HEART DISEASE PREDICTION

A

MAJOR PROJECT REPORT

Submitted by

Manak
(06114802718)

Manaswi
(06214802718)

Pankaj Kumar
(07414802718)

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Under the guidance

of

Ms. Garima Gupta

Department of Computer Science and Engineering

Maharaja Agrasen Institute of Technology,
PSP Area Sector - 22, Rohini, New Delhi - 110086
(Affiliated to Guru Gobind Sing Inderprastha, New Delhi)

MAHARAJA AGRASEN INSTITUTE OF TECHNOLOGY

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that this Major project report “**HEART DISEASE PREDICTION**” is submitted by MANAK (06114802718), MANASWI (06214802718), PANKAJ (07414802718) who carried out the project work under my supervision.

I approve this Major project for submission.

Ms. Garima Gupta

(Associate Professor, CSE)

(Project Guide)

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to human beings given how our lives and schedules are evolving into more sedentary ones with the advent of technologies originally made to make our lives easier.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, and obesity, and increase the risks of colon cancer, high blood pressure, depression and anxiety.

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

ACKNOWLEDGEMENT

It gives us immense pleasure to express our deepest sense of gratitude and sincere thanks to our respected guide, Ms. Garima Gupta, Assistant Professor, CSE, MAIT, Delhi, for their valuable guidance, encouragement and help for completing this work. Their useful suggestions for this whole work and co-operative behaviour are sincerely acknowledged.

We also wish to express our indebtedness to our parents as well as our families whose blessings and support always helped us to face the challenges ahead.

Place: Delhi

Date:

Manak
(06114802718)

Manaswi
(06214802718)

Pankaj Kumar
(07414802718)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

SYMBOL	**MEANING**
CP	Chest pain
RESTBPS	Resting blood pressure
CHOL	Cholesterol level
FBS	Fasting blood sugar
RESTECG	Resting ECG
THALACH	Maximum heart rate
EXANG	Exercise induced angina
OLDPEAK	Depression in ST induced by exercise relative to rest
TP	True positives
TN	True negatives
FP	False positives
FN	False negatives
KNN	K-Nearest Neighbour
SVM	Support Vector Machine
CVD	Cardio Vascular Diseases

CHAPTER 1: INTRODUCTION

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. There could be other reasons behind a threatening heart disease as well including but not limiting to smoking, family history, high cholesterol, overwhelming stress and bad diets. 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, and obesity, and increase the risks of colon cancer, high blood pressure, depression and anxiety. 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 have a serious threat to their life.

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 [3]. 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 project.

In this research, 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 the following algorithms

1. Random Forest Classifier (criterion = 'entropy')
2. K nearest neighbour (n=9)
3. Logistic Regression
4. Decision Tree Classifier
5. Naïve Bayes
6. XGBoost (n_estimators=500)
7. Support Vector Classifier (kernel='linear')
8. Soft Voting

In the end, we compare the outcomes and numbers from all these algorithms to find out the most suitable 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.

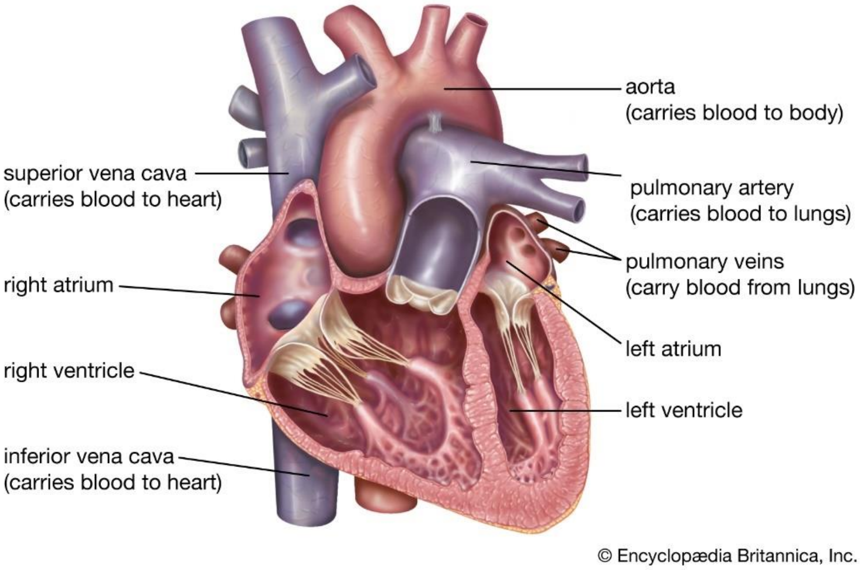


Figure 1.1: Cross sectional diagram of heart

1.1 SCOPE

It is impractical for a common man to frequently undergo costly tests like the ECG and thus there needs to be a system in place which is handy and at the same time reliable, in predicting the chances of a heart disease. Thus, we propose to develop an application which can predict the vulnerability of a heart disease given basic symptoms like age, sex, pulse rate etc. Some techniques have proven to be the most accurate and reliable algorithm and hence used in the proposed system.

In the epidemic of heart diseases, our proposed algorithm can be useful as the medical attributes required for this diagnosis can easily be measured at home. Therefore, the physical availability of the patient is not required

CHAPTER 2: LITERATURE SURVEY

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.

CHAPTER 3: RESEARCH APPROACH

3.1 Tools and Technologies

The experiment has been performed using Intel (R) Core i7 CPU and 8 GB of RAM.

3.1.1 Python

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985- 1990. Like Perl, Python source code is also available under the GNU General Public License (GPL). This tutorial gives enough understanding on Python programming language. Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

Python is a MUST for students and working professionals to become a great Software Engineer specially when they are working in Web Development Domain. It supports functional and structured programming methods as well as OOP. It can be used as a scripting language or can be compiled to byte-code for building large applications. It provides very high-level dynamic data types and supports dynamic type checking. It supports automatic garbage collection. It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

Python is:

- Easy-to-learn
- Easy-to-read
- Easy-to-maintain
- A broad standard library
- Interactive Mode
- Portable
- Extendable
- GUI Programming
- Scalable

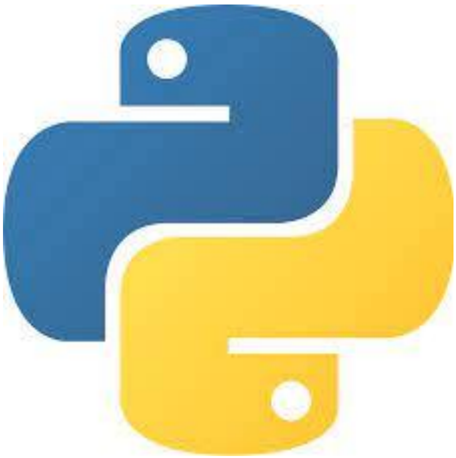


Figure 3.1: Python Logo

3.1.2 Jupyter Notebook

Project Jupyter is a project and community whose goal is to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages". It was spun off from IPython in 2014 by Fernando Pérez and Brian Granger. Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are Julia, Python and R, and also a homage to Galileo's notebooks recording the discovery of the moons of Jupiter. Project Jupyter has developed and supported the interactive computing products Jupyter Notebook, JupyterHub, and JupyterLab. Jupyter is financially sponsored by NumFOCUS.

