Heart Disease Prediction

Manak*	Manaswi	Pankaj Kumar	Garima Gupta
CSE Department	CSE Department	CSE Department	CSE Department
MAIT	MAIT	MAIT	MAIT
Rohini, Delhi	Rohini, Delhi	Rohini, Delhi	Rohini, Delhi
manakbansal@gmail.com	13.manaswi@gmail.com	pankajkumar.mait@gmail.com	garimagupta@mait.ac.in

Abstract- This document explores the possibility of the prediction whether a person is susceptible to various heart diseases like Coronary Artery Disease (CAD), Heart Arrhythmias, Heart Failure, Heart Valve Disease, Pericardial Disease, Cardiomyopathy (Heart Muscle Disease), Congenital Heart Disease and many more which has put a great threat to human beings given how our lives and schedules are evolving into more sedentary ones with the advent of technologies originally made to make our lives easier. A passive lifestyle puts not only our heart at risk, but also is a direct cause of more physical and mental illnesses and diseases like osteoporosis, lipid disorders diabetes, depression and anxiety, and increase the risks of different types of cancer and a higher blood pressure. In this article, we have aimed to study the various different factors that may or may not be in direct correlation of heart diseases. These factors are as follows: Age, sex, chest pain type, resting blood pressure, cholesterol in mg/dl, fasting blood sugar, resting electrocardiography results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise, slope of the peak exercise ST segment, number of major vessels and maximum heart rate. We have also compared the correlation of these factors with the possibility of a heart related illness. These factors are elaborated in a more detailed way in this paper. And for the same, we have used multiple algorithms (logistic regression, naïve bayes, Support vector machine, KNN, decision tree, random forest and artificial neural network) and compare the results to find out the most accurate one. We are using dataset from kaggle.com.

Finally, we found out the results to be around 90% accurate so there is a 90% probability of the results being accurate when we feed custom data to the algorithm.

Keywords- angina, heart diseases, random forest, heart diseases prediction, classification, ensemble

I. *INTRODUCTION*

As we all know, heart is a vital organ of our body. It has a very important task in our body, pumping blood to every part of our body. If it fails to function correctly, then the rest of the organs will stop working, and within a few minutes, the person will die. There could be multiple potential reasons behind a threatening heart disease including but not limiting to sedentary lifestyle, overwhelming stress and bad diets.

Heart diseases and related illnesses are seen as one of the most prominent causes of death all around the world. According to the World Health Organisation, heart related diseases are responsible for taking 17.7 million lives every year, 31% of all global deaths. Situation in India too is not different, heart related diseases have become the leading cause of mortality. There has been a constant rise in the heart diseases related deaths in our country. The numbers rose around 34% from 1990 to 2016 [1]. According to the Global Burden of Disease study, CVD has a death rate of around 272 per 1,00,000 people which is significantly higher than the global average [2]. CVD directly increases the pressure on any nation's health care and adversely impacts the productivity. As estimated by the World Health Organisation (WHO), CVD have costed India up to $237 billion, from 2005-2015. Thus, there is a grave need of a model that could predict, and in turn help prevent, the cardiovascular diseases. And that is exactly what we have endeavoured to do in this article.

II. *LITERATURE REVIEW*

There are many literature contributions under the same field, using a wide variety of technologies and methods to achieve better and efficient results. In the paper by Enriko et al, they achieve an accuracy of almost 82% and argue that the factors that we are including are sufficient enough for prediction. In, Boshra et al used K-Nearest Neighbors, J48 Decision tree, SMO and Naive Bayes and the 8 attributes with WEKA validation. The highest accuracy came out to be 83.7%.

