Heart Disease Prediction Using Techniques of Classification in Machine Learning

by

Sadia Afrose
17101546
Farah Rubaiat
17101540
Homayra Tabassum
17301068

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
June 2021

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Sadia Afrose
17101546

Farah Rubaiat
17101540

Homayra Tabassum
17301068

Approval

The thesis/project titled "Heart Disease Prediction Using Techniques of Classification in Machine Learning" submitted by

1. Sadia Afrose (17101546)
2. Farah Rubaiat (17101540)
3. Homayra Tabassum (17301068)

Of Spring, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on June 10, 2021.

Examining Committee:

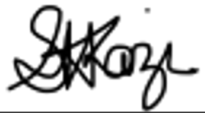
Supervisor:
(Member)

Dr. Amitabha Chakrabarty
Associate Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)



Sadia Hamid Kazi
Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

In this thesis we have examined the accuracy of various classifiers to predict heart disease and heart vessel blockage. We have also analyzed the key features contributing to heart vessel blockage. We have used a dataset containing 14 attributes related to heart disease of 1025 patients. From our study we found that the Decision Tree, Random Forest and KNN algorithm gave the highest accuracy for detecting heart disease. For predicting heart vessel blockage, the Decision tree had the highest accuracy. While analyzing the features contributing to heart vessel blockage, we found that patients' age and cholesterol level has the highest contribution. Hence, monitoring the patient's cholesterol level may help prevent heart vessel blockage.

Keywords: Heart Disease, Coronary Artery Blocks, Chest Pain, Machine Learning, Angina, Disease Prediction, Machine learning, CatBoost, lightGBM, XGBoost, Gradient, AdaBoost, Decision Tree, Random Forest, SVM, KNN, Naive Bayes, Logistic Regression, Linear Regression.

Acknowledgement

All praise to the Almighty for whom we have completed our undergrad thesis amidst this pandemic. We are thankful to our advisor Dr. Amitabha Chakrabarty for his support and guidance in our work. We have learned valuable lessons from him that will help us in our future academic and personal lifes.
We are thankful towards our family members for providing love and support throughout our undergrad. Without their support we could not have completed our journey.
We would like to thank senior lecturer, Mr.Annajiat Alim Rasel and Lecturer, Rubayat Ahmed Khan for motivation and introducing us to the fundamentals of computer science and engineering during our first year of undergraduate.They have inspired us to enjoy and acknowledge the beauty of studying computer science.
Our undergraduate journey has been pleasant, thanks to the faculty of computer science and engineering. They have been supportive throughout our entire journey of undergraduate.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Heart is one among the foremost prime organs of our body. It forces our body to stay alive by pumping blood to each area of our whole body. In case it ceases to behave properly, various organs of our body will stop working, specially the brain. The person may die within a few minutes because of this. Our lifestyle plays a very significant role in maintaining our heart healthy. Stress, food habits, sleep, diabetes, high blood pressure, high cholesterol, abnormal pulse rate, almost everything contributes to several heart related diseases. Heart diseases now become one of the most prominent causes of death all around the world. According to the World Health Organization (WHO), Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, an estimated 31% of all deaths worldwide. It was estimated to account for 7.1 million deaths worldwide in 2016 which is really concerning because only within 4 years It has increased noticeably [23]. Doing nothing about it may lead to early death. The viewpoint of medical science and data mining are used for focusing on various types of metabolic syndromes. Data mining for classification plays a vital role for prediction of heart condition and data investigation. Medical organizations all around the world, accumulate data on several health related issues. For effective comprehension, these data are often used for various machine learning process. But the data collection is not an easy process at all always. Most of the time, it gives noisy data or outputs to use or making decisions. Beacuse of this dataset issue, it gets harder and complicated for an average human mind to attain the ability to comprehend and this issue can be simply inspected by applying profuse machine learning approach. Thus, we can easily predict or make an assumption about a possibility of having a disease by implementing the algorithms. It also provides an additional source of data for creating or predicting decisions to the healthcare professionals