In 2014, Fernando Pérez announced a spin-off project from IPython called Project Jupyter. IPython continues to exist as a Python shell and a kernel for Jupyter, while the notebook and other language-agnostic parts of IPython moved under the Jupyter name. Jupyter is language agnostic and it supports execution environments (aka kernels) in several dozen languages among which are Julia, R, Haskell, Ruby, and of course Python (via the IPython kernel). In 2015, GitHub and the Jupyter Project announced native rendering of Jupyter notebooks file format (.ipynb files) on the GitHub platform.

3.1.3 Numpy

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices.

NumPy was created in 2005 by Travis Oliphant. It is an open-source project and you can use it freely. NumPy stands for Numerical Python. In Python we have lists that serve the purpose of arrays, but they are slow to process. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists. The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy.

Arrays are very frequently used in data science, where speed and resources are very important.

NumPy arrays are stored at one continuous place in memory unlike lists, so processes can access and manipulate them very efficiently. This behavior is called locality of reference in computer science. This is the main reason why NumPy is faster than lists. Also, it is optimized to work with latest CPU architectures. NumPy is a Python library and is written partially in Python, but most of the parts that require fast computation are written in C or C++.

3.1.4 Pandas

Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

Advantages

- Fast and efficient for manipulating and analyzing data.
- Data from different file objects can be loaded.
- Easy handling of missing data (represented as NaN) in floating point as well as non-floating-point data
- Size mutability: columns can be inserted and deleted from DataFrame and higher dimensional objects
- Data set merging and joining.
- Flexible reshaping and pivoting of data sets
- Provides time-series functionality.
- Powerful group by functionality for performing split-apply-combine operations on data sets.

3.1.5 Matplotlib

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Matplotlib comes with a wide variety of plots. Plots helps to understand trends, patterns, and to make correlations. They’re typically instruments for reasoning about quantitative information.

3.1.6 Seaborn

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and colour palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas. Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset.

Plots are basically used for visualizing the relationship between variables. Those variables can be either be completely numerical or a category like a group, class or division. Seaborn divides plot into the below categories –

- Relational plots: This plot is used to understand the relation between two variables.
- Categorical plots: This plot deals with categorical variables and how they can be visualized.
- Distribution plots: This plot is used for examining univariate and bivariate distributions
- Regression plots: The regression plots in seaborn are primarily intended to add a visual guide that helps to emphasize patterns in a dataset during exploratory data analyses.
- Matrix plots: A matrix plot is an array of scatterplots.

3.1.7 Sklearn

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Scikit-learn is a NumFOCUS fiscally sponsored project.

The scikit-learn project started as scikits.learn, a Google Summer of Code project by French data scientist David Cournapeau. Its name stems from the notion that it is a "SciKit" (SciPy Toolkit), a separately-developed and distributed third-party extension to SciPy. The original codebase was later rewritten by other developers. In 2010 Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort and Vincent Michel, all from the French Institute for Research in Computer Science and Automation in Rocquencourt, France, took leadership of the project and made the first public release on February the 1st 2010. Of the various scikits, scikit-learn as well as scikit-image were described as "well-maintained and popular" in November 2012. Scikit-learn is one of the most popular machine learning libraries on GitHub.

Scikit-learn is largely written in Python, and uses NumPy extensively for high-performance linear algebra and array operations. Furthermore, some core algorithms are written in Cython to improve performance. Support vector machines are implemented by a Cython wrapper around LIBSVM; logistic regression and linear support vector machines by a similar wrapper around LIBLINEAR. In such cases, extending these methods with Python may not be possible. Scikit-learn integrates well with many other Python libraries, such as Matplotlib and plotly for plotting, NumPy for array vectorization, Pandas dataframes, SciPy, and many more.

3.1.8 Keras

Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library.

Up until version 2.3, Keras supported multiple backends, including TensorFlow, Microsoft Cognitive Toolkit, Theano, and PlaidML. As of version 2.4, only TensorFlow is supported. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible. It was developed as part of the research effort of project ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System), and its primary author and maintainer is François Chollet, a Google engineer. Chollet is also the author of the XCeption deep neural network model.

Keras contains numerous implementations of commonly used neural-network building blocks such as layers, objectives, activation functions, optimizers, and a host of tools to make working with image and text data easier to simplify the coding necessary for writing deep neural network code. The code is hosted on GitHub, and community support forums include the GitHub issues page, and a Slack channel.

Keras allows users to productize deep models on smartphones (iOS and Android), on the web, or on the Java Virtual Machine. It also allows use of distributed training of deep-learning models on clusters of Graphics processing units (GPU) and tensor processing units (TPU).

3.1.8 Outlier detection

Outlier detection is a key consideration within the development and deployment of machine learning algorithms. Models are often developed and leveraged to perform outlier detection for different organisations that rely on large datasets to function.

3.1.9 Soft Voting:

In soft voting, every individual classifier provides a probability value that a specific data point belongs to a particular target class. The predictions are weighted by the classifier's importance and summed up. Then the target label with the greatest sum of weighted probabilities wins the vote.

3.1.10 Feature Importance:

Feature Importance refers to techniques that calculate a score for all the input features for a given model — the scores simply represent the “importance” of each feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable.

