PREDICTION OF HEART DISEASE USING

MACHINE LEARNING

A PROJECT REPORT

Submitted by

SUREHA S(171421601033)

AZARUDEEN A(171421601004)

MOHAMMED BASHEER D(171421601017)

Under the guidance of

Ms.KEERTHANA

In partial fulfillment for the award of the degree of

Bachelor of Computer Applications

JULY 2020

BONAFIDE CERTIFICATE

Certified that this project report "**PREDICTION OF HEART DISEASE USING MACHINE LEARNING"** is the bonafide work of **SUREHA S(171421601033),AZARUDEEN A(171421601004), MOHAMMED BASHEER D(171421601017)** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Ms.KEERTHANA

SUPERVISOR

Assistant professor

Department of Computer Application

B.S. Abdur Rahman Crescent Institute

of Science and Technology

Vandalur, Chennai – 600 048

SIGNATURE

Dr. GUFRAN AHMAD ANSARI

HEAD OF THE DEPARTMENT

Professor and Head

Department of Computer Application

B.S.Abdur Rahman Crescent Institute

of Science and Technology

Vandalur, Chennai – 600048.

B.S. Abdur Rahman Crescent Institute of Science & Technology
Deemed to be University u/s 3 of the UGC Act, 1956

VIVA-VOICE EXAMINATION

The viva-voice examination of the project work titled "**PREDICTION OF HEART DISEASE USING MACHINE LEARNING**" submitted by **SUREHA.S(171421601033), AZARUDEEN.A(171421601004), MOHAMMED BASHEER.D(171421601017)**is held on _____.

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

I thank the **Almighty** for showering His blessings upon me in completing the project. I submit this project with a deep sense of gratitude and reverence for my beloved parents for their moral support and encouragements.

I sincerely express my heartfelt gratitude to Prof. Dr. A. Peer Mohamed **Pro Vice Chancellor, B.S. Abdur Rahman Crescent University and** Prof. Dr.A. Azad**, Registrar, for furnishing every essential facility for doing my project.**

I owe my sincere gratitude to **Prof. Dr. Venkatesan Selvam.,** Dean of School of Computer, Information and Mathematical Science (SCIMS), **Dr. Gufran Ahmad Ansari,** Professor and Head, Department of Computer Applications for providing strong oversight of vision, strategic direction, encouragement and valuable suggestions in completing my project work.

I convey my earnest thanks to my project guide **Ms.keerthana.G,**Assistant professor**,**Department of Computer Applications, for his/her valuable guidance and support throughout the project.

I express my gratitude to the Project Coordinators and the project Committee Members of the Department of Computer Applications for their support and continued assistance in the process.

I extend my sincere thanks to all my faculty members for their valuable suggestions, timely advice and support to complete the project.

Mohammed basheer.D

Azarudeen.A

Sureha.S

ABSTRACT

Health care field usually has a vast amount of data and for processing those data certain techniques are used.Heart disease is the leading cause of death worldwide.This disease is quite common nowadays,We used different features which causes the heart disease to predict its presence.

Effective heart disease prediction system (EHDPS) is developed using machine learning algorithms. The dataset used are classified in terms of 15 medical parameters like chest pain type,cholesterol,maximum heart rate achieved.These parameters are considered as input to machine learning algorithms and used for prediction

The project focuses on implementation of the codes in python programming.Supervised machine learning algorithms such as k nearest neighbor and random forest are used for the prediction. The outcome of the system provides the presence of heart disease in terms of percentage i.e accuracy level.Visualizing software,tableau is used to visualize the patients with high risk of disease using major features causing the disease.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Machine learning is widely used nowadays in many business applications like ecommerce and manymore.Prediction is one of the area where machine learning is used.In this system,a heart disease dataset is used.The main aim of this system is to predict the possibilities of occurring heart disease of the patients in terms of percentage. This is performed through data mining classification techniques. The classification technique is used for classifying the entire dataset into two categories namely yes and No. Classification techniques applied to the dataset through the machine learning classification algorithm namely K-Nearest Neighbor Classification models and Random Forest Classification models. These models are used to enhance the accuracy level of the classification technique. This model performs both the classification and prediction methods. These models are performed using python Programming Language.Different outcomes of the project is visualized in tableau.

1.1 GENERAL

The project is to predict the percentage(i.e accuracy) by training the patients data and to compare which machine learning algorithm gives the best accuracy. Tableau is used to visualize the patients with high risk of the disease by analyzing the features causing the disease.

1.2. EXISTING SYSTEM:

The existing system performs a analysis technique through confusion matrix with calculation of percentages of the accuracy, precision, and recall. The method used was K-Nearest Neighbor (KNN), in which prediction on the trending topic was determined based on the membership distance of a class.

1.2.1 LITERATURE SURVEY

- **‘Efficient Heart Disease prediction system using decision tree’**

The author **Purushottam** in this paper discusses the performance of the system is evaluated in terms of classification accuracy using DECISION TREE

algorithm and the results shows that the system has great potential in predicting the heart disease risk level more accurately.

II ‘Prediction of Heart Disease Using Machine Learning Algorithms’

The author Mr.Santhana Krishnan in this paper discusses about the Classification technique that is applied to the dataset through the machine learning classification algorithm namely Decision tree and Naïve Bayes Classification models.These models are used to enhance the accuracy level of the classification technique. This model performs both the classification and prediction methods.

III ‘ Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning’

The author Reddy Prasad,Pidaparthi Anjali presents.In this paper prediction was done by building a model for logistic regression and naïve bayes. Here, they use sigmoid function which helps in the graphical representation of the classified data.By using logistic regression, naïve bayes the accuracy rate increases.

1.2.1 Disadvantages of the Existing System:

The following are the disadvantages of the existing system,

- Less accuracy was achieved
- Lacks visualization

1.3 PROPOSED SYSTEM:

Prediction of heart diease is done by trained data(i.e patients data) and the accuracy rate is found for two machine learning algorithms.Algorithm which gives the

best accuracy is found.Visualization is done to visualize the patients with high risk of disease using major features causing the disease..

1.3.1 Advantages of the Proposed System:

- Accuracy is found for two machine learning algorithms
- Higher accuracy is obtained
- Tableau is used to visualize different outcomes

1.4 ORGANISATION OF THE CHAPTER

The general introduction, existing system and proposed system for this project was discussed. The working and implementation of this system are discussed further in the chapters. Chapter 2 discusses about the problem definition and methodology. The development process, requirement analysis, design, implementation of the project are discussed in chapter 3.The results of the project are analyzed and concluded in the final chapter 4.

CHAPTER 2

PROBLEM DEFINITION AND METHODOLOGY

In the previous chapter, the existing system and the proposed system for this project are discussed . This chapter deals with the problem definition and the methodology. The problem definition discusses about the objective of the project and the methodology used to develop the project.

2.1. PROBLEM DEFINITION

Problem definition describes in details about the various requirements that are required for building a model and the implementation of an algorithm to enhance the accuracy.Visulaization is one of the major part where the users can understand easliy what the data represents.

- Each and every feature of the data must be clearly understood based on its types,only then visualization can be done

2.2 METHODOLOGY

The methodology followed in this project is top down approach. Top down approach emphasize planning and complete understanding of the system. This project is separated into four modules. Each module is processed that generate the result from the given data.