1.1 Motivation

Machine Learning (ML) utilizes data from the real world like weather, movies, medical diagnosis, image recognition etc. It has been used in multiple industries and fields. For instance, possibilities, classifications, regression, learning association etc. The number of data is generated everyday all over the world can be very useful for machine learning. Almost all the same types of data can generate a pattern which can be used to predict the future by using machine learning. Since it is the study

of computer algorithms which gets better from time to time automatically through experience and by the use of data, we can say that the more data we can get, the better prediction it can give. So, for real world problems, machine learning is a really useful and helpful tool. For example, a film maker wants to know what type of films audience will like the most. For detecting a disease, which factors play a notable role is always wanted to be known. If we look into our day to day activities, we will see ML is also being used to predict the weather. How much loss or gain can be achieved by an industry or a company is also determined by ML.
We can talk about Yandex,Taxi. This is a Russian company and they used machine learning for their ease of work. Machine learning is also used in Facebook, twitter. Twitter's AI examines every tweet in real time and then it "scores" them according to various metrics.
Spam(often explained as unsolicited email messages) detection is also handled by ML. to distinguish between spam and legitimate email messages, spam categorization is important. Decision tree can be used to handle this. Again, Support Vector Machines (SVMs) are a new learning method and achieve a solid upgrade over the currently preferred methods [1].
Almost every company or industry, sections are getting more interested in ML due to its easiness, simplicity and effectiveness, the medical sector is not an exception and connecting human bodies with machines is always very fascinating to us. We need more associations between the health sector and the technology, as it is the faster way to handle big chunks of data. For the fast prediction ability, not only medical but also almost every sector is getting interested in machine learning. That is why, in ten years we might be operating almost everything under machine learning and then we will need a lot of people knowing how to deal with it and how to tackle new problems using machine learning. Therefore, Our motivation is simple. We want to catch up with the upcoming machine learning wave that can make our life easier and better. Besides, there are lots of projects which are open to everyone to contribute, that motivates us even more to engage in to serve as a part of the bigger group.

1.2 Thesis Contribution

1.2.1 Problem Statement

Millions of people are being affected by heart diseases every year. Medical diagnosis is essential when it comes to treating heart disease. The medical service experts find it easier to provide effective treatment to the patients if the disease is diagnosed in their early stages. It is difficult and expensive to perform medical tests to diagnose heart diseases. Our goal is to apply different data mining techniques to help medical service experts to predict heart disease efficiently and accurately.

1.2.2 Solution

There has been a lot of work in the field of predicting heart disease using various machine learning algorithms. As we are moving forward in the field of machine learning there are new algorithms being introduced to use that can be used to help us predict heart diseases. Therefore, constant analysis and comparison among these algorithms is important to keep upto date about the best possible solution for predicting heart diseases. In this research we have worked on a dataset consisting of information of 1025 numbers of patients with features such as age, blood pressure, total cholesterol etc . In our work we have implemented a total of 12 algorithms such as CatBoost, lightGBM, XGBoost, AdaBoost, Random Forest, Linear and Logistic Regression, etc. Furthermore, we have used these algorithms to predict heart vessels block using the same dataset. Among these algorithms the ones with highest accuracy were used to find out top feature importance for predicting heart vessels block. Using these feature importance we can understand the key contributors to heart block among the patients. Using these information doctors can advise their patients to monitor such key features to reduce the chances of heart vessels block.

1.2.3 Methodology

First step of our work would be to collect a data set that contains risk factors of heart disease of patients. Such risk factors are age, blood pressure, total cholesterol, diabetes, hypertension, family history of heart disease, obesity and lack of physical exercise, fasting blood sugar etc [7]. We will clean up null data from the data set. Next, we will create visual representation of our data through various charts and graphs to have a better idea about our data set. After that we will separate the data into two sections, training data set and testing data set. Testing data set will have a higher proportion and the subset will be created randomly. We will use the training data set to train different models using different ML classification algorithms such as Neural Network, Decision tree, Random Forest, Native Bayes, Logistic Regression, Adaboost etc.