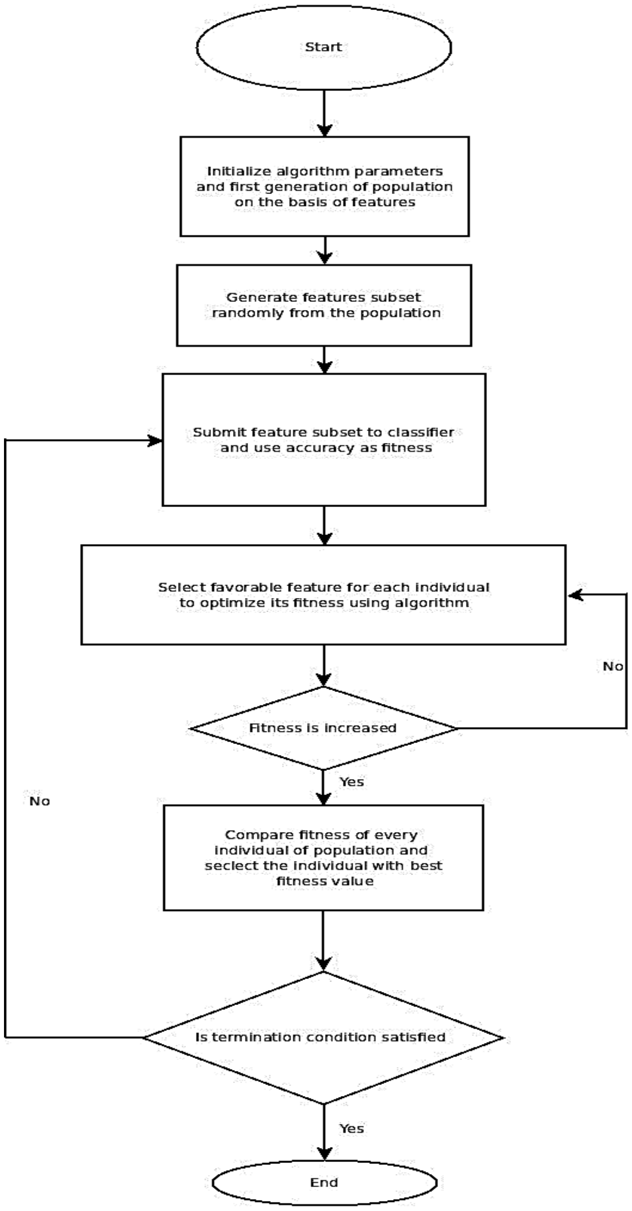

Start
Initialize algorithm parameters and first generation of population on the basis of features
Generate features subset randomly from the population
Submit feature subset to classifier and use accuracy as fitness
Select favorable feature for each individual to optimize its fitness using algorithm
No
Fitness is increased
Yes
No
Compare fitness of every individual of population and seclect the individual with best fitness value
Is termination condition satisfied
Yes
End

Fig 3.2: The general flow process

3.2 The attributes

1. Age: Patients Age in years (Numeric)
2. Sex: Gender of patient (Male - 1, Female - 0) (Nominal)
3. Chest Pain Type: Type of chest pain experienced by patient categorized into 1 typical, 2 typical angina, 3 non- anginal pain, 4 asymptomatic (Nominal)
4. resting bp s: Level of blood pressure at resting mode in mm/HG (Numerical)
5. cholestrol: Serum cholestrol in mg/dl (Numeric)
6. fasting blood sugar: Blood sugar levels on fasting > 120 mg/dl represents as 1 in case of true and 0 as false (Nominal)
7. resting ecg: Result of electrocardiogram while at rest are represented in 3 distinct values 0 : Normal 1: Abnormality in ST-T wave 2: Left ventricular hypertrophy (Nominal)
8. max heart rate: Maximum heart rate achieved (Numeric)
9. exercise angina: Angina induced by exercise 0 depicting NO 1 depicting Yes (Nominal)
10. oldpeak: Exercise induced ST-depression in comparison with the state of rest (Numeric)
11. ST slope: ST segment measured in terms of slope during peak exercise 0: Normal 1: Upsloping 2: Flat 3: Downsloping (Nominal)

Target variable

12. target: It is the target variable which we have to predict 1 means patient is suffering from heart risk and 0 means patient is normal.

3.3 Dataset

The dataset used in this project was gathered from kaggle.com. It is a combination of multiple datasets combined to form one to make a bigger dataset so as to give more training points to the algorithm. In this section, we discuss about the dataset used and some statistics of our dataset.

3.3.1 Shape and sample from the dataset

In this we talk about the size of the dataset used. We have attached screenshots from the Jupyter notebook, depicting the various statistics and information about the dataset. This information is very important and absolutely crucial since we should have and understanding of our dataset before we put it to some use. Datasets are fundamental to foster the development of several computational fields, giving scope, robustness, and confidence to results. And not only that, but a Good dataset allows for efficient analysis, limits errors and inaccuracies that can occur to data during processing, and makes all processed data more accessible to users. It's also gotten easier with new tools that enable any user to cleanse and qualify data on their own.

3.3.2 Correlation of the attributes

Now, we discuss about the correlation of the different attributes of our dataset and we use seaborn library to depict the heatmap which displays the correlations between attributes using the different shades of colours. Here, the darker shade implies that the two attributes are more related to each other as compared to the other attributes. All the values in this map are between 0 and 1. And diagonally, we see that correlation is '1' because an attribute has the highest relation to itself, hence we see one and the darkest shade in that box.

3.3.3 Exploratory data analysis

In the final subsection of this section, we study our datasets. We will talk about how they vary and how they look plotted against our "target" variable. We will keep the target value separated from our rest of dataset so we can study different trends in the numbers and make a general assumption about how different values can affect our target variable.

1. Target

   This variable shows whether the person in our dataset has a heart disease or not. This is the most crucial attribute as it helps in training and testing our algorithm.int the below countplot made using matplotlib, we see the distribution of people suffering from heart diseases and those who do not. We notice how both number of people are almost equal. This fact makes sure there is no bias and our algorithm gives out the as accurate results as possible.

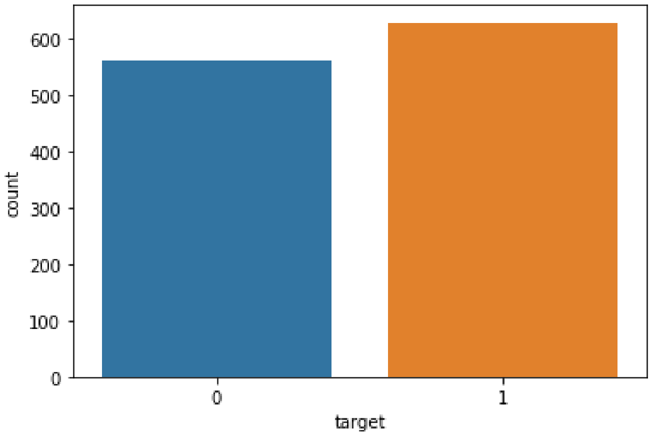


Fig. 3.3 Target distribution

2. Age

   Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately a tripling of risk with each decade of life. Coronary fatty streaks can begin to form in adolescence. It is estimated that 82 percent of people who die of coronary heart disease are 65 and older. Simultaneously, the risk of stroke doubles every decade after age 55.

