3] A. P. Pawlovsky, "An ensemble based on distances for a kNN method for heart disease diagnosis," *2018 International Conference on Electronics, Information, and Communication (ICEIC)*, 2018, pp. 1-4, DOI: https://doi.org/10.23919/ELINFOCOM.2018.8330570.

[4] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath, "Heart disease prediction using machine learning algorithms," 2021 *IOP Conference Series: Materials Science and Engineering*, 2021, pp. 1-4, DOI: https://doi.org/10.1088/1757-899X/1022/1/012072.

[5] C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," 2022 Advances in Science and Engineering Technology International Conferences (ASET), 2022, pp. 1-6, Doi: 10.1109/ASET53988.2022.9734880.

[6] Dangare Chaitrali S and Sulabha S Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10 (2012): 44-8.

[7] Kiyasu J Y (1982). U.S. Patent No. 4,338,396. Washington, DC: U.S. Patent and Trademark Office.