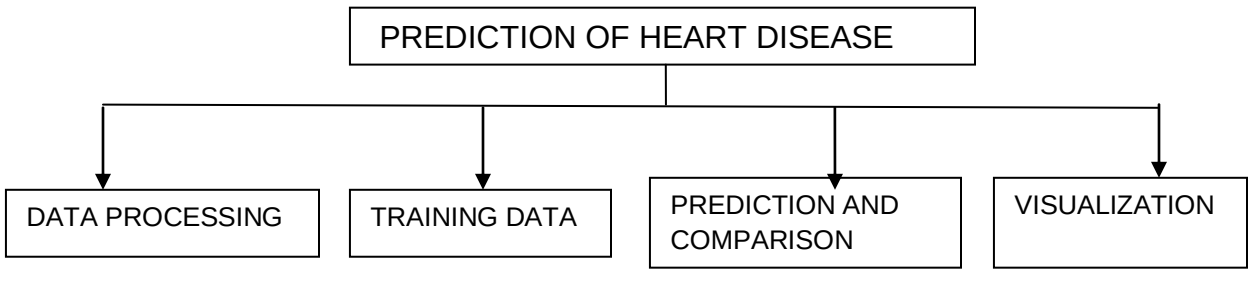


Fig 2.1 Methedology diagram

CHAPTER – 3

DEVELOPMENT PROCESS

3.1 REQUIREMENT ANALYSIS:

3.1.1 Input Requirements:

The input requirements of this project is a heart disease dataset,which is trained and a model is built.Machine learning algorithms have been used for obtaining the accuracy.

3.1.2 Output Requirements

Output of the project is the accuracy we obtained in terms of percentage.Respective graphs are obtained for the algorithms.Visualizaton is done by analyzing the various features.

3.1.3 Resource Requirements

Software Requirements:

- Programming Language: Python
- Tools:python 3.7.,tableau

Hardware Requirements:

- Operating System:Windows 10
- RAM:8 GB
- Memory:500 GB

3.2 DESIGN:

3.2.1 Architectural Design:

Fig 3.1 depicts the system architecture of the entire project.It defines the structure of the system in different modules along with the principle and elements.

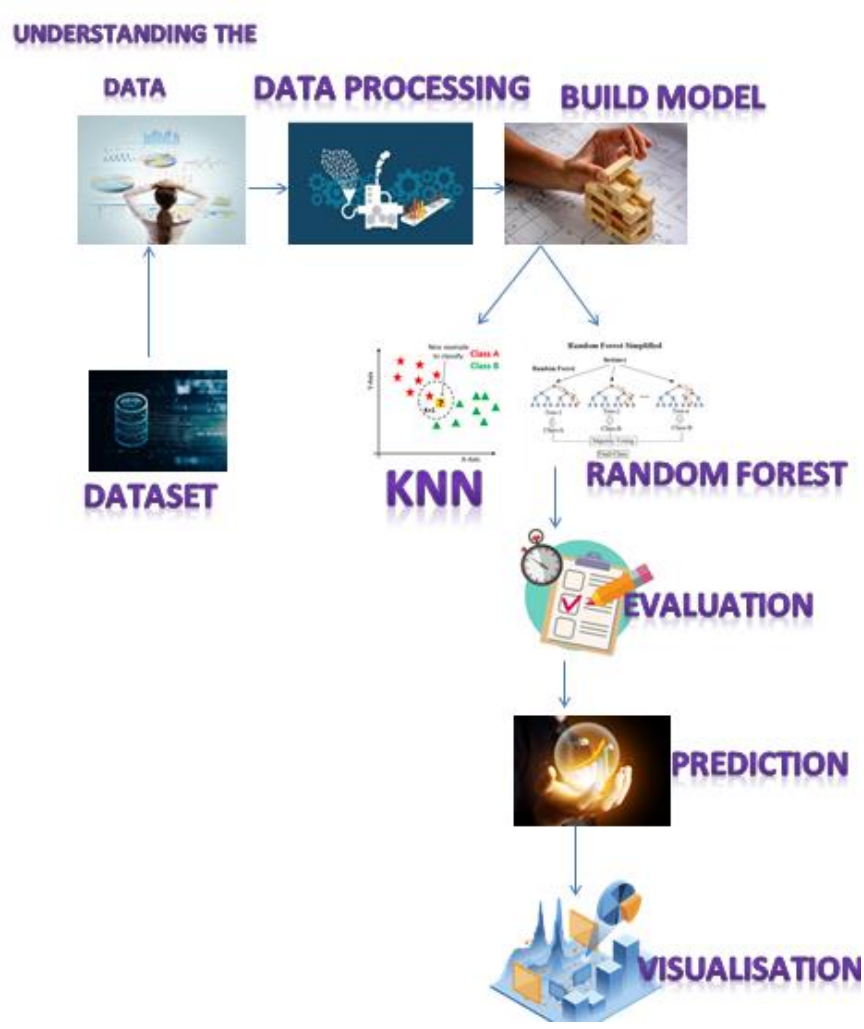


Fig 3.1 Architectural diagram

Prediction of heart disease consists of several modules are as follows:

Dataset :

Heart disease dataset is used for the prediction of heart disease.The dataset consist of several features which causes the heart disease.

Understanding the data:

Before processing a data,one must understand and analyse the data.Understanding the data is the firt step in accessing the data.Only through understanding the data,we can come to an idea how to proceed with the data.Understanding data visually helps to relate more with the data.

Data processing:

Data processing is the stage where the data should be processed by doing all the required processes to build a model.Required processes include giving dummy variables and scaling the data

Build model:

The model is build by training the data.The two machine learning algorithms such as k nerarest neighbour and random forest are considered for building a model.

Evaluation :

Evauation is a stage which allows us to make judgements and proper decisions with respect to the output of the project.Here in this project,accuracy level is found for two machine learning algorithms and an algorithm which gives the best accuracy for prediction is found.

Prediction :

Prediction is the final outcome of the project.It is an act of predicting the output of the whole project. In this project, the above mentioned machine learning algorithms are used for the prediction and it is obtained in terms of percentage (i.e accuracy level)

Visualization:

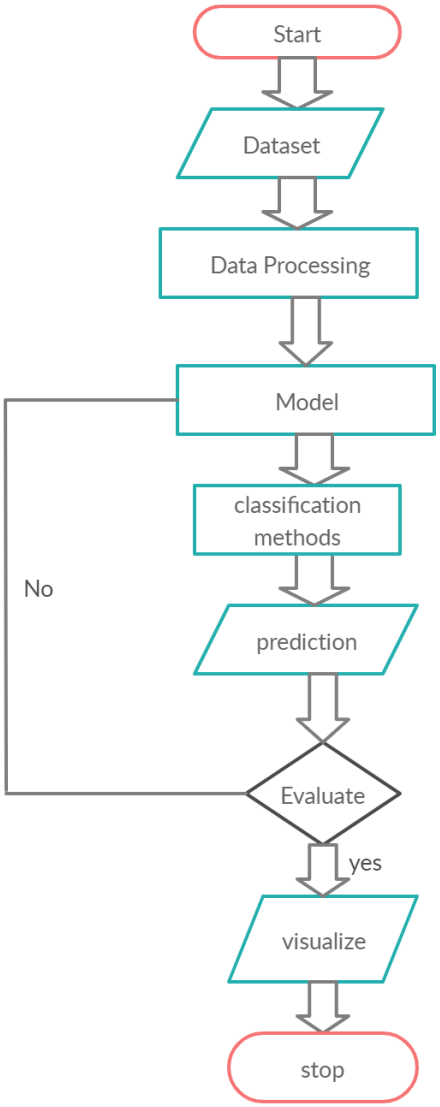
Data visualization is the representation of data or information in a graph,chart,or other visual format.It is a technique used to communicate with the data.The goal of data visualization is to communicate information clearly and efficiently.The final outcome (i.e prediction of accuracy) is visualized clearly.Different outcomes of the features are analyzed and visualized efficienty in visualizing software,Tableau.

3.2.2 Dataset design

Heart disease dataset is used for the prediction.Dataset includes patients details with 14 features causing the heart disease like chest pain,cholesterol level,electro cargiographic results,etc.

3.2.3 Data Flow design

Fig 3.2 is a data flow diagram which depicts the whole project.