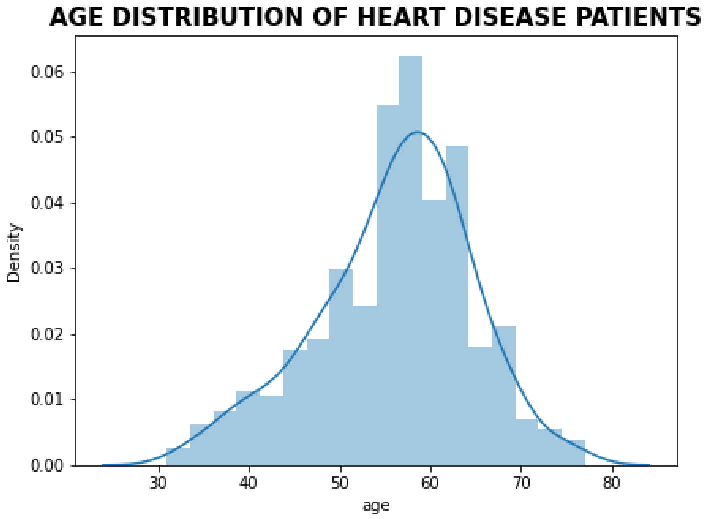


Fig. 3.4 Age distribution of heart disease patients

3. Sex

   In this plot, we observe how more men in our dataset suffers from a heart disease as compared to women. It is also an accepted fact that men are more likely to get a heart disease as compared to women.

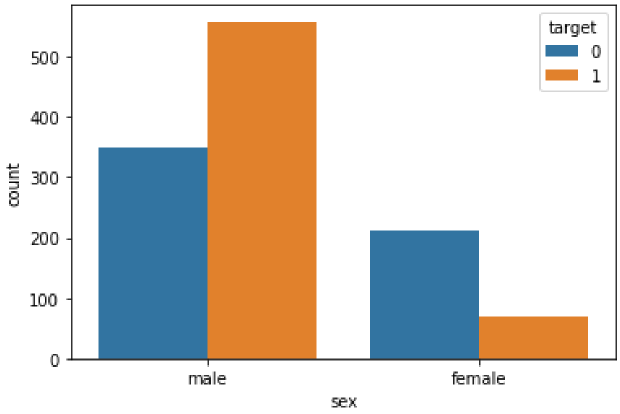


Fig. 3.5 Gender distribution wrt target

4. Chest pain

   Angina is chest pain or discomfort caused when your heart muscle doesn't get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest. The discomfort also can occur in your shoulders, arms, neck, jaw, or back. Angina pain may even feel like indigestion. In our dataset, we have divided this category into four parts. (0) asymptomatic, (2) atypical angina, (3) non-anginal pain and (4) typical angina

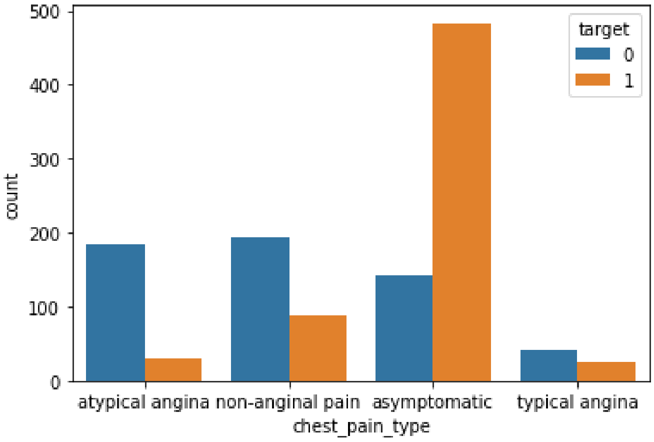


Fig 3.6 Chest pain distribution wrt target

5. Fasting blood sugar

   Not producing enough of a hormone secreted by your pancreas (insulin) or not responding to insulin properly causes your body's blood sugar levels to rise, increasing your risk of a heart attack. In our dataset, if blood sugar level is above 120ml/dl, it is represented by 1, otherwise by 0.

6. Resting ECG

For people at low risk of cardiovascular disease, the USPSTF concludes with moderate certainty that the potential harms of screening with resting or exercise ECG equal or exceed the potential benefits. For people at intermediate to high risk, current evidence is insufficient to assess the balance of benefits and harms of screening.

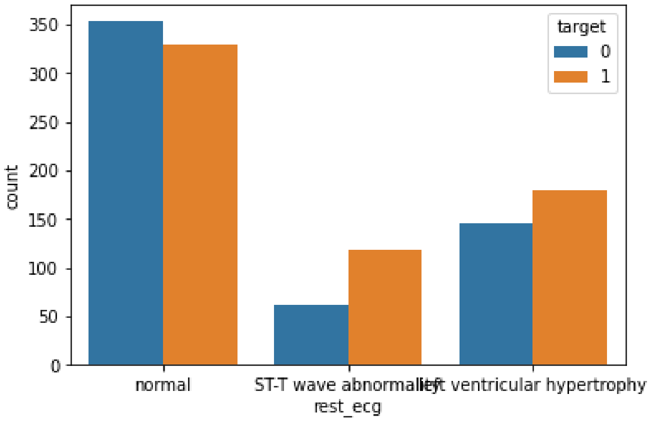


Fig. 3.7 Rest ECG distribution wrt target

7. Exercise induced Angina

   The pain or discomfort associated with angina usually feels tight, gripping or squeezing, and can vary from mild to severe. Angina is usually felt in the center of your chest but may spread to either or both of your shoulders, or your back, neck, jaw or arm. It can even be felt in your hands. o Types of Angina a. Stable Angina / Angina Pectoris b. Unstable Angina c. Variant (Prinzmetal) Angina d. Microvascular Angina.

8. Slope of the peak exercise ST segment

   A treadmill ECG stress test is considered abnormal when there is a horizontal or down-sloping ST-segment depression ≥ 1 mm at 60–80 ms after the J point. Exercise ECGs with up-sloping ST-segment depressions are typically reported as an ‘equivocal’ test. In general, the occurrence of horizontal or down-sloping ST-segment depression at a lower workload (calculated in METs) or heart rate indicates a worse prognosis and higher likelihood of multi-vessel disease. The duration of ST-segment depression is also important, as prolonged recovery after peak stress is consistent with a positive treadmill ECG stress test. Another finding that is highly indicative of significant CAD is the occurrence of ST-segment elevation > 1 mm (often suggesting transmural ischemia); these patients are frequently referred urgently for coronary angiography.