In one article which used ANN with feature correlation analysis was studied on Sixth Korea National Health and Nutrition Examination dataset and the authors found that chronic renal failure and triglyceride were closely related to coronary heart disease, it showed the accuracy of 82.51%. In a study by Lavanya et al, J48, CART, Naïve Bayes were implemented using the WEKA tool to obtain an accuracy of almost 86%. Haq et al in talks about the usage of plethora of machine learning predictive

models such as k-nearest neighbor, logistic regression, AdaBoos, Naive Bayes, XRBoost to get a best accuracy of 89% using LR with relief. In a study by Hidayet et al, 4 different feature selection methods and 12 classification algorithms has been used. This got the accuracy of around 85% using SVM and Naïve bayes.

If we compare ANN and neuro-fuzzy algorithms, as done by Abushariah et al, we discover that ANN has a better accuracy at 87% as compared to its counterpart which stands at around 76%. A three-phase model based on the ANN was proposed to achieve an 88.89% accuracy in a paper by Olaniyi et al. When authors of another study proposed integrated decision based on ANN and Fuzzy AHP, the attained 89.1% accuracy.

III. DATASET DESCRIPTION

Our dataset comprises of 11 different features and the target value. It has 6 categorical variables and 5 variables with numeric values. The description of attributes used is given below:

- Age: Age of subject in years (Numeric)
- Sex: Gender of subject (1-Male, 0-Female) (Nominal)
- Chest Pain Type of chest pain suffered by subject (1-typical, 2-typical angina, 3-non-anginal pain, 4-asymptomatic) (Nominal)
- resting blood pressure: blood pressure at rest in mm/HG (Numerical)
- chol: Serum cholestrol in mg/dl (Numeric)
- fasting blood sugar: Blood sugar levels on fasting > 120 mg/dl (1-true, 0-false) (Nominal)
- rest ecg: ECG results while at rest (0-Normal, 1-Abnormality, 2-Left ventricular hypertrophy) (Nominal)
- max hr: Max heart rate achieved (Numeric)
- exercise angina: Exercise induced angina (0-No, 1-Yes) (Nominal)
- oldpeak: ST-depression during exercise as compared to rest state (Numeric)
- ST slope: ST plot slope during peak exercise (0-Normal, 1-Upsloping, 2-Flat, 3-Downsloping) (Nominal)

Target variable

- target: Variable storing the result value (1-higher heart disease risk, 0-No heart disease risk).

IV. *EXPLORATORY DATA ANALYSIS*

In this section, we explore and study our data. Here are some of the attributes that we are keeping in mind.

A. Target

Here, we observe that the target values in our dataset are almost equal, hence giving it a better chance for going through all the possibilities and being as accurate as possible with the given conditions.

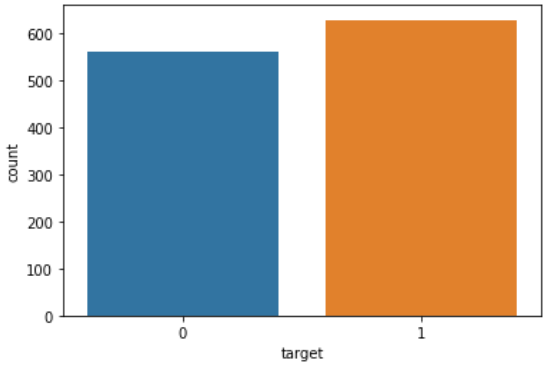

Fig. 1. Distribution of target values

B. Sex

In the dataset, when we plot the genders with respect to the target values, we observe that a higher number of males have heart problems as compared to females.

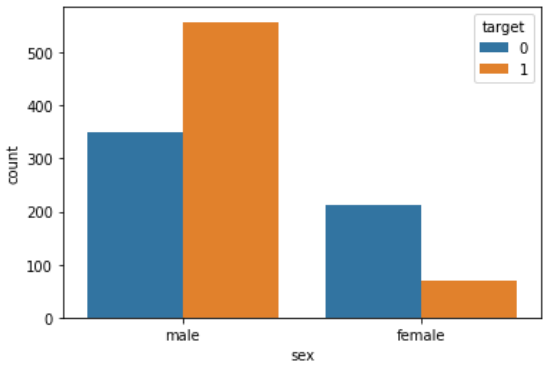

Fig. 2. Distribution of sex and target

C. Chest pain

The distribution of people experiencing chest pains with respect to target is shown below.