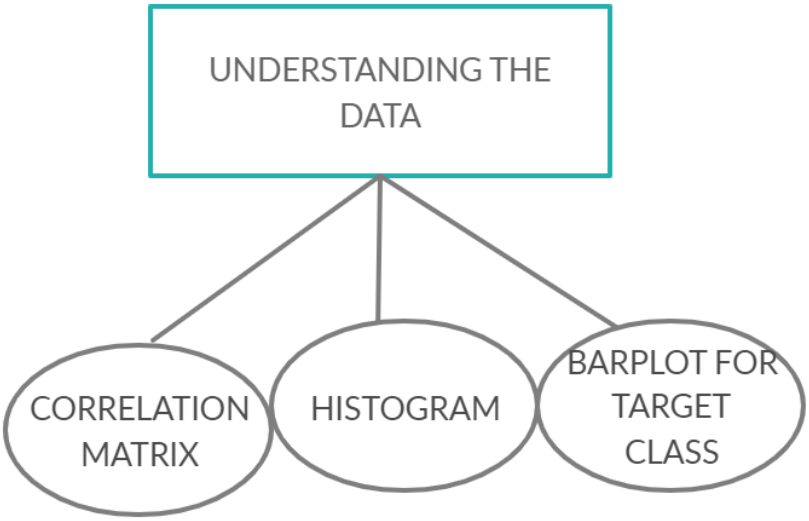
.

Fig 3.2 Data Flow diagram

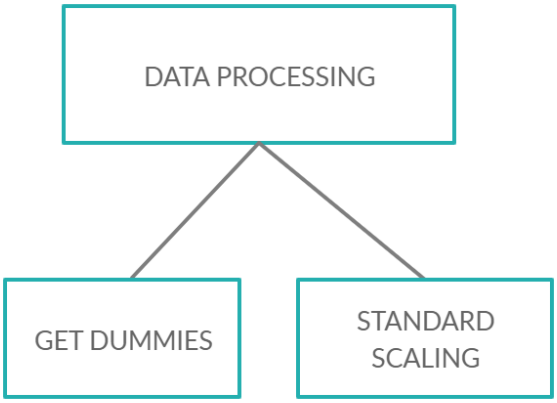
3.2.4 Detailed design

Detailed design gives a detailed view of the modules in the architectural diagram.

Understanding the data :



Data Processing:



Build Model:

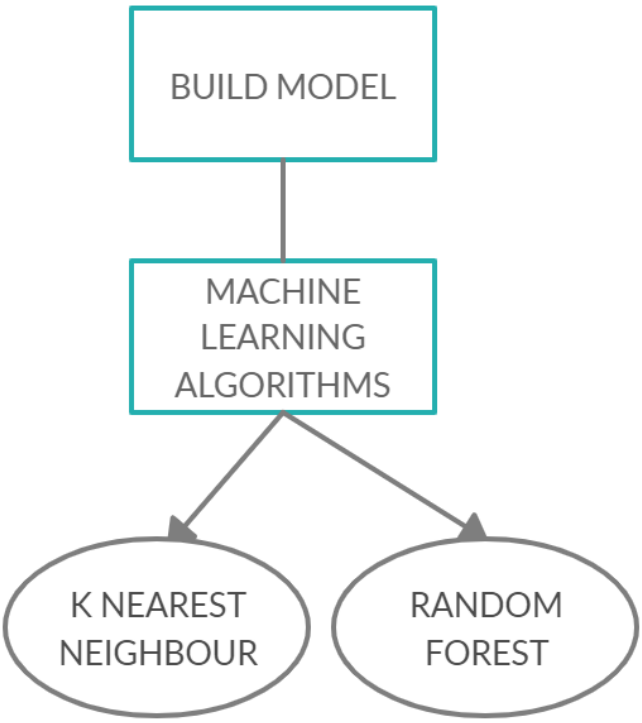


Fig 3.3 Detailed design

3.3 IMPLEMENTATION

3.3.1 Language specification

Python Technology

Python is a high level, interpreted and general purpose dynamic programming language that focuses on code readability.It has fewer steps when compared to Java and C.It was founded in 1991 by developer Guido Van Rossum.It is used in many organizations as it supports multiple programming paradigms.It also performs automatic memory management.

Advantages :

1) Presence of third-party modules

2) Extensive support libraries(NumPy for numerical calculations, Pandas for data analytics etc)

3) Open source and community development

4) Easy to learn

5) User-friendly data structures

6) High-level language

7) Dynamically typed language

8) Object-oriented language

9) Portable and Interactive

10) Portable across Operating systems

Following are some useful features of Python language:

- It uses the elegant syntax, hence the programs are easier to read.
- It is a simple to access language, which makes it easy to achieve the program working.
- The large standard library and community support.
- The interactive mode of Python makes its simple to test codes.
- In Python, it is also simple to extend the code by appending new modules that are implemented in other compiled language like C++ or C.
- Python is an expressive language which is possible to embed into applications to offer a programmable interface.
- Allows developer to run the code anywhere, including Windows, Mac OS X, UNIX, and Linux.
- It is free software in a couple of categories. It does not cost anything to use or download Pythons or to add it to the application.

Python for Data Science

Python is open source, interpreted, high level language and provides great approach for object-oriented programming. It is one of the best language used by data scientist for various data science projects/application. Python provide great functionality to deal with mathematics, statistics and scientific function. It provides great libraries to deals with data science application.
One of the main reasons why Python is widely used in the scientific and research communities is because of its ease of use and simple syntax which makes it easy to adapt for people who do not have an engineering background. It is also more suited for quick prototyping.

According to engineers coming from academia and industry, deep learning frameworks available with Python APIs, in addition to the scientific packages have made Python incredibly productive and versatile. There has been a lot of evolution in deep learning Python frameworks and it's rapidly upgrading.

In terms of application areas, ML scientists prefer Python as well. When it comes to areas like building fraud detection algorithms and network security, developers leaned towards Java, while for applications like natural language processing (NLP) and sentiment analysis, developers opted for Python, because it provides large collection of libraries that help to solve complex business problem easily, build strong system and data application.

3.3.2 Visualization tool

Tableau

Tableau is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry. It helps in simplifying raw data into the very easily understandable format.

Data analysis is very fast with Tableau and the visualizations created are in the form of dashboards and worksheets. The data that is created using Tableau can be understood

by professional at any level in an organization. It even allows a non-technical user to create a customized dashboard.

The best feature Tableau are

- Data Blending
- Real time analysis
- Collaboration of data

The great thing about Tableau software is that it doesn't require any technical or any kind of programming skills to operate. The tool has garnered interest among the people from all sectors such as business, researchers, different industries, etc.

3.3.3 Implementation technique:

3.3.3.1 Machine learning:

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans.Machine learning is actively being used today, perhaps in many more places than one would expect.

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

But, using the classic algorithms of machine learning, text is considered as a sequence of keywords; instead, an approach based on semantic analysis mimics the human ability to understand the meaning of a text.

Some machine learning methods

Machine learning algorithms are often categorized as **supervised or unsupervised**.

Supervised machine learning algorithms

Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

Unsupervised machine learning algorithms

In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

Semi-supervised machine learning algorithms

Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data requires

skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabeled data generally doesn't require additional resources.

Reinforcement machine learning algorithms

Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

Machine learning enables analysis of massive quantities of data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly. Combining machine learning with AI and cognitive technologies can make it even more effective in processing large volumes of information.

3.3.3.2.Exploratory data analysis:

Exploratory data analysis (**EDA**) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand,before getting them dirty with it.

Exploratory Data Analysis is one of the important steps in the data analysis process. Here, the focus is on making sense of the data in hand – things like formulating the correct questions to ask to your dataset, how to manipulate the data sources to get the required answers, and others. This is done by taking an elaborate look at trends,

patterns, and outliers using a visual method.