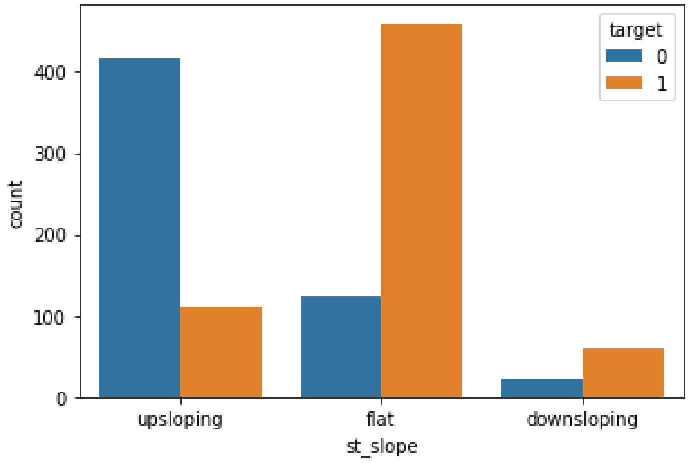


Fig. 3.8 ST slope distribution wrt target

9. Maximum heart rate achieved

   The increase in cardiovascular risk, associated with the acceleration of heart rate, was comparable to the increase in risk observed with high blood pressure. It has been shown that an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%, and this increase in the risk is similar to the one observed with an increase in systolic blood pressure by 10 mm Hg.

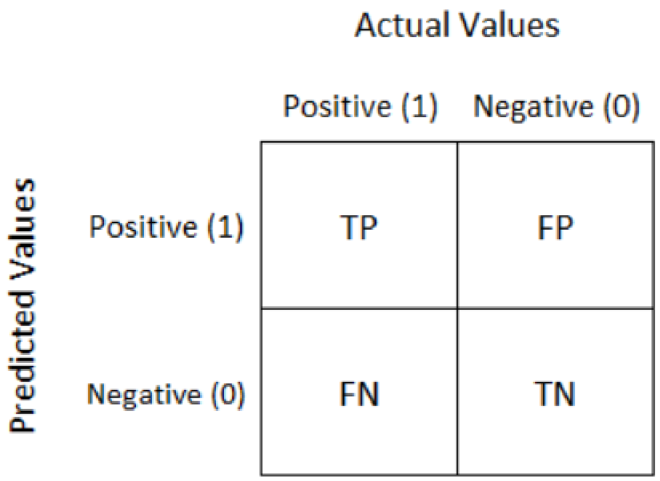
3.4 Classification

Now, we will use the following classification algorithms and feed them our dataset to train the algorithms and study the outcomes of each algorithm.

1. Random Forest Classifier (criterion = 'entropy')
2. K nearest neighbour (n=9)
3. Logistic Regression
4. Decision Tree Classifier
5. Naïve Bayes
6. XGBoost (n_estimators=500)
7. Support Vector Classifier (kernel='linear')
8. Soft Voting

But before that, we drop the “target” attribute from our data and rename the target as “Y” and the rest of data as “X”. We do that to help the training and testing process of the algorithms easier. After that, we split the data into training and testing sets, which finally divides the data into four parts, namely, X_train, X_test, Y_train and Y_test. This is done with the help of sklearn library method called “train_test_split”. Then we will use multiple parameters to judge a model. These parameters are listed as below:

1. Confusion matrix



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig 3.9: Confusion matrix Sample

Classification is the process of categorizing a given set of data into classes. In Machine Learning (ML), you frame the problem, collect and clean the data, add some necessary feature variables (if any), train the model, measure its performance, improve it by using some cost function, and then it is ready to deploy. But how do we measure its performance? Is there any particular feature to look at? A trivial and broad answer would be to compare the actual values to the predicted values. But that does not solve the issue.

2. Accuracy

Accuracy is the ratio of the True predicted values to the Total predicted values.

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

3. Precision

The precision for class 1 is, out of all predicted class values like 1, how many actually belong to class 1.

$$\text{Precision} = \frac{\text{TP}}{(TP + FP)}$$

4. Recall

   Recall for class 1 is, out of all the values that actually belong to class 1, how much is predicted as class 1.

$$\text{Recall} = \frac{\text{TP}}{(TP + FN)}$$

5. F1

   To have a combined effect of precision and recall, we use the F1 score. The F1 score is the harmonic mean of precision and recall.

$$\text{F1 Score} = \frac{2}{({}^{1}/_{Precision} + {}^{1}/_{Recall})}$$

3.5 SOFTWARE AND HARDWARE REQUIREMENTS

Hardware Requirements:

- Intel i3 Generation 3 Processor or equivalent or higher
- 4 GB RAM or higher
- 20 GB HDD space or higher

Software Requirements:

- Jupyter Notebook
- Python

3.6 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. We have used a combination of multiple data sets available online to compile one bigger document with more points. Our final data contains about 1189 rows and 13 columns. We will discuss about the columns, the factors impacting the probability of heart disease in a person.

1. Age: Patients Age in years (Numeric)
2. Sex: Gender of patient (Male - 1, Female - 0) (Nominal)
3. Chest Pain Type: Type of chest pain experienced by patient categorized into 1 typical, 2 typical angina, 3 non- anginal pain, 4 asymptomatic (Nominal)
4. resting bp s: Level of blood pressure at resting mode in mm/HG (Numerical)
5. cholestrol: Serum cholestrol in mg/dl (Numeric)
6. fasting blood sugar: Blood sugar levels on fasting > 120 mg/dl represents as 1 in case of true and 0 as false (Nominal)
7. resting ecg: Result of electrocardiogram while at rest are represented in 3 distinct values 0 : Normal 1: Abnormality in ST-T wave 2: Left ventricular hypertrophy (Nominal)
8. max heart rate: Maximum heart rate achieved (Numeric)
9. exercise angina: Angina induced by exercise 0 depicting NO 1 depicting Yes (Nominal)
10. oldpeak: Exercise induced ST-depression in comparison with the state of rest (Numeric)
11. ST slope: ST segment measured in terms of slope during peak exercise 0: Normal 1: Upsloping 2: Flat 3: Downsloping (Nominal)

Target variable

12. target: It is the target variable which we have to predict 1 means patient is suffering from heart risk and 0 means patient is normal.

3.6.1 Working

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.