Chapter 2

Machine Learning

2.1 Introduction of Machine Learning

Machine learning is a way of programming the computers which will help computers to learn from data. Using Machine learning, we fit a dataset into models that can be understood and utilized by people.
In traditional computing, algorithms are designed with specific instructions to solve a problem. Machine learning algorithms instead allow the computer to train on a dataset and use statistical analysis to give output values. This is an automatic decision-making system that does not follow a set of instructions to reach its goal.
Anyone using technology is being benefited from machine learning. People are using facial recognition technology on social media to tag or share pictures. Based on user preference various websites recommend movies. songs, books, articles etc. Self-driving cars use machine learning algorithms to navigate, recognise objects on streets etc [20].

Using machine learning we can develop a model that can make decisions on its own.The process of machine learning can be explained in 7 steps [18].

1. Data collection
2. Data preparation
3. Choosing a model
4. Training
5. Evaluation
6. Parameter tuning
7. Prediction

Figure 2.1: Steps of Machine Learning [18]

2.2 Various Types of Machine Learning Systems

There are several types of machine learning systems. We can classify them based on the following criteria:

- Is the system trained under human supervision or not
- Can the system learn on the fly incrementally
- Does the system compare new data with previously known data points or it tries to find patterns in the training set and builds a model to predict.

2.2.1 Supervised or Unsupervised Learning

There are four major categories based on the supervision of the system in training :

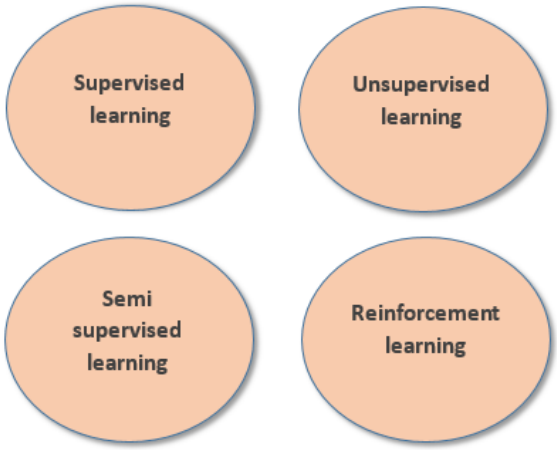


Figure 2.2: Catagory based on the Supervision of the system in training

Supervised Learning

In supervised learning, the training set includes the expected outcomes which is called labels . These data sets are designed to train or "supervise" algorithms into classifying data or predicting outcomes accurately [27]. Classification is a typical supervised learning task and widely used such as spam filters. Another task is to predict a target value. This is called regression. Some regression algorithms can also be implemented for classification. Such as linear and logistic regression.

Unsupervised Learning

In unsupervised learning, the training data set does not have the desired solutions included. In supervised learning the system tries to find the hidden structure of the inlabled dataset [27]. Like, clustering algorithms try to find groups of similar data. Visualization algorithms are also good for using unsupervised learning. The unlabeled data is represented in 2D or 3D graphs. Dimensionality reduction is a similar task. Here the goal is to simplify data with most minimum information loss. Another unsupervised task is anomaly detection. It is an important task. For example, preventing credit card fraud, identifying manufacturing defects etc.

Semi Supervised Learning

Semi supervised learning uses partially labeled training data. Google photos is a good example for semi supervised learning. Here the user may label some people in most photos, and some photos stay unlabeled. Mostly a combination of supervised and unsupervised algorithms are used for semi supervised learning. For example, deep belief networks (DBNs) use semi supervised learning.

