Prediction of Heart Disease and Determining the Factors Influencing it

Smriti Tilak
Computer Science and Engineering
PES University
Bangalore, India
smrititilak@gmail.com

Kakumani Sai Deepika
Computer Science and Engineering
PES University
Bangalore, India
ksdrao7@gmail.com

Baddela Swathi Reddy
Computer Science and Engineering
PES University
Bangalore, India
swathireddybaddela@gmail.com

Deepa S
Computer Science and Engineering
PES University
Bangalore, India
deepadhruthi13@gmail.com

***Abstract*—The goal of this project is to determine if a patient could potentially have heart disease or a failure on the basis of his or her previous medical records. It also aims to identify the factors that influence the patient's risk of having heart failure using various statistical methods and picking the one which provides the most accurate results. This paper includes a brief introduction to our problem along with a review of other papers that discuss a similar problem. It also includes the visualizations for the dataset and our initial analysis and insights in order to accomplish a better understanding of it. In addition to this, it includes the classification models we chose to solve the problem and the insights we gained from it.**

Keywords—heart disease, heart failure, prediction, classification, data analysis, XGBoost, Decision Trees, Logistic Regression.

I. Introduction

The heart is one of the most important muscles of the body responsible for pumping blood received from veins and into arteries throughout the body. It is responsible for oxygenating the blood and which is essential to keep the human body strong and healthy. The heart, like any other human organ, is prone to diseases. These are collectively called cardiovascular diseases. Cardiovascular diseases affect the structures or functions of the heart, such as Arrhythmias, congenital heart disease, etc. Heart failure (HF), or congestive heart failure (CHF) or congestive cardiac failure (CCF), are manifestations that are due to the failure of the functionality of the human heart; its symptoms are the consequences of a structural and/or functional abnormality of the heart. HF is one of the major causes of death today. According to the Centers for Disease Control and Prevention, about 6.2 million people in the United States of America have HF. In 2018, HF was stated as the cause of death in 379800 death certificates. HF cost the nation an estimated total of 30.7 billion USD in 2012. The total included the cost of healthcare services, medicines and missed days of work. It is known to us that with age, people are prone to HF even more. In addition to this, the lifestyle of the people plays a very important role in their lifespan. The economic burden is also predicted to rise particularly, given the chronic nature of HF and the high risk of hospitalisation. Therefore, it would be extremely beneficial to healthcare providers and payers if the risk of future outcomes can be predicted associated with HF and a number of risk prediction models have been developed for the same.

For our model, we considered a dataset [1] obtained from Kaggle. It has 12 attributes in total, of which one of them is a binary attribute stating if the patient has heart disease or not which is our target attribute for classification. It has 918 records which amounts to a total of 11016 values. We use various statistical models for classification and determining influential factors that can help predict HF. The aim is to build a model that can accurately predict if a patient has the risk of a heart failure given his medical record containing data like his age, cholesterol levels, type of chest pain, etc. In addition to this, we also analyse which attributes or factors of the patient's history are found to influence the risk of him having an HF. This analysis was done to uncover any patterns in the patients' medical histories.

II. Literature Review

A. *Machine Learning Techniques for Heart Failure Prediction*

This research paper [2] aims to determine which Machine Learning (ML) model was appropriate for early HF prediction. The paper compares four ML models – Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR) and Naïve Bayes (NB). Various performance metrics like accuracy, precision, specificity were used to determine how well the models predicted the dataset. This study also investigates to find which factors influenced coronary heart failure the most, out of the 13 features that it considered. It checked to see if the elimination of any of the factors still gave accurate predictions. The dataset for the study was obtained from Kaggle and was called the heart disease dataset containing 303 patients with 14 variables: age, sex, cp (chest pain

experience), chol (cholesterol), target (1 - heart failure, 0 - no failure) etc.

Random Forest (RF) is a learning technique that uses Decision Trees (DT) for classification and is known to work well on large datasets. Multiple DTs are constructed for a single RF model. The final prediction result is based on votes given by the individual DTs to every intermediate prediction. Support Vector Machine (SVM) is an ML technique (supervised) that uses a hyper-plane to separate the classes. The hyper-plane serves as the decision boundary. Naïve Bayes (NB) uses Bayes Theorem for classification. The probability of different classes is predicted using NB based on attributes. Logistic Regression (LR) is used to measure the relationship between multiple dependent variables and the independent ones. In this problem considered, the target class was the dependent variable with the thirteen attributes being independent.