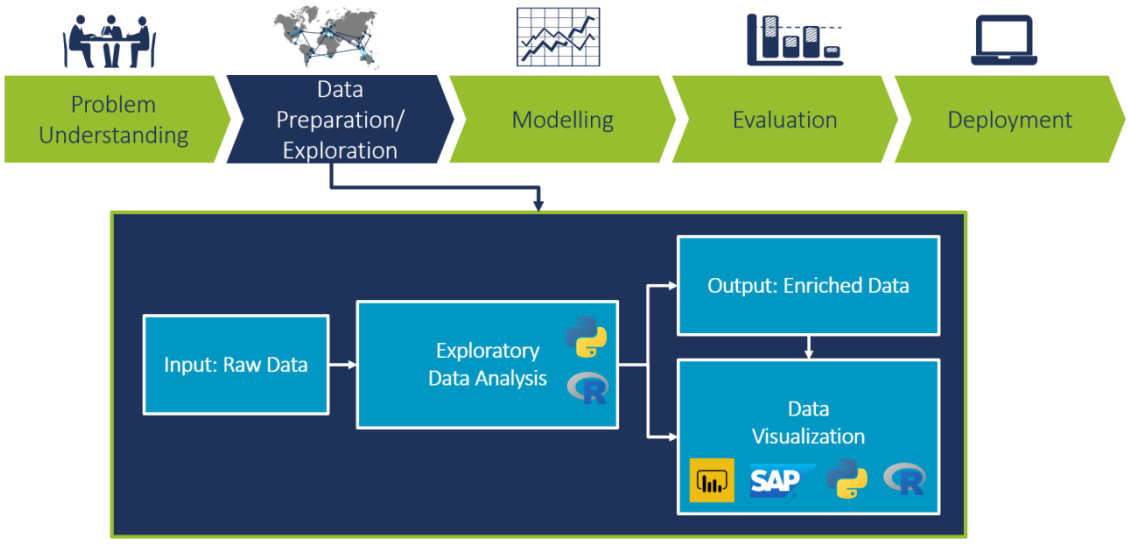


Fig 3.4 EDA technique

Exploratory Data Analysis is a crucial step before you jump to machine learning or modeling of your data. It provides the context needed to develop an appropriate model – and interpret the results correctly.

3.3.3.3.Data visualization:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps.Data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines or bars) contained in graphics. The goal is to communicate information clearly and efficiently to users. It is one of the steps in data analysis or data science.

The Importance of Data Visualization:

We are an inherently visual world, where images speak louder than words. Data visualization is especially important when it comes to big data and data analyzation projects.

Nowadays more and more companies are using machine learning to collect mass amounts of data. While it's great that they're able to do this so quickly and effectively, it also calls for a way to sort through, comprehend, and explain this data in a way that makes sense to both the business owners and stakeholders.

The same concept applies to advanced data analyzation projects.When data scientists are in the midst of a complex project, they need a way to understand the data that's being collected so that they can monitor and tweak their process to ensure it's performing the way it should.The results from complex algorithms are much easier to understand in a visual format as opposed to lines and lines of text and numbers.

Data visualization is truly important for any career; from teachers trying to make sense of student test results to computer scientists trying to develop the next big thing in artificial intelligence, it's hard to imagine a field where people *don't* need to better understand data.

3.3.4 Algorithms

3.3.4.1 K nearest neighbor:

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). K nearest neighbor has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.

K-nearest neighbors (KNN) algorithm is a type of supervised machine learning algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well −

- **Lazy learning algorithm** − It is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- **Non-parametric learning algorithm** – It is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (**K**) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

Working of KNN Algorithm

K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. We can understand its working with the help of following steps −

Step 1 − For implementing any algorithm, we need dataset. So during the first step of KNN, we must load the training as well as test data.

Step 2 − Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.

Step 3 − For each point in the test data do the following −

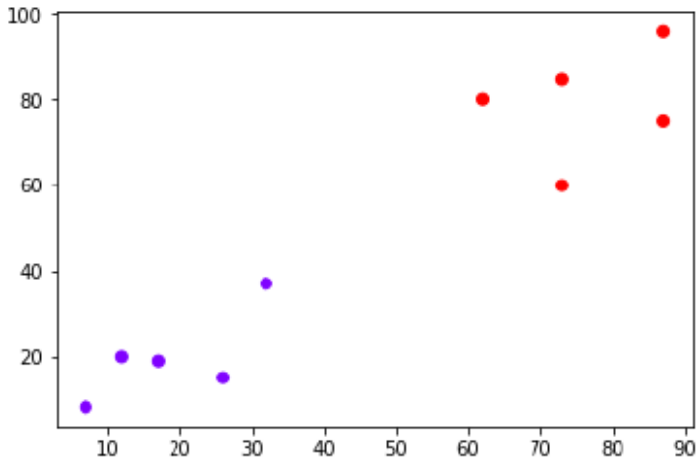
- **3.1** − Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.
- **3.2** − Now, based on the distance value, sort them in ascending order.

- **3.3** − Next, it will choose the top K rows from the sorted array.
- **3.4** − Now, it will assign a class to the test point based on most frequent class of these rows.

Example

The following is an example to understand the concept of K and working of K nearest neighbor algorithm −

Suppose we have a dataset which can be plotted as follows −



Now, we need to classify new data point with black dot (at point 60,60) into blue or red class. We are assuming K = 3 i.e. it would find three nearest data points. It is shown in the next diagram −

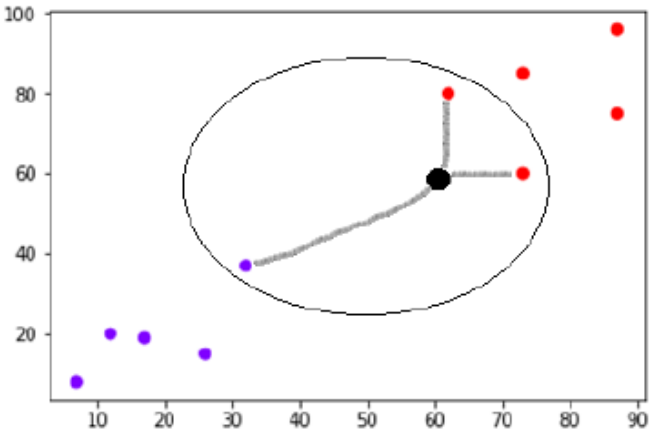


Fig 3.5 K nearest neighbor representation

We can see in the above diagram the three nearest neighbors of the data point with black dot. Among those three, two of them lies in Red class hence the black dot will also be assigned in red class.

Pros and Cons of K Nearest Neighbor

Pros

- It is very simple algorithm to understand and interpret.
- It is very useful for nonlinear data because there is no assumption about data in this algorithm.
- It is a versatile algorithm as we can use it for classification as well as regression.
- It has relatively high accuracy but there are much better supervised learning models than KNN.

Cons

- It is computationally a bit expensive algorithm because it stores all the training data.

- High memory storage required as compared to other supervised learning algorithms.
- Prediction is slow in case of big N.
- It is very sensitive to the scale of data as well as irrelevant features.

3.3.4.2.Random forest

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

The name suggest, this algorithm creates the forest with a number of trees.In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.

Working of Random Forest Algorithm

We can understand the working of Random Forest algorithm with the help of following steps −

- **Step 1** − First, start with the selection of random samples from a given dataset.
- **Step 2** − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- **Step 3** − In this step, voting will be performed for every predicted result.
- **Step 4** − At last, select the most voted prediction result as the final prediction result.