Reinforcement Learning

In reinforcement educating the knowledge of the learning system, called an agent. The environment can be observed from this. Selecting and performing the actions, and get rewards and penalties in return based on the effect of their actions are also included here. The agent tried to maximise rewards in a situation taking suitable action. For example, robots can use reinforcement algorithms during walking.

2.3 Batch and Online Learning

Machine learning system can be classified based on if incrementally learnt by the process from a flow of incoming data [27].

Batch Learning

The process in batch learning can not learn incrementally. It has to be trained using all the available data [27]. It takes a lot of time and computing resources to train the system. Therefore in offline it can be done. Before introducing into production the process is informed. While it continues it does not learn anymore. Batch learning is also called offline learning. In order to modify or improve the system we have to train the system with a new dataset from scratch. Then the new version will replace the old one.

Online Learning

This process is care trained increasingly. Data cases could be received by the process sequentially, one by one or in short groups. Every instruction step is speedy and also requires small computation resource [27]. So that Data on the fly can be learned by the process. For system which requires new data, for them online learning can be helpful continuously and for changing quickly they are in necessity to adjust, such as stock prices. Once an online system is trained with new data there is no need to save these data anymore. Since online learning algorithms does not need to store all data in the system. Usually it uses on large datasets to train which is not able to adjust in any machine's leading memory.

2.4 Challenges of Machine Learning

To design a good machine learning system we need to select suitable algorithms with appropriate dataset. There can be many challenges in designing a good system.

- Low Quantity of training data : For most machine learning algorithms it takes a lot of data to train the model properly. Lack of enough data will cause the accuracy of the model.
- Poor-Quality Data : Even if we have a large enough dataset the quality of the dataset needs to be ensured. If the data is filled with error, outliers and noise the model will fail to detect the underlying pattern.
- Irrelevant Features : The system will be able to learn if the dataset is filled with relevant creatures. Irrelevant features will cause complications in the training.
- Overfitting the training data : the model performs well on training data but fits poorly on the new dataset.
- Underfitting the training data : when the model is too simple to learn the underlying structure of the data then underfitting occurs.

Chapter 3

Literature Review

Machine learning is a powerful tool for identifying diseases. We have many classifier algorithms that can be implemented for heart disease and heart vessel blockage detection. However the efficiency of the model depends on the nature of the dataset. Throughout our literature review we have found implementation of various classification algorithms for heart disease prediction.

Sellappan Palaniappan et al [3] proposed an Intelligent Heart Disease Prediction System (IHDPS). It is developed by using data mining techniques such as Neural Network, Decision Trees and Naive Bayes. Each method has its own strength or strategy to get appropriate outcomes. To build this system relationships between them and hidden patterns are used.

Mohammed Abdul Khaleel et al[6] has published a paper that focuses on information mining procedures. These are used for information mining to find local illness. Such as, heart infirmities, bosom disease etc. Naive Bayes algorithm has been implemented here. Bayes theorem was used in the Naive Bayes algorithm. Hence Naive Bayes can make independent assumptions. A leading diabetic research institution of Chennai, Tamilnadu provided the dataset used in their paper, with patient's information. weka tool is used and 70% of the split is used for classification. Naive Bayyes showed 86.419% accuracy.

Haik Kalantarin, Mohammad Pourhomayoun, Costas Sideris and Nabil Alshurafa et al[8] have written a paper titles "Remote health monitoring outcome success prediction using baseline and first month intervention data".RHS systems are helpful in bringing down the illness and saving cost. They have portrayed an upgraded RHD framework named Wanda CVD. Here it is a cell phone based framework. Their goal is to provide remote instructure help to members. Social welfare associations focus on CVD counteractive action measures.

B. Gomathy and M.A. Nishara Banu et al [4] have published a paper "Disease predicting system using data mining techniques" where they talk about MAFIA (Maximal Frequesnt Itemset Algorithm) and K-means clustering. Here their classifier has high accuracy.