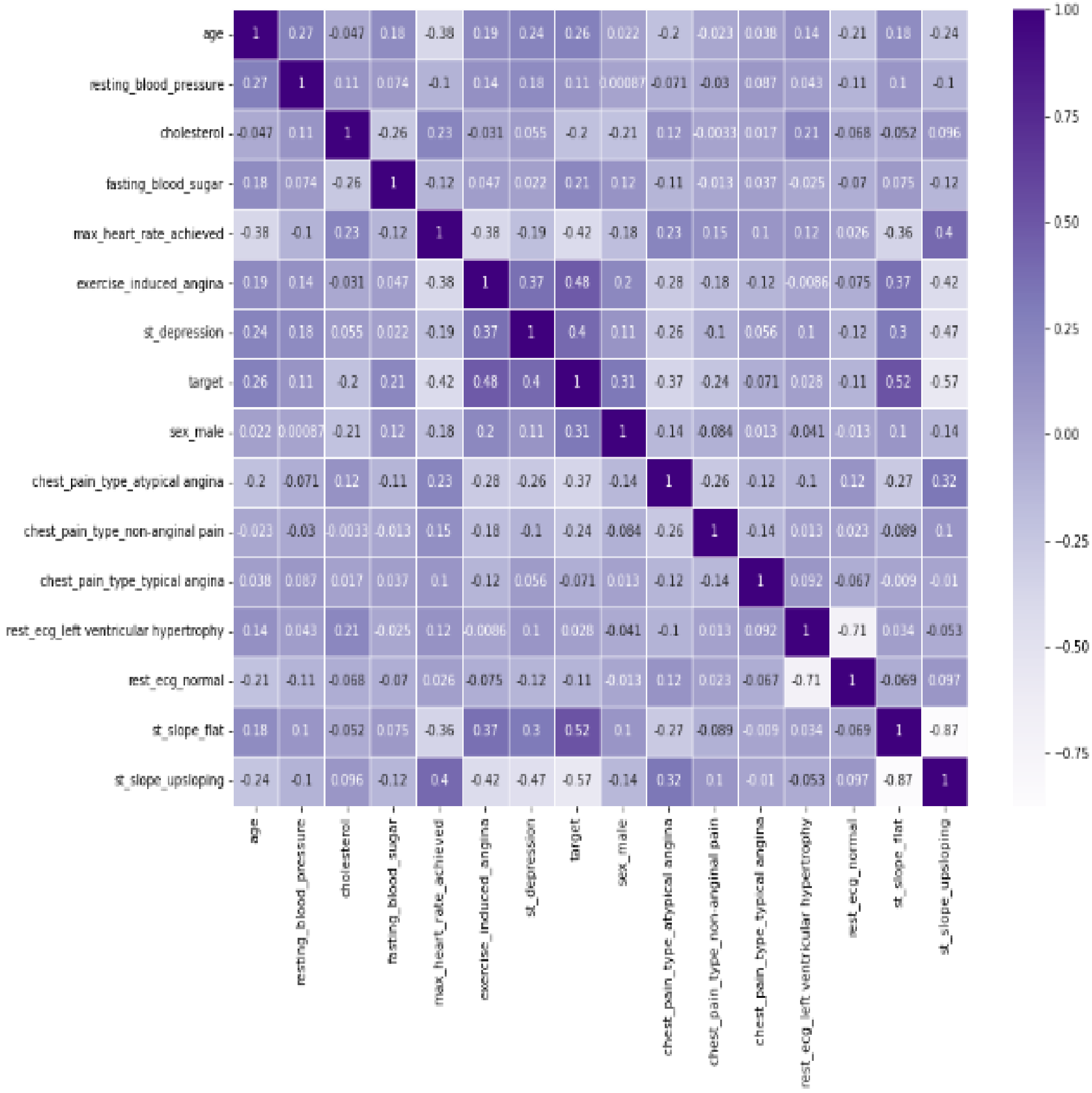


Fig. 3.10 Correlation Heatmap

3.6.2 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. It would perform the fitting procedure a total of 10 times, with each fit being performed on a training set containing 90% of the total, while remaining is used for validation.

3.6.3 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 very powerful technique which is used to combine the results of different algorithms using weighted importance and the summed up. Then the target label with the greatest sum of weighted probabilities, wins the vote.

In this project, we have used the following algorithms:

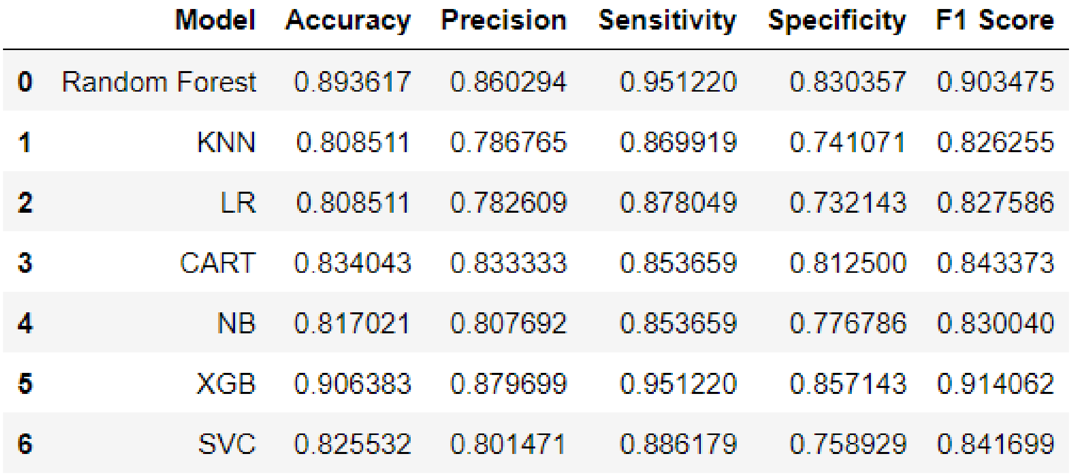
- Random Forest
- Decision Tree
- XG Boost
- ExtraTrees Classifier
- Naïve Bayes

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

CHAPTER 4: RESULTS

On running all the following algorithms, the given results were obtained.

1. Random Forest Classifier (criterion = 'entropy')
2. K nearest neighbour (n=9)
3. Logistic Regression
4. Decision Tree Classifier
5. Naïve Bayes
6. XGBoost (n_estimators=500)
7. Support Vector Classifier (kernel='linear')
8. Soft Voting



	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score
0	Random Forest	0.893617	0.860294	0.951220	0.830357	0.903475
1	KNN	0.808511	0.786765	0.869919	0.741071	0.826255
2	LR	0.808511	0.782609	0.878049	0.732143	0.827586
3	CART	0.834043	0.833333	0.853659	0.812500	0.843373
4	NB	0.817021	0.807692	0.853659	0.776786	0.830040
5	XGB	0.906383	0.879699	0.951220	0.857143	0.914062
6	SVC	0.825532	0.801471	0.886179	0.758929	0.841699

Fig 4.1: Results for classification 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:

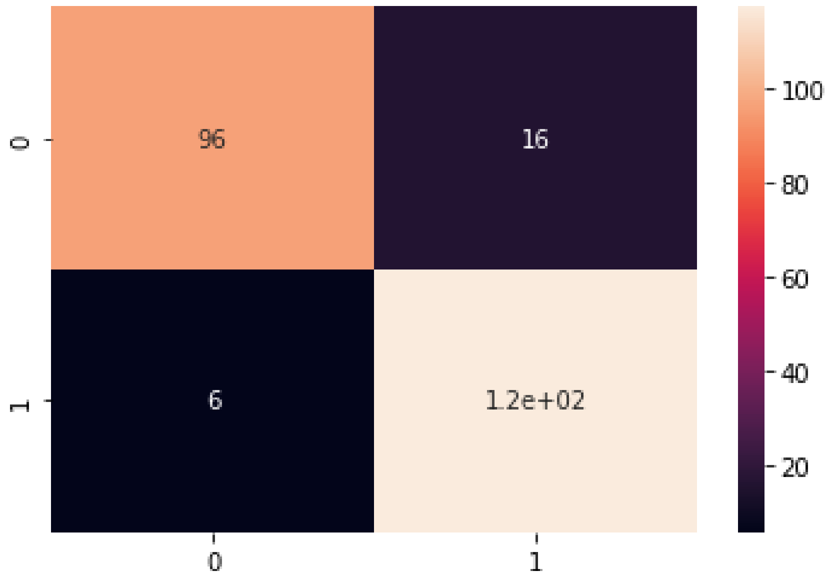
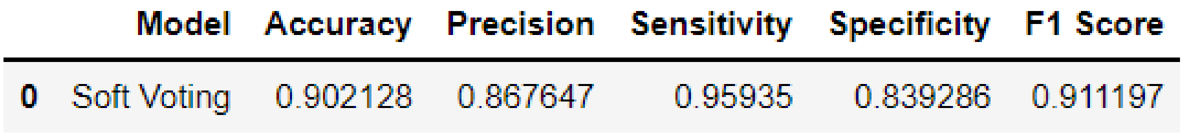


Fig 4.2: Confusion matrix for XG Boost

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

The confusion matrix for soft voting is given below

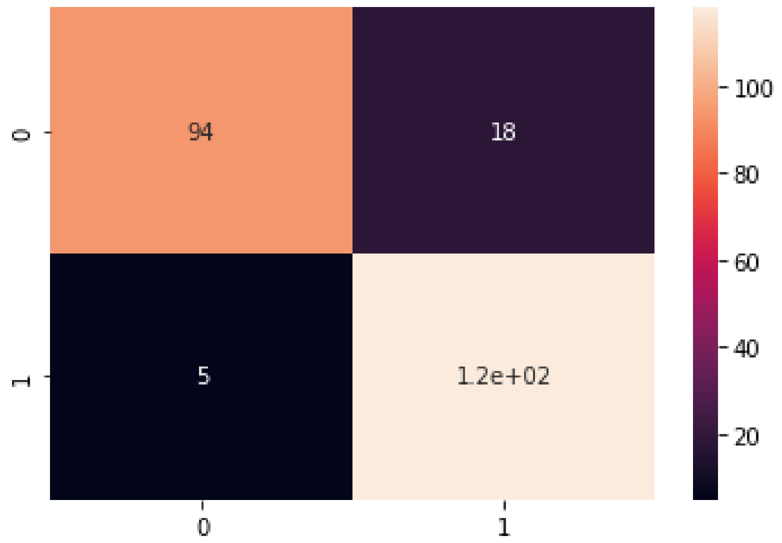
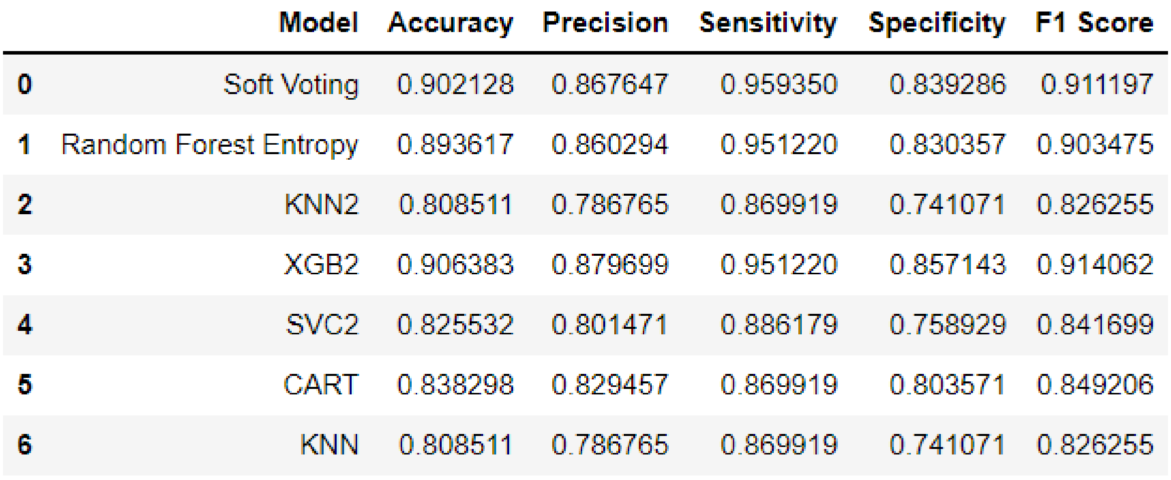


Fig 4.3: Confusion matrix for soft voting



	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score
0	Soft Voting	0.902128	0.867647	0.959350	0.839286	0.911197
1	Random Forest Entropy	0.893617	0.860294	0.951220	0.830357	0.903475
2	KNN2	0.808511	0.786765	0.869919	0.741071	0.826255
3	XGB2	0.906383	0.879699	0.951220	0.857143	0.914062
4	SVC2	0.825532	0.801471	0.886179	0.758929	0.841699
5	CART	0.838298	0.829457	0.869919	0.803571	0.849206
6	KNN	0.808511	0.786765	0.869919	0.741071	0.826255

Fig 4.4: Comparative analysis of results against 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.