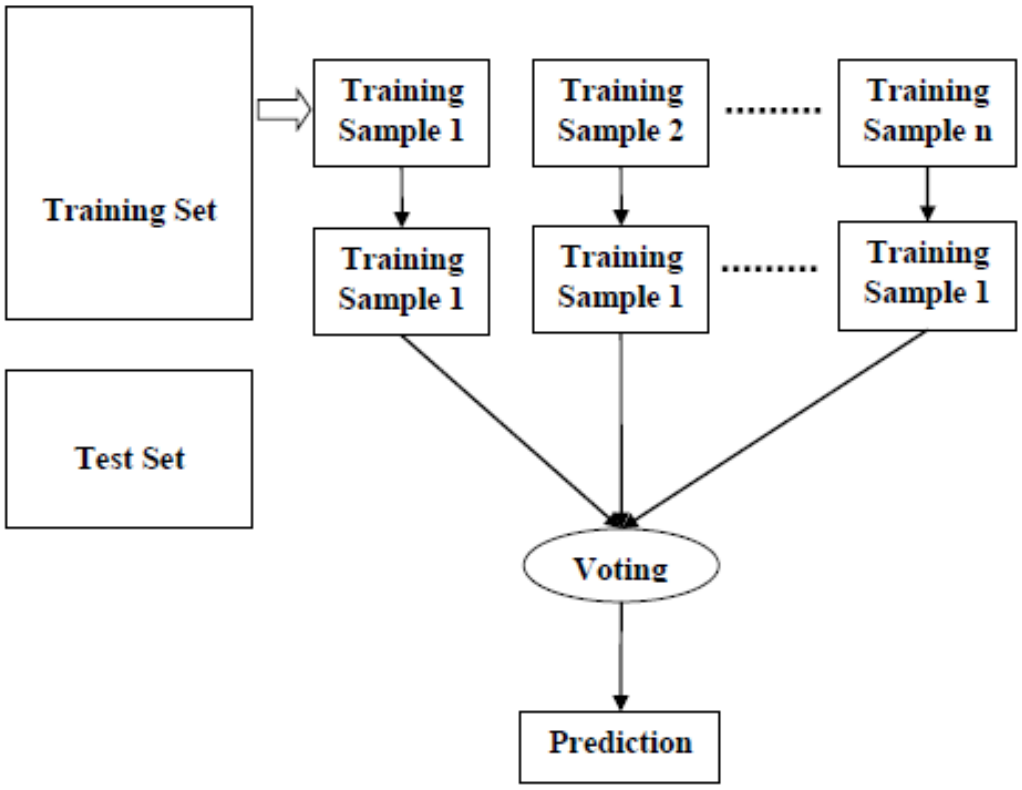


Fig 3.6 Working of random forest

Pros and Cons of Random Forest

Pros

The following are the advantages of Random Forest algorithm −

- It overcomes the problem of overfitting by averaging or combining the results of different decision trees.
- Random forests work well for a large range of data items than a single decision tree does.
- Random forest has less variance then single decision tree.
- Random forests are very flexible and possess very high accuracy.

- Scaling of data does not require in random forest algorithm. It maintains good accuracy even after providing data without scaling.
- Random Forest algorithms maintains good accuracy even a large proportion of the data is missing.

Cons

The following are the disadvantages of Random Forest algorithm −

- Complexity is the main disadvantage of Random forest algorithms.
- Construction of Random forests are much harder and time-consuming than decision trees.
- More computational resources are required to implement Random Forest algorithm.
- It is less intuitive in case when we have a large collection of decision trees.
- The prediction process using random forests is very time-consuming in comparison with other algorithms.

3.4 Plots and graphs

3.4.1 Heat map

A **heat map** (or **heatmap**) is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. A heatmap is a graphical representation of data that uses a system of color-coding to represent different values.The primary purpose of **Heat Maps** is to better visualize the volume of locations/events within a dataset and assist in directing viewers towards areas on data visualizations that matter most. But they're much more than that.

Heatmaps visualise data through variations in colouring. When applied to a tabular format, Heatmaps are useful for cross-examining multivariate data, through placing variables in the rows and columns and colouring the cells within the table. Heatmaps are

good for showing variance across multiple variables, revealing any patterns, displaying whether any variables are similar to each other, and for detecting if any correlations exist in-between them.

Fig 3.7 is a heat map representation to find the correlation matrix of features.From the below figure we can see that **cp**(chest pain),**thalach**(maximum heart rate achieved),**slope** (the slope of the peak exercise ST segment) are positively(green) correlated with the target(green)

From this we can understand that they are the major features causing the disease.

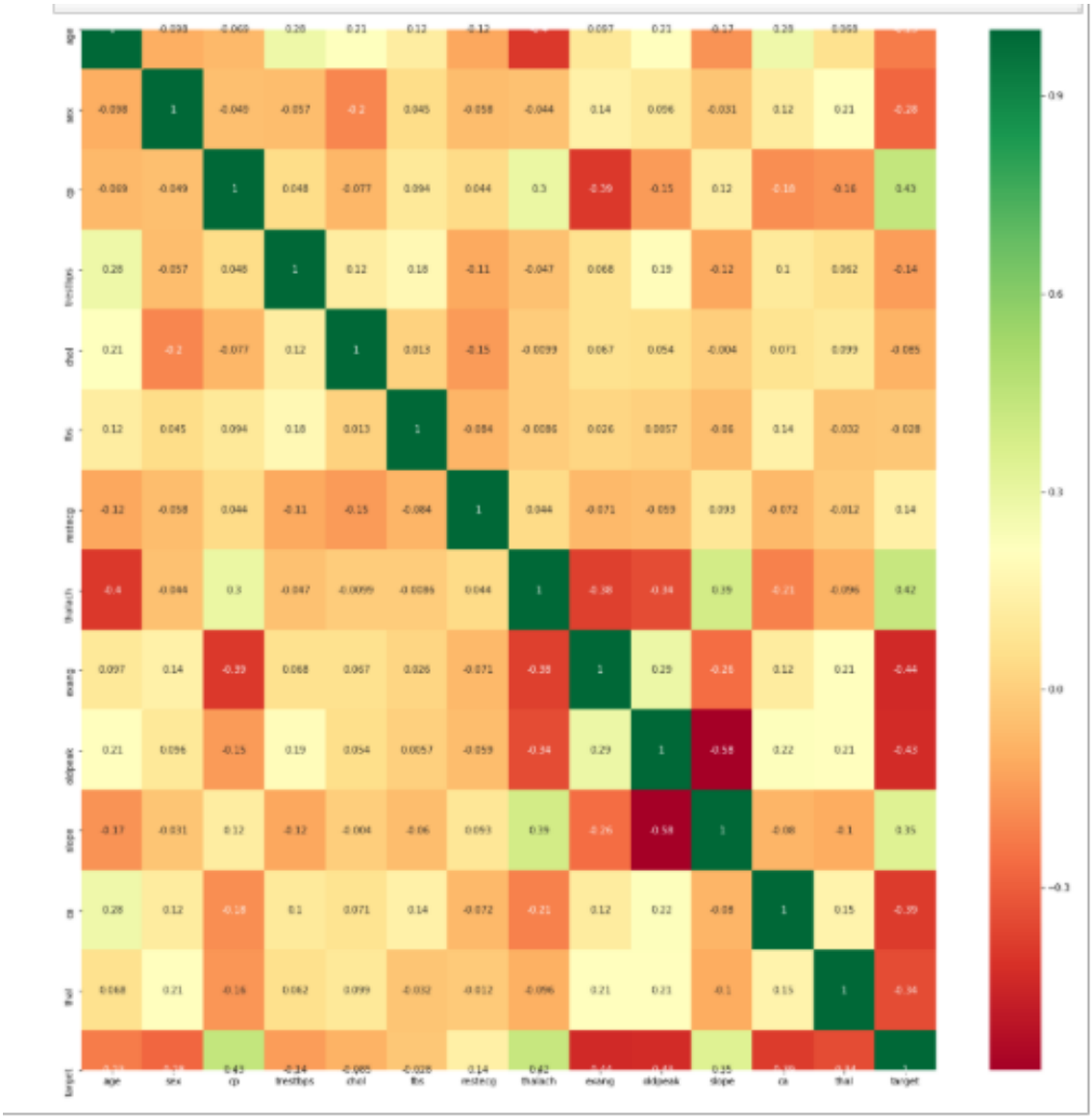


Fig 3.7 Heat map representation

3.4.2 Histogram

A **histogram** is an approximate representation of the distribution of numerical or categorical data.A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data. This allows the inspection of the data for its underlying distribution (e.g., normal distribution), outliers, skewness, etc.

Fig 3.8 is a histogram that shows how each feature and label is distributed along different ranges,which further confirms the need for scaling.