Hari Kusnanto and Wiharto el st [9] published a paper "Intelligence system for diag-

nosis level of coronary heart disease with K-star algorithm". The paper have exhibit a framework utilizing learning vector quantization neural system calculation. The framework predicts that there is a nearness or nonattendance of coronary illness in the patients and various execution measures.

Niti Guru et al[29] presented " Decision support system for heart disease diagnosis using neural network sease diagnosis using neural network niti guru* anil dahiya". The dataset used in the paper has 13 attributes. The supervised networks i.e. For training and testing of data, Neural Network with back propagation algorithm is applied.

Krishnan et al [15] proved that the decision tree is more accurate as compared to the naïve bayes classification algorithm in their project.

Gavhane et al [12] has implemented CAD technology on a multi layer perception model to predict heart diseases. As more people use the prediction system the awareness about heart disease will increase. Which will lead to reduction of the death rate of heart patients.

Senthilkumar Mohan and his fellow members et al [17] worked with various types of algorithms and they compared those algorithms' results at the end. They not only compared but also proposed a new idea that gave better accuracy than other algorithms they used in their paper.

We have noticed that the dataset used in Senthikumar Mohan and his fellow members work et al [17] is similar to the dataset we have worked on. In this paper they have implemented 9 algorithms for heart disease detection and compared the results. Since the structure of their work is similar to us we have chosen this paper for our results comparison.

Chapter 4

Collecting and Processing the Dataset

For figuring out which algorithm is more suitable for predicting heart disease, from UCI repository, cleveland dataset is used, which is accessible at mentioned link (i)http://archive.ics.uci.edu/ml/datasets/Heart+Disease

And also at
(ii)https://www.kaggle.com/rajnikant2020/heart-data
The dataset(i) has 76 number of attributes and 303 number of records mostly used in the papers we have read. However, 14 attributes and 1025 records [from (ii)] are used for this study and testing. The reason for using the second dataset is that the dataset(i) is already used multiple times. Therefore, using an updated and different dataset can give different outcomes.

Table 4.1: Data description table of the dataset

No.	Attributes	Type	Description
1	age	Integer(continuous)	Age given in years
2	sex	Integer(Discrete)	1=male; 0=female
3	cp	Integer (Discrete)	chest pain type 0= no chest pain; 1= typical angina; 2= atypical angina; 3= non angina pain; 4= asymptomatic
4	trestbps	Integer(continuous)	Resting blood pressure (in mm Hg on admission to the hospital
5	chol	Integer (continuous)	Cholesterol Serum cholesterol in mg/dl
6	fbs	Integer (Discrete)	FBS over 120 mg/dl 1 = true; 0 = false; (fasting blood sugar greater than 120 mg/dl)
7	restecg	Integer (Discrete)	Resting electrocardiographic results 0 = normal; 1 = having ST-T wave abnormality. (T wave inversions and/or ST elevation ordepression of $>$ 0.05 mV); 2 = showing probable or definite left ventricular hypertrophy
8	thalach	Integer (Continuous)	Maximum heart rate achieved
9	exang	Integer (Discrete)	1 = yes; 0 = no (exercise induced angina)
	oldpeak	Decimal	ST depression induced by exercise relative to rest
11	slope	Integer (Discrete)	Slope of ST 1 = up sloping; 2 = flat; 3 = down sloping The slope of the peak exercise segment
12	ca	Integer (Continuous)	Number of major vessels colored by fluoroscopy that ranged between 0 and 4
13	thal	Integer (Discrete)	Thallium 3 = normal; 6 = fixed defect; 7= reversible defect