Using these ML techniques, four prediction models were built. For SVM, linear kernel function gives the most accurate predictions among other kernel functions, hence, this prediction was used for comparison with that of other techniques (RF, NB and LR). RF showed the best results for sensitivity and recall, and both were 0.97. RF and LR achieved the same values of accuracy and f1-score which are 0.88 and 0.9 respectively. SVM achieved the highest validation accuracy among all models and was 0.9 and NB didn't outperform any of the models in terms of any of the metrics. Hence, the study concluded that "the highest average performance was achieved by RF followed by LR, NB and SVM. Another experiment was conducted to determine the least significant feature so as to eliminate these one at a time. Fasting blood sugar was found out to have the least influence on the prediction model and it was followed by Resting ECG value."

B. *Assessing risk prediction models for predicting heart failure in adults.*

The main aim of this research paper [3] was to build risk prediction models for predicting heart failure in adults. In order to achieve these many risk prediction models were developed. For this, they searched for literature databases from March 2013 to May 2018. They found that many prediction models were developed based on multi variables instead of single/uni variables. All these models were developed like this in order to overcome the problem of the risk of bias. Bias was assessed with a tool called PROBAST. When measured for bias, if the overall result was low, then the model was considered to be a good model and if it was high, the model was not considered totally bad but it was confirmed that some factors influenced it.

Out of the data that they considered, they found only a few publications that had relevant and consistent information. Some of the factors like mortality, heart failure and many other composite endpoints were calculated. The c-statistics for each one of these factors were calculated and the results for heart failure ranged between 0.59 and 0.826, composite points ranged between 0.6 and 0.7 while the median for all cause mortality ranged between 0.6 and 0.8. From this study, it was observed that most of the models were concerned only about the risk of bias and the reasons for these were lack of authentication and standardisation.The usefulness of ML tools for developing prediction tools is yet to be fixed.

C. *Congestive heart failure detection using random forest classifier*

The main motive of this paper [4] is to create a machine learning model which assorts normal and congestive heart failure (CHF) on the long-term ECG time series. According to this paper heartbeat classification is considerable for diagnosing heart failure. So they considered automatic electrocardiogram as a measure of heartbeat to predict the classification of the heart failure.

The study was performed in stages: feature extraction, classification phase. In the feature extraction phase, they have used the Autoregressive Burg method for extracting features. As a part of the model building phase they have experimented with many models like K Nearest Neighbours, SVM , Random Forest Classifier, Artificial Neural Network, C4.5 Decision Tree. The ECG signals are classified by applying various models. The data set was evaluated in terms of several performance measures like sensitivity, ROC curve, specificity, F-measure, accuracy and exhibited 100% classification accuracy while using random forest classifier. And they have finally concluded that the performance of the random forest method proves that it plays an foremost role in classifying congestive heart failure (CHF) and can be valuable in denoting knowledge which is useful in medicine.

D. *Congestive Heart Failure Symptoms in Patients with Preserved Left Ventricular Systolic Function: Analysis of the CASS Registry*

In this paper [5] In total two goals were set and were analyzed. One was to do a comparison between characteristics and how long a patient survives with heart failure. Second one was to check how the presence of coronary artery disease affects the survival of patients. Analysis was made on each cardiac segment and those were graded from 1 to 6. Final score was calculated and if the score is 5 it concludes that it is good left ventricular systolic function. The chi-square and Student's t tests for continuous variables along with categorical variables were used to compare the clinical characteristics between two groups. Kaplan-Meier method along with the statistical comparisons between groups was used to compute survival function. It was found that patients who got heart disease were older ones and are more likely women. Also, they have high hypertension, diabetics and chronic lung disease. As the Population continues to rise, the number of patients with symptoms of heart failure and coronary disease will be more. The patients will have to be carefully tested and treated for long term survival.

III. The Dataset

The dataset we've considered for our study is taken from Kaggle [1]. It contains 12 attributes with 918 observations. Each of these observations is a medical record of a patient indicating various health measures and it also includes a variable target that indicates if the patient has faced a heart failure. Using this data we aim to predict if a patient's medical record should be classified as one that can have a heart attack. The variables included in the dataset are:

- Age: Patient's current age.
- Sex: Patient's gender/sex.
- Chest Pain: Qualitative variable indicating the type of pain in the chest (TA - Typical Angina, ATA - Atypical Angina, NAP - Non Anginal Pain or ASY - Asymptomatic).
- RestingBP: Patient's Resting Blood Pressure
- Cholesterol: Indicates the cholesterol levels of a patient.
- FastingBS: Binary variable indicating fasting blood sugar (0 - if less than 120 mg/dL, 1 - if more than 120 mg/dL)
- RestingECG: Qualitative variable that can take the values - normal, ST (having ST-T wave abnormality) and LVH (showing probable or definite left ventricular hypertrophy).
- MaxHR: Indicates the max. heart rate that the patient's heart can achieve.
- ExerciseAngina: Exercise-induced angina (Y - yes or N - no)
- Oldpeak: ST (positions on the ECG plot) depression induced by exercise relative to rest.
- ST_Slope: Qualitative variable indicating slope of the peak exercise ST segment. It can take 3 values (Up - upsloping, Flat and Down - downsloping)
- HeartDisease: Binary variable indicating if the person had heart problem/failure or not

IV. Proposed Solution

Our data had no missing values, hence the next step in our preprocessing was dealing with outliers. Initially, we tried to visualize the distributions of the various attributes through their histograms and pie charts. We noticed that the attribute RestingBP and Cholesterol had outliers as shown in figures 1 and 2 respectively.

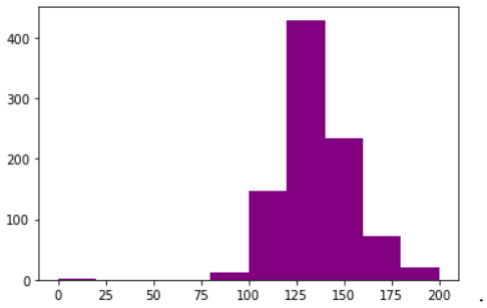


Figure 1

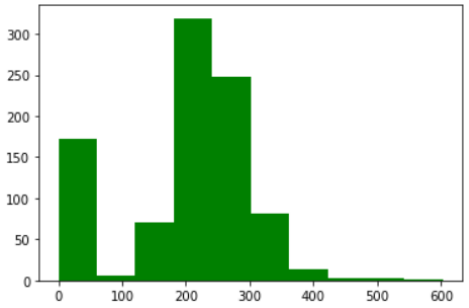


Figure 2

The presence of outliers in a prediction model can drastically affect its accuracy. Hence, it was important to nullify the effect of the outliers. These outliers were dealt with, using the interquartile range method (IQR method). In this method, the only accepted values fall within the range of the tenth and ninety-fifth percentile. The rest of the values are replaced with the median of the accepted range. Once, the outliers are handled, we get the following graphs for RestingBP (Figure 3) and Cholesterol (Figure 4) respectively:

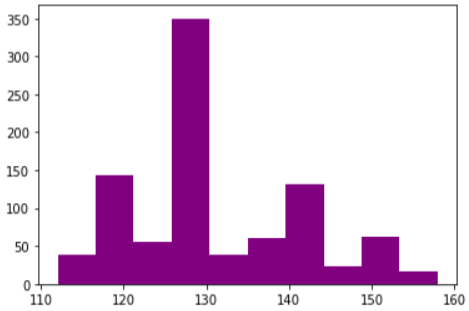


Figure 3

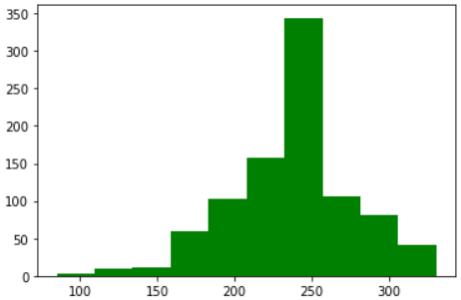


Figure 4

As it can be observed from these graphs, outliers seem to have been successfully removed. After the removal of outliers, we analysed the correlation between the variables as we aim to understand the influence of these factors over the target variable. The following correlation matrix was obtained.

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
Age	1.000000	0.193793	0.063797	0.198039	-0.382045	0.258612	0.282039
RestingBP	0.193793	1.000000	-0.001148	0.033950	-0.045088	0.142213	0.073347
Cholesterol	0.063797	-0.001148	1.000000	0.016227	0.009569	0.047711	0.068786
FastingBS	0.198039	0.033950	0.016227	1.000000	-0.131438	0.052698	0.267291
MaxHR	-0.382045	-0.045088	0.009569	-0.131438	1.000000	-0.160691	-0.400421
Oldpeak	0.258612	0.142213	0.047711	0.052698	-0.160691	1.000000	0.403951
HeartDisease	0.282039	0.073347	0.068786	0.267291	-0.400421	0.403951	1.000000

From the correlation matrix, we can infer that MaxHR and the target variable (HeartDisease) have a negative correlation, indicating that a person who can achieve higher heart rates could be less prone to heart failure. This is true in general, as younger people can achieve much higher heart rates compared to elders and heart failure is low or rare among the youth. The same is indicated by the negative correlation between the variables Age and MaxHR.