We can also observe that there are discrete bars(cp,thal,etc)which means that they are categorical variables.We will need to handle these categorical variables before applying machine learning algorithms.

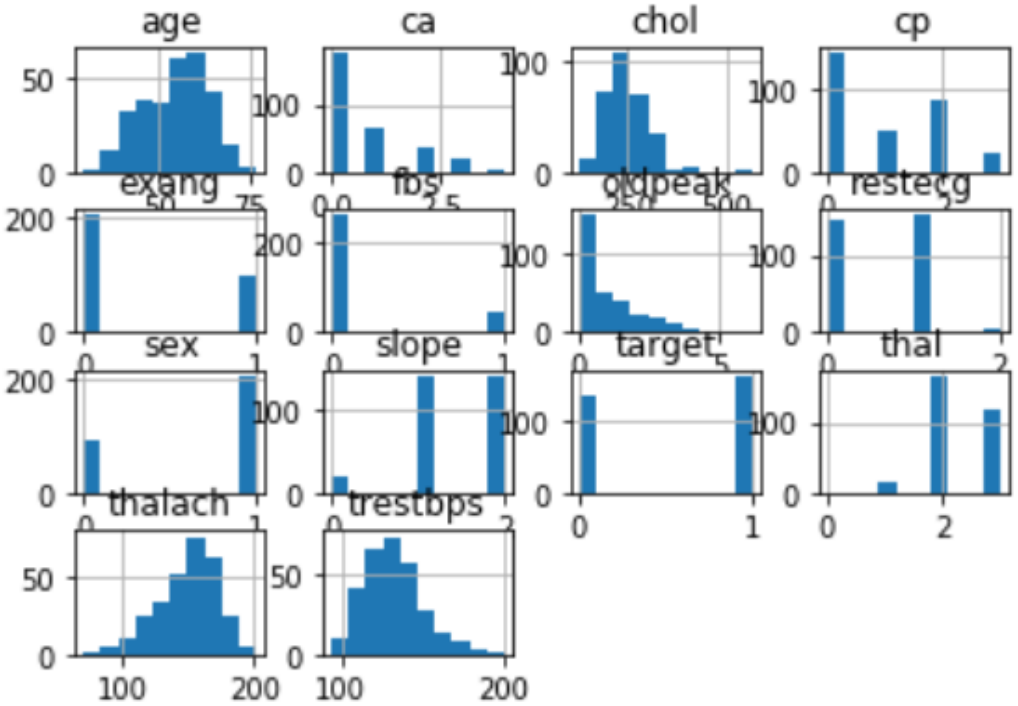


Fig 3.8 Histogram representaion

3.4.3 Bar graph

A **bar chart** or **bar graph** is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a **column chart**.

A bar graph shows comparisons among discrete categories. One axis of the chart shows the specific categories being compared, and the other axis represents a measured value

Fig 3.9 is a bar gragh that is used to check whether the dataset is balanced or not.

From this we can observe that 0(i.e absence of disease) is nearly 140 and 1(i.e presence of disease) is nearly 160.So the dataset is almost balanced.So we can procedd with the data processing.

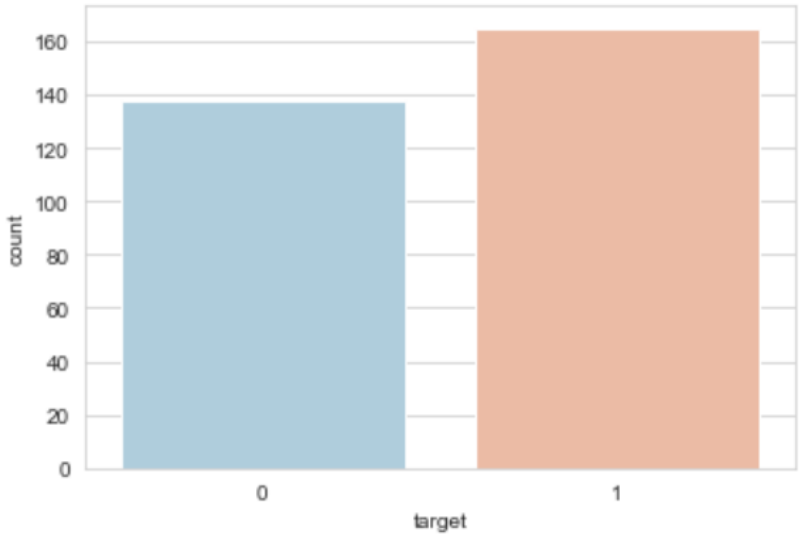


Fig 3.9 Bar graph

3.4.4 Line graph

A **line chart** or **line plot** or **line graph** or **curve chart**[1] is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. A line graph is a type of chart used to show information that changes over time. We plot line graphs using several points connected by straight lines. We also call it a line chart. The line graph comprises of two axes known as ‘x’ axis and ‘y’ axis

Fig 3.10 is a simple example of line graph.

In our project, a line graph is used to visualize the final output (i.e accuracy of the algorithm)

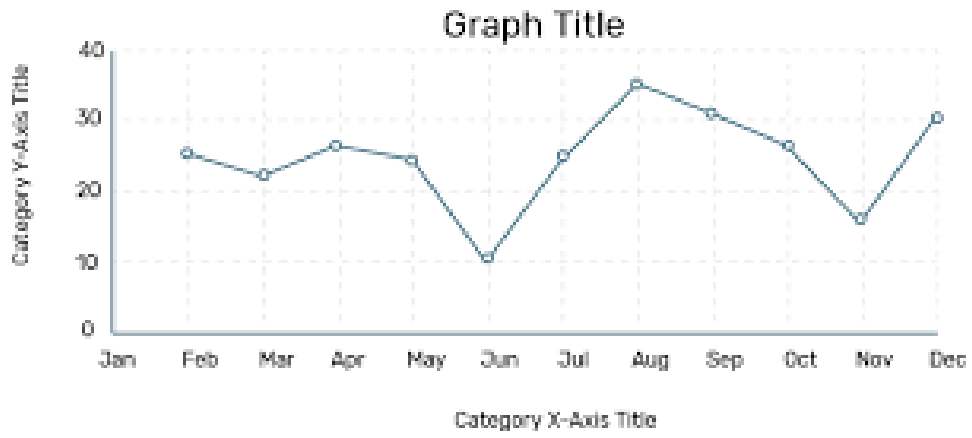


Fig 3.10 Line graph

3.5 VISUALIZATION USING TABLEAU

Patients with high risk of the disease has been visualized using tableau

Process of visualization:

- From 14 features,9 major features causing the disease has been taken for this visualization.Those 9 features include age,cp(chest pain type),chol(serum

cholesterol),trestbps(resting blood pressure),fbs(fasting blood sugar),restcg(resting electrocardiographic results),thalach(maximum heart rate),thal,exang(exercise induced angina)

- These features are then filtered with the limit causing the disease.
- The filters used for the visualization are
- Age – above 50

- Fbs – 1(true)
- Cp – 3(typical angina)
- Bp – above 100
- Chol – above 240
- Restcg – 2(ST-T wave abnormality)
- Thalach – above 100
- Exang – 1(yes)
- Thal – 3(reversible defect)
- In **Fig 3.11** common factors like age,chol,fbs,trestbps are visualized