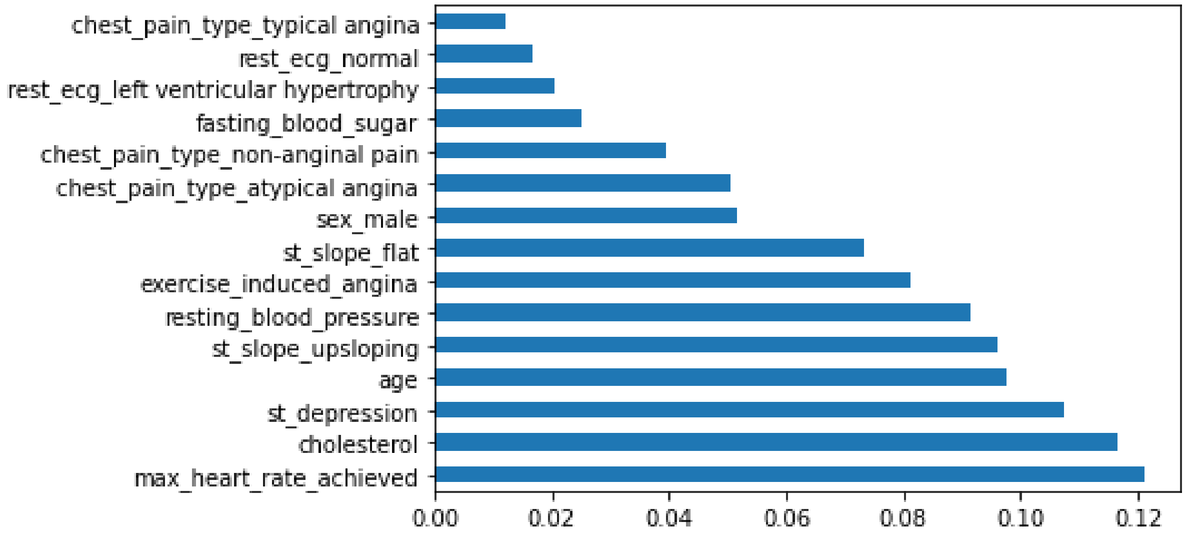


Fig 4.5: Feature importance plot

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

CHAPTER 5: CONCLUSION AND FUTURE SCOPE

As we have seen, stacked ensemble of power machine learning algorithms resulted in lower performance than an individual machine learning model. Hence, we see a disadvantage of soft voting in action.

We have also interpreted second best performing algo i.e., XG Boost

The top 5 most contribution features are:

Max heart Rate achieved

Cholestrol

st_depression

Age

exercise_induced_angina

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.

CHAPTER 6: REFERENCES

[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)

[3] 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

[4] Mostafa A, Xiao-Yan G, Amin Ali, Hassan S, Eman M., (2021), Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method, 6663455, 2021/02/10, 1076-2787, https://doi.org/10.1155/2021/6663455, Hindawi

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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

[11] 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

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

[13] 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

[14] 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

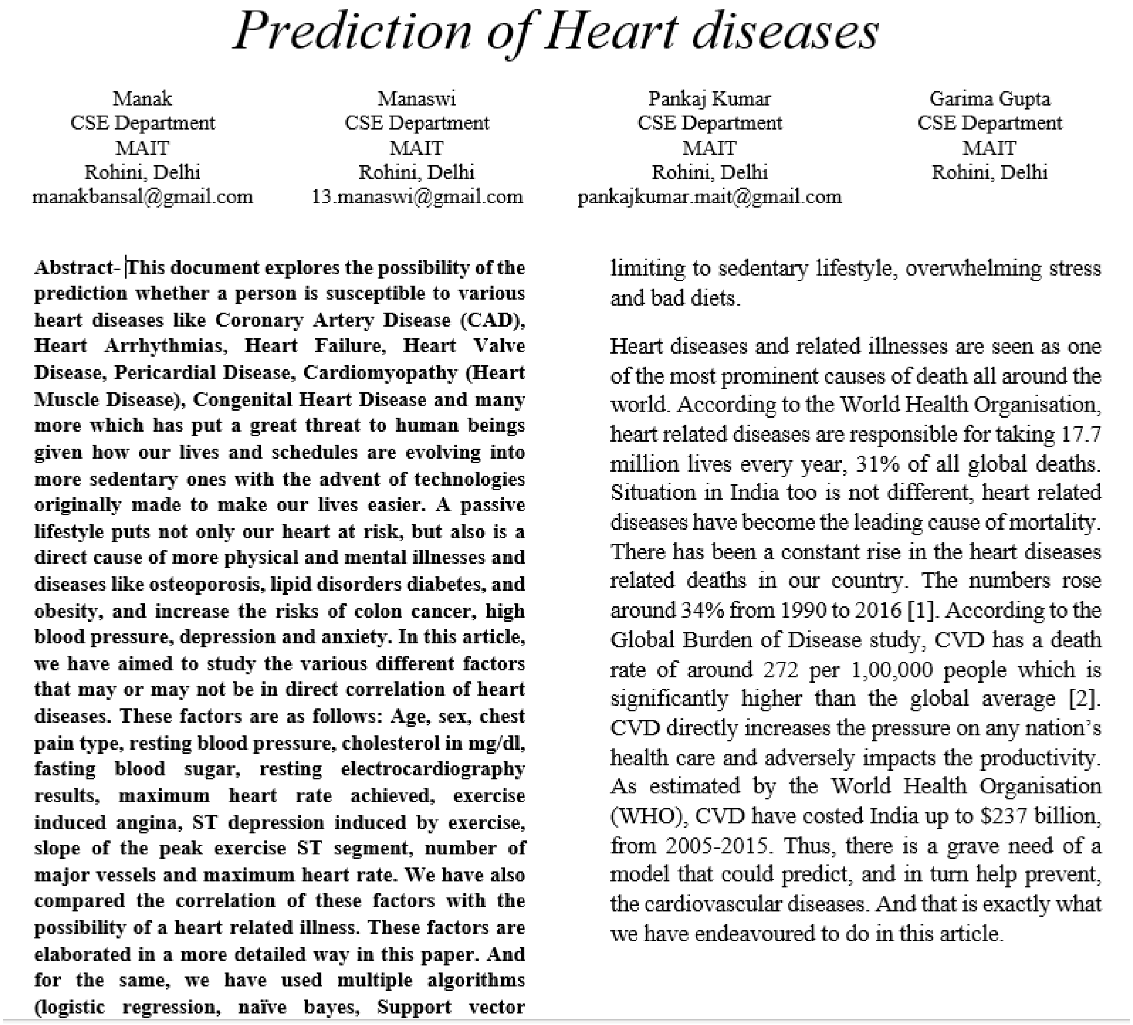
[15] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, “An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction,” Expert Systems with Applications, vol. 68, pp. 163–172, 2017.

[16] wikipedia.org

[17] geeksforgeeks.org

[18] w3schools.com

GLIMPSE OF OUR RESEARCH PAPER



Prediction of Heart diseases

Manak
CSE Department
MAIT
Rohini, Delhi
manakbansal@gmail.com

Manaswi
CSE Department
MAIT
Rohini, Delhi
13.manaswi@gmail.com

Pankaj Kumar
CSE Department
MAIT
Rohini, Delhi
pankajkumar.mait@gmail.com

Garima Gupta
CSE Department
MAIT
Rohini, Delhi

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, and obesity, and increase the risks of colon cancer, high blood pressure, depression and anxiety. 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

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.

Fig 18: Glimpse of Research Paper