We went on to build four different models for the dataset. They are Logistic Regression, Decision Trees, Random Forest and XGBoost. All three models were known to work well with binary classification problems, hence we opted to choose them for our analysis.

Logistic Regression (LR) - We started our analysis by building an LR model for the dataset. We split our dataset into 80% and 20% for training and testing the model. The categorical attributes in our dataset were encoded using one-hot encoding. A logistic regression method was modeled on our training data and we predicted the classification for testing data using this model. We used confusion metrics and few other statistics to understand the performance of the model. The classification accuracy turned out to be 88.15% on training data and 82.61% on testing data. As the performance digits were not substantial we proceeded to build a few other models.

Decision Trees (DT) - We proceeded by building a decision tree model for our dataset. We start off by splitting our dataset into 85% for training and 15% for testing. Categorical attributes like sex, ChestPainType etc., were encoded using label encoder for further processing. The target class was dropped from the original dataset and saved separately for future comparisons.

A DT model was built. We fitted it on the training data split and then used this model to predict the target class for the test split. We separately wrote some common functions to obtain various performance metrics like accuracy, f1 score, precision and recall. The initial results were observed as follows:

	train_set	test_set
Accuracy	1.0	0.804348
Precision	1.0	0.845070
Recall	1.0	0.789474
f1	1.0	0.816327

As it can be observed on the test set, the accuracy was only about 80% and precision was 84.5%. We tried to improve on these metrics by tuning the hyperparameters using grid search and performance metrics for the same are given below:

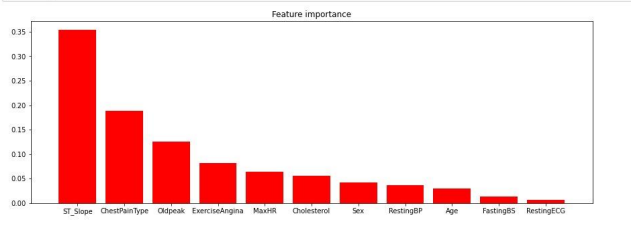
	train_set	test_set
Accuracy	0.857692	0.891304
Precision	0.839323	0.850575
Recall	0.918981	0.973684
f1	0.877348	0.907975

As it can be seen from the image, the accuracy improved to 89.13% and the precision improved to 85.05%. Recall saw the most improvement and hence we concluded that after tuning the hyperparameters, the model performed much better and these were the best performance metric results we obtained for DT. After tuning the hyperparameters, the overfitting was reduced as can be observed from the reduction of performance metrics for the train split.

XGBoost - We proceeded to build the gradient boosted decision trees i.e., XGBoost. We decided on this model as we thought that using gradient boost on the DT would help us achieve better accuracy. The same training and test splits were used for the DT model. We built the XGBoost model and trained it using the train split and predicted the values for the test split. The accuracy obtained was 86.23%. We performed grid search for hyperparameter tuning and obtained an accuracy of 87.68%. Hence, DT performed better than XGBoost.

Random Forest - Finally, We take measures to build the Random Forest Classifier. We decided on this model as one of the papers we reviewed earlier used this model on some other dataset and obtained 100% classification accuracy. The same training and test splits were used for the DT model. We built the Random forest model and trained it using the train split and predicted the values for the test split. The accuracy obtained was 86.47%.

Using RF we also analysed which factors influenced the risk of HF the most. The following graph was obtained:



From the graph, we can conclude that ST_Slope has the most influence over the classification, followed by ChestPainType.

V. Conclusion

From our analysis of the four models, Decision tree performed gave the highest accuracy of about 89.13% on the test split. The rest of the models performed decently

with all of the final models after hyperparameter tuning having an accuracy of atleast 82%. We conclude that decision tree performs the best on the given data. The factors that affcet HF the most are - ST_Slope closely followed by ChestPainType. All the models and analysis can be found on GITHUB.

REFERENCES

[1] URL to the dataset to be used for our analysis: Heart Failure Prediction Dataset | Kaggle

[2] https://www.researchgate.net/publication/353834616_Machine_Learning_Techniques_for_Heart_Failure_Prediction

[3] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224135

[4] https://www.sciencedirect.com/science/article/abs/pii/S0169260715303369

[5] https://www.sciencedirect.com/science/article/pii/0735109791905892

[6] smrititilak/Heart-Failure-Prediction-and-Analysis (github.com)