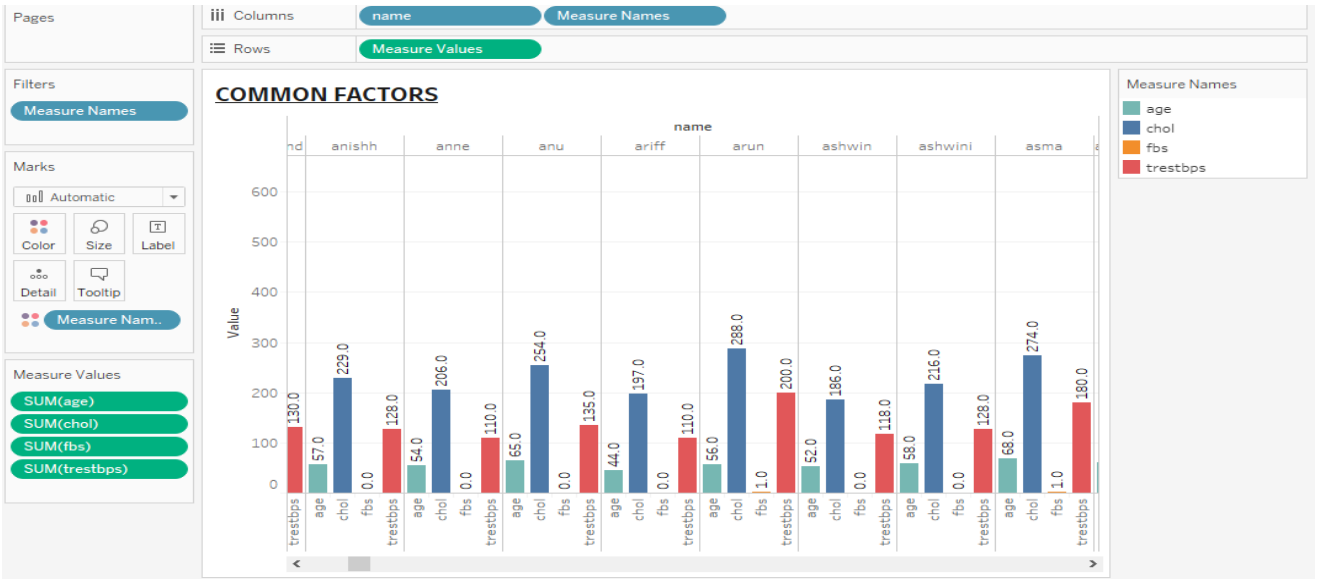


Fig 3.11 common factors

- In **Fig 3.12** common factors are filtered using the filters mentioned above and remaining(major factors)are visualized.

Fig 3.12 Major factors

- And finally all the features are filtered using the filters and the patients with high risk of the disease is obtained

CHAPTER 4

RESULTS AND CONCLUSIONS

4.1 ACCURACY

Accuracy is the final output of the project.Two supervised machine learning algorithms are used for prediction and accuracy is obtained in terms of percentage.Finally,an algorithm which gives the best accuracy is found

4.1.1 K NEAREST NEIGHBOUR

This classifier looks for the classes of k nearest neighbours of a given data point and based on the majority class,it assigns a class to this data point.However the number of neighbours can be varied.We varied them from 1 to 20 neighbours and calculated the test score(i.e accuracy) in each case.

Fig 4.1 is a line graph, plotted to visualize the number of neighbours and test score achieved in each case.

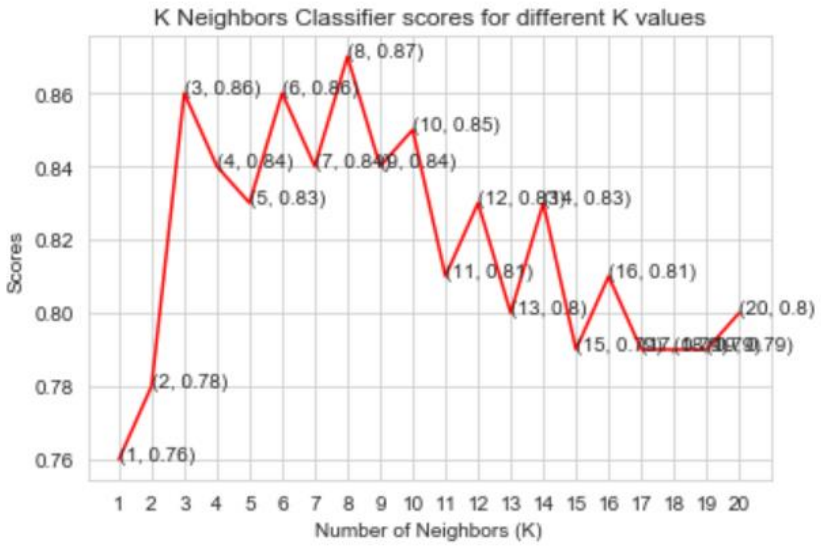


Fig 4.1 K nearest neighbor score for different k values

From the line graph ,we can observe that maximum accuracy of **87%** is achieved when the neighbor is **8.**

4.1.2 RANDOM FOREST

This algorithm creates a forest of trees where each tree is formed by random selection of features from the total features.Here we can vary the number of trees that will be used to predict the class.We calculated the test score(accuracy) over 10,100,200,500,1000 trees.

Fig 4.2 is a bar graph,plotted to visualize the number of estimators with test score achieved in each case.

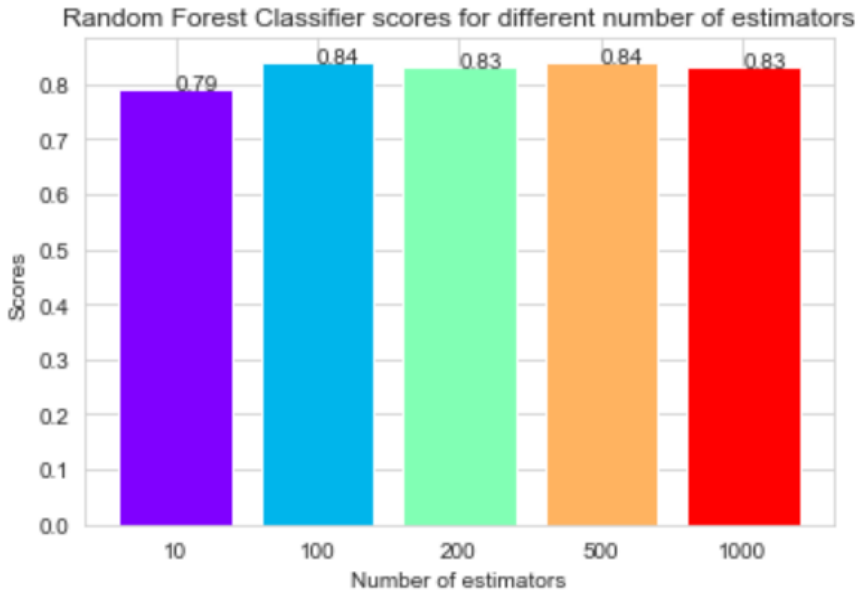


Fig 4.2 Random Forest score for different number of estimators

From the bar graph we can observe that an accuracy of **84%** is achieved for both **100** and **500** trees.

4.1.3 BEST ACCURACY

Fig 4.3 shows the accuracy obtained by the prediction of two supervised machine learning algorithms.

ALGORITHM	ACCURACY
K nearest neighbour	87
Random forest	84

Fig 4.3 Accuracy of both Machine Learning Algorithms

From **Fig 4.3** we can observe that **K Nearest Neighbour** is a supervised machine learning algorithm which gives the **best accuracy** for the prediction of heart disease.