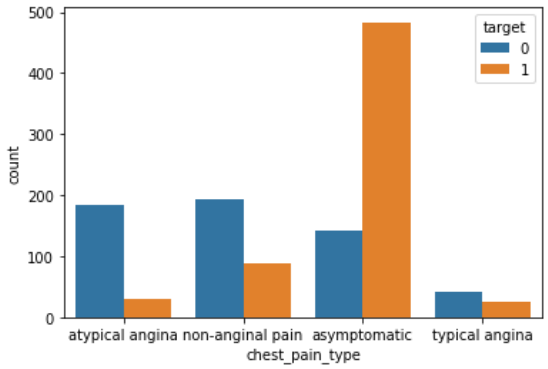

Fig. 3. Distribution of chest pain types and target

D. Resting ECG

Here we see that most of people studied had normal ECG results, values 0 and 2 shows abnormality

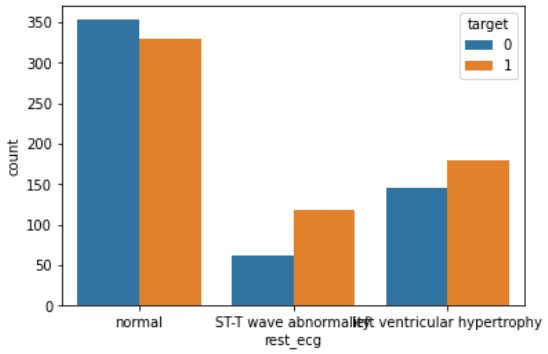

Fig. 4. Distribution of Rest ECG curve and target

G. Slope of the peak exercise ST segment

During exercise, a healthy heart should have an upslope in ST segment

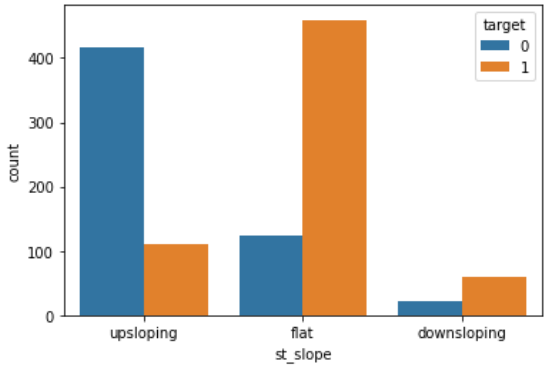

Fig. 5. Distribution of ST curve and target

V. *DATASET MANIPULATION*

We have obtained our dataset from a very popular website, kaggle.com. We made sure the datapoints and the factors were relevant to our study and made sense. Our final data contains about 1189 rows and 11 columns.

A. Missing values

First, we check for the missing values. If present, the data may have a bias. But in our dataset, there was no null values.

B. Renaming

We rename the data to proper name so they are easy to understand and work with.

C. Outlier detection and removal

Outlier detection is the process where we catch the unusual datapoints that may cause a bias due to their extreme values, and consider just usual values, in their expected range. We filter only numeric features as age, resting bp, cholestrol and max heart rate achieved has outliers as per EDA.

Once outliers are detected, these records are dropped from the table which gives the final shape of our dataset to be (1172,12), including the target value.

Now before splitting dataset into train and test we first encode categorical variables as dummy variables and segregate feature and target variable. We do this to convert categorical data into numerical values, so that the model gives an improved prediction. Once done, we get yet another dimension for our dataset, (1172,16).

Once our dataset is perfect for our use, we study the co-relation pattern between the columns and our target value. The heatmap below shows the same.