4.2 VISUALIZATION USING TABLEAU

Patients with high risk of the heart disease has been visualized using tableau

Fig 4.4 is the final output of the patients with high risk of disease

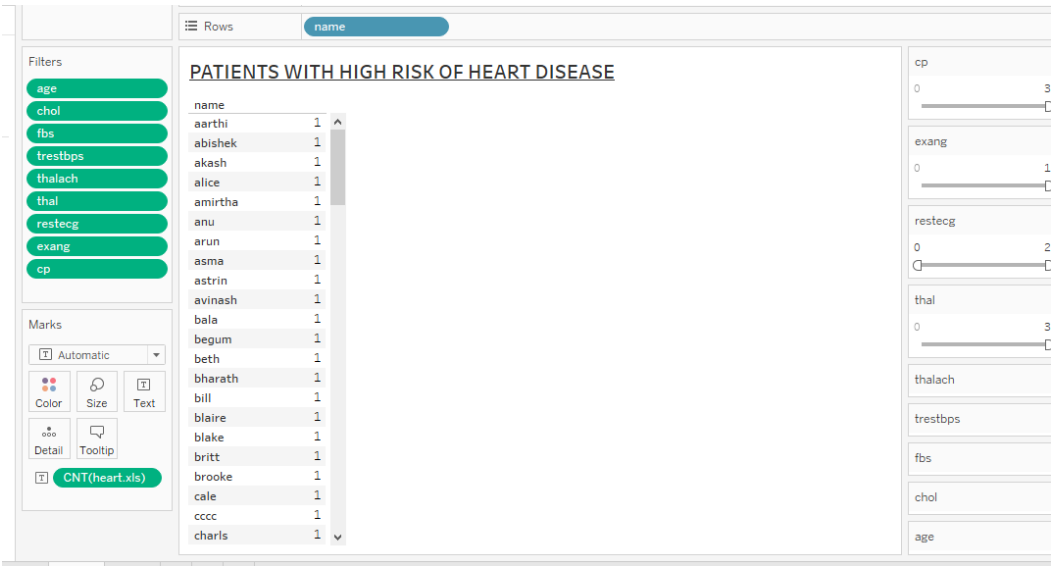


Fig 4.4 Patients with high risk of heart disease

4.3 CONCLUSION

In this project two supervised machine learning algorithms are used to predict the heart disease in terms of accuracy.The k nearest neighbor has predicted the heart disease with an accuracy of 87% and random forest with 84%.Algorithm which gives the best accuracy is found.Patients with high risk of disease is also visualized using tableau.

4.4 FUTURE ENHANCEMENTS

In future**,**the designed system is used to predict and diagnose other diseases.So it will be usefull in the medical field.Selection of features can be done in the future while dealing with the prediction.Other machine learning algorithms can also be used for the prediction and to enhance the accuracy.

REFERENCES

[1] Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., &Lamgunde, A. (2017, June). Heart disease prediction using data mining techniques. In 2017 International Conference on Intelligent Computing and Control (I2C2) (pp. 1-8). IEEE.

[2] Gandhi, Monika, and Shailendra Narayan Singh."Predictions in heart disease using techniques of data mining." In 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), pp. 520-525.IEEE, 2015

[3]Rajesh,T Maneesha,shaik Hafeez,Hari Krishna “prediction of heart disease using machine learning algorithms”.In 2018 International Journal of Engineering and technology,(2.32)363-366

[4] Purushottam,prof.dr kanak saxena,Richa sharma” Efficient Heart Disease Prediction System using Decision Tree”.In 2015, International Conference on Computing, Communication and Automation(ICCCA2015)

[5] Reddy Prasad,Pidaparthi Anjali, S.Adil, N.Deepa” Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning”.In february 2019 International Journal of Engineering and Advanced Technology (IJEAT)

[6] Monika Gandhi,2015. Prediction in heart disease using techniques of data mining, International conference on futuristic trend in computational analysis and knowledge management(ABLAZE- 2015)

[7] Dinesh Kumar G, 2018 . Prediction of cardiovascular disease using machine learning algorithms, procedingof 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India.

[8] Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., & Lamgunde, A. (2017, June). Heart disease prediction using data mining techniques. In 2017 International Conference on Intelligent Computing and Control (I2C2) (pp. 1-8). IEEE.

APPENDIX

SOURCE CODE

```
import numpy as np

import matplotlib.pyplot as plt

from matplotlib import rcParams

from matplotlib.cm import rainbow

import seaborn as sns

import pandas as pd
```

IMPORTING DATASET

```
df=pd.read_csv('c:/Users/8460p/Downloads/heart.csv')

df.info()

df.describe()
```

UNDERSTANDING THE DATA

```
corrmat = df.corr()

top_corr_features = corrmat.index

plt.figure(figsize=(20,20))
```

```
g=sns.heatmap(df[top_corr_features].corr(),annot=True,cmap="RdYlGn")

df.hist()

sns.set_style('whitegrid')

sns.countplot(x='target',data=df,palette='RdBu_r')
```

DATA PROCESSING

```
dataset = pd.get_dummies(df, columns = ['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal'])
```

STANDARD SCALING

```
from sklearn.preprocessing import StandardScaler

standardScaler = StandardScaler()

columns_to_scale = ['age', 'trestbps', 'chol', 'thalach', 'oldpeak']

dataset[columns_to_scale] = standardScaler.fit_transform(dataset[columns_to_scale])

dataset
```

TRAINING AND TESTING DATA

```
from sklearn.model_selection import train_test_split

y = dataset['target']
```

```
X = dataset.drop(['target'], axis = 1)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.33, random_state =
0)
```

K NEAREST NEIGHBOR

```
from sklearn.neighbors import KNeighborsClassifier

knn_scores = []


for k in range(1,21):

 knn_classifier = KNeighborsClassifier(n_neighbors = k)

 knn_classifier.fit(X_train, y_train)

knn_classifier.score(X_test, y_test)

knn_scores.append(knn_classifier.score(X_test, y_test))


plt.plot([k for k in range(1, 21)], knn_scores, color = 'red')

for i in range(1,21):

plt.text(i, knn_scores[i-1], (i, knn_scores[i-1]))

plt.xticks([i for i in range(1, 21)])

plt.xlabel('Number of Neighbors (K)')
```

```
plt.ylabel('Scores')

plt.title('K Neighbors Classifier scores for different K values')


knn_scores


knn_classifier = KNeighborsClassifier(n_neighbors = 8)

knn_classifier.fit(X_train, y_train)

knn_classifier.score(X_test, y_test)
```

RANDOM FOREST

```
 from sklearn.ensemble import RandomForestClassifier

rf_scores = [ ]


estimators = [10, 100, 200, 500, 1000]

for i in estimators:

 rf_classifier = RandomForestClassifier(n_estimators = i, random_state = 0)

 rf_classifier.fit(X_train, y_train)

 rf_classifier.score(X_test, y_test)

 rf_scores.append(rf_classifier.score(X_test, y_test))
```

```
colors = rainbow(np.linspace(0, 1, len(estimators)))


plt.bar([i for i in range(len(estimators))], rf_scores, color = colors, width = 0.8)

for i in range(len(estimators)):

plt.text(i, rf_scores[i], rf_scores[i])

plt.xticks(ticks = [i for i in range(len(estimators))], labels = [str(estimator) for estimator in estimators])

plt.xlabel('Number of estimators')

plt.ylabel('Scores')

plt.title('Random Forest Classifier scores for different number of estimators')


rf_scores


rf_classifier = RandomForestClassifier(n_estimators = 100, random_state = 0)

rf_classifier.fit(X_train, y_train)

rf_classifier.score(X_test, y_test)
```

TECHNICAL BIOGRAPHY

AZARUDEEN.A(RRN:171421601004) was born on 3rd March 2000. Currently pursuing Bachelor of Computer Applications degree programme with specialization in Data Science from BS Abdur Rahman Crescent Institute of Science and Technology, Vandalur. Chennai -600048.

Email id:azarazad7078401@gmail.com

Mobile No:9600537556

TECHNICAL BIOGRAPHY

MOHAMMED BASHEER.D(RRN:171421601017) was born on 7th August 1999. Currently pursuing Bachelor of Computer Applications degree programme with specialization in Data Science from BS Abdur Rahman Crescent Institute of Science and Technology, Vandalur. Chennai -600048.

Email.id:d.mdbasheer1999@gmail.com

Mobile number:8667083016

TECHNICAL BIOGRAPHY

SUREHA.S(RRN:171421601033) was born on 17th November 1999. Currently pursuing Bachelor of Computer Applications degree programme with specialization in Data Science from BS Abdur Rahman Crescent Institute of Science and Technology, Vandalur. Chennai -600048.

Email.id:surehasrinivasagan@gmail.com

Mobile No:9094481171