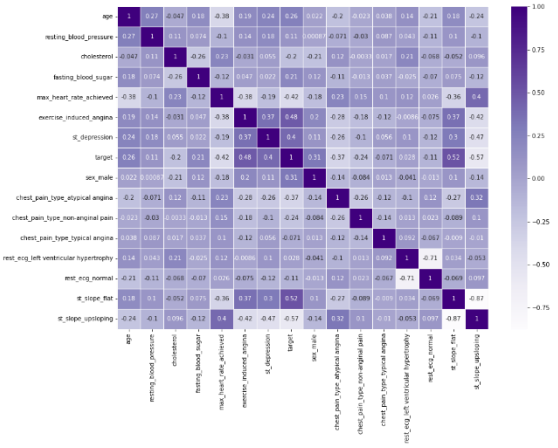

Fig. 6. Correlation heatmap between categories

VI. *PREPROCESSING*

A. Test-Train split

In the previous section, we saw how the various factors are gathered and how much few factors can impact the possibility of a person having heart disease. Each of these factors are important in determining our final results. Hence, we move forward with our project where we separate the

target variable from the rest of the table so that we can feed these values to the algorithm for it to train and test itself. We will name the target table to be "y" and the rest of the table left after splitting will be called as "x".

And finally, we further split our data into two parts, one for training and the other part for testing. After a few runs, trials and errors, we came to the conclusion that splitting the data to keep 20% of it for testing after 80% of the data is used for training the algorithm. We do all this with the help of sklearn library with a predefined set of functions. Now, we have our data in four parts, X_train, X_test, Y_train and Y_test.

B. Feature Normalisation

Feature normalisation is an important technique through which we normalise the data, that is we scale the numeric data values in the dataset to use a common scale. It is done so that all the values carry equal weight and the model does not get skewed in one direction just because of a larger value.

C. Cross Validation

Here, we built different baseline models and perform a 10-fold cross validation to filer the best performing baseline models, which then can be used in stacked ensembled model. The model performs the fitting procedure 10 times in total, every time the fit being performed on a training set containing 90% of the total data points, while remaining is used for validation.

VII. MODEL BUILDING

Now we will be using all the algorithms mentioned one after the other and study the results we get after each of the classification algorithms. Each algorithm results are stored in a specifically assigned variable. The algorithms used are from sklearn library.

A. Soft voting

Soft voting is a powerful technique which is used to combine the results of different algorithms using weighted importance and summed up results. Individual algorithm is arranged in increasing order and then assigned weights accordingly. The algorithms and weights are combined and then finally the output of every algorithm is combined to create a desired output.

In this project, we have used the following algorithms:

- Random Forest
- Decision Tree
- XG Boost
- ExtraTrees Classifier
- Naïve Bayes

The results from these algorithms are then combined with the weights, which is assigned as the algorithm with best accuracy gets the highest weight.

VIII. *RESULTS*

Finally, after running the algorithms, we calculate the results. And we compare all the algorithms. There are various parameters used to compare the algorithms. They are as follows:

- *Accuracy*

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- *Precision*

$$Precision = \frac{TP}{TP + FP}$$

- *Specificity*

$$Specificity = \frac{TN}{TN + FP}$$

- *Sensitivity*

$$Sensitivity = \frac{TP}{TP + FN}$$

- *F-1 Score*

$$F - 1\ Score = \frac{2 \times precision \times recall}{precision + recall}$$

Below is the table containing all these parameters for different algorithms.

	Model	Accuracy	Precision	Sensitivity	Specificity	F1-Score
1	RF	0.893617	0.860294	0.951220	0.830357	0.903475
2	KNN	0.808511	0.786765	0.869919	0.741071	0.826255
3	LR	0.808511	0.782609	0.878049	0.732143	0.827586
4	CART	0.834043	0.833333	0.853659	0.812500	0.843373
5	NB	0.817021	0.807692	0.853659	0.776786	0.80040
6	XGB	0.906383	0.879699	0.951220	0.857143	0.914062
7	SVC	0.825532	0.801471	0.886179	0.758929	0.841699

Table 1. Results for individual algorithms

Here we notice that from the selected algorithms, XG Boost has the highest accuracy at 90.64% with a sensitivity of 0.951 and specificity of 0.857 and highest f1-score of 0.914. It also tops in the precision which comes out at around 88%. Random forest comes in at the second place with and accuracy just a shy higher that 89%. Third is decision tree classifier with an approximate accuracy of 83.4%. The other parameters can be read from the table above.

The confusion matrix showing TP, TN, FP and FN for XG Boost is as below:

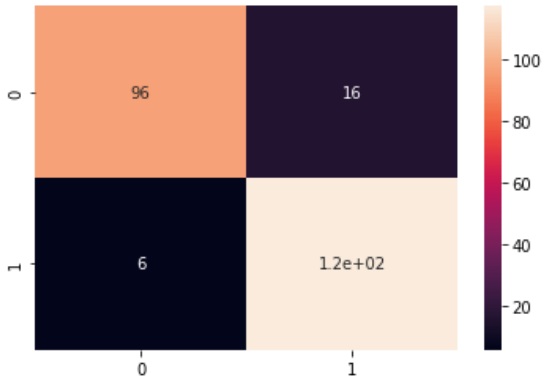

Fig. 7. Confusion matrix for XG Boost

And now, we apply the soft voting algorithm, combine the algorithms with their respective weights and gain a cumulative accuracy. The results are as follows:

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score
0	Soft Voting	0.902128	0.867647	0.95935	0.839286	0.911197

Fig. 8. Soft voting results

The confusion matrix for soft voting is given below

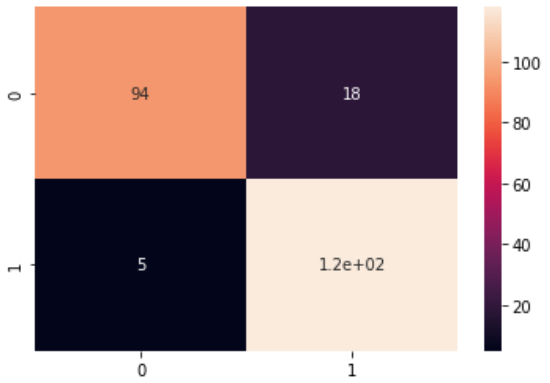

Fig. 9. Confusion matrix for soft voting

	Model	Accuracy	Precision	Sensitivity	Specificity	F1-Score
1	RF	0.893617	0.860294	0.951220	0.830357	0.903475
2	KNN	0.808511	0.786765	0.869919	0.741071	0.826255
3	LR	0.808511	0.782609	0.878049	0.732143	0.827586
4	CART	0.834043	0.833333	0.853659	0.812500	0.843373
5	NB	0.817021	0.807692	0.853659	0.776786	0.80040
6	XGB	0.906383	0.879699	0.951220	0.857143	0.914062
7	SVC	0.825532	0.801471	0.886179	0.758929	0.841699
8	SV	0.902128	0.867647	0.959350	0.839286	0.911197

Table 2. Results for individual algorithms and soft voting

Finally, we calculate the importance and relevance of the attributes, with respect to our model. This study will reveal which features are more important than the others and hence, others can be weighted less as per their relevance. Below is the bar chart showing the importance where higher value indicate a higher importance.

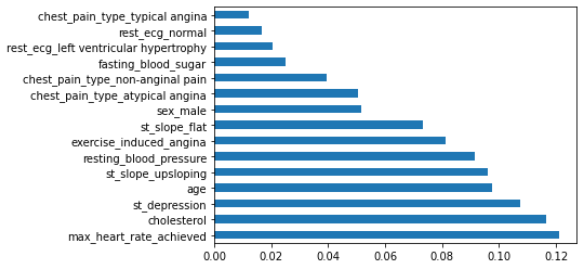

Fig. 10. Attribute importance plot

The top 5 most contribution features are as below, in the same order of ranking. This fact can also be studied through the correlation chart given above.

1. Max heart Rate achieved
2. Cholestrol
3. st_depression
4. Age
5. exercise_induced_angina

IX. *CONCLUSIONS*

There is an urgency in controlling the rising heart and allied disease problem all over the globe, especially in developed and developing countries and because of that, real-time, quick, cheap and easy detection techniques and applications are extremely valuable and important. The authors have tried to demonstrate a working model for a fairly accurate and real time heart disease prediction application using various algorithms and tools available online. After studying various other published work about the same situation, the authors reviewed the existing work and technology within the same domain. The dataset was gathered and it was modelled according to the needs and requirements, before being made sure that there are no inaccuracies or defects in the dataset. Then the model was constructed and trained using the dataset. The results were analysed and studied on wide variety of factors like sensitivity, precision, accuracy, F1 score, specificity and the model exhibiting very promising results.

There is still a huge gap for future development in the very same technology and further room for improvement. For the future upgrades, this model could be made accessible on various platforms like mobile phone applications and web applications, which would only make it more accessible to general public, hence reducing the financial stress.

X. *REFRENCES*

[1]Heart disease deaths rise in India by 34% in 26 years | India News - Times of India (indiatimes.com)

[2]Cardiovascular disease in India: A 360 degree overview (nih.gov)

Dorairaj P., Panniyammakal J.and Ambuj R., Cardiovascular Diseases in India, 19 Apr 2016 https://doi.org/10.1161/CIRCULATIONAHA.114.008729Circulation. 2016;133:1605–1620

Enriko I. K. A., Muhammad, Suryanegara. and Dadang, Gunawan. (2016). Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters. Journal of Telecommunication, Electronic and Computer Engineering, 8(12), 59-65.

Boshra, Bahrami., Mirsaeid, Hosseini, Shirvani. (2015). Prediction and Diagnosis of Heart Disease by Data Mining Techniques. Journal of Multidisciplinary Engineering Science and Technology, 2(2), 164-168.

Janardhanan, P., L, Heena. and Sabika, F. (2015). Effectiveness of Support Vector Machines in Medical Data Mining. Journal of Communications Software and Systems, 11(1), 25-30.

Karayılan, T. and Kılıç, Ö. (2017). Prediction of heart disease using neural network. International Conference on Computer Science and Engineering (UBMK), Antalya, 719-723. doi: 10.1109/UBMK.2017.8093512.

Lavanya, M., Gomathi, P, M. (2016). Prediction of Heart Disease using Classification Algorithms. International Journal of Advanced Research in Computer Engineering & Technology, 5(7), 2173-2175.

A. H. Chen, S.-Y. Huang, P.-S. Hong, C.-H. Cheng, and E.-J. Lin, "Hdps: heart disease prediction system," In 2011 computing in Cardiology, vol. 557–560, 2011.View at: Google Scholar

A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heartdisease using machine learning algorithms," Mobile Information Systems, vol. 2018, Article ID 3860146, 21 pages, 2018.View at: Publisher Site | Google Scholar

Hidayet, Takci. (2018). Improvement of Heart Attack Prediction by the Feature Selection Methods. Turk J Elec Eng & Comp Sci, 26, 1-10

M. A. Abushariah, A. A. Alqudah, O. Y. Adwan et al., "Automatic heart disease diagnosis system based onartificial neural network (ann) and adaptive neuro-fuzzy inference systems (anfis) approaches," Journal of Software Engineering and Applications, vol. 7, p. 1055, 2014.View at: Publisher Site | Google Scholar

E. O. Olaniyi and O. K. Oyedotun, "Heart diseases diagnosis using neural networks arbitration," International Journal of Intelligent Systems and Applications, vol. 7, no. 12, pp. 75–82, 2015.View at: Publisher Site | Google Scholar

O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," Expert Systems with Applications, vol. 68, pp. 163–172, 2017.

wikipedia.org

geeksforgeeks.org

w3schools.com