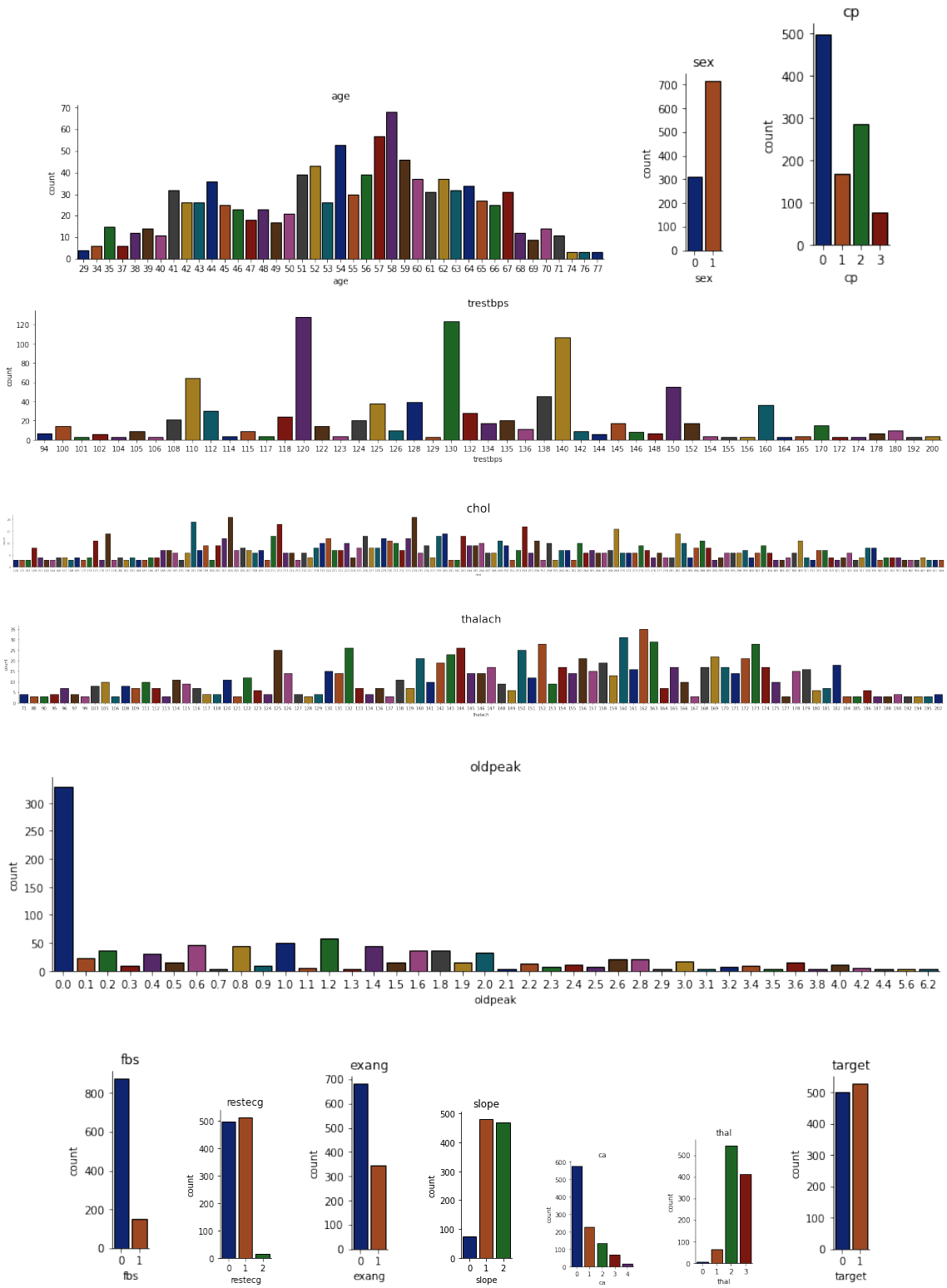


Figure 4.1: Data representation

Chapter 5

Research Methodology

5.1 Boosting

The terminology 'Boosting' is directed to a category of algorithms that transforms weak learners to strong learners. For achieving this every weak learner uses some methods. They are using average or using weighted average. Considering higher voted prediction is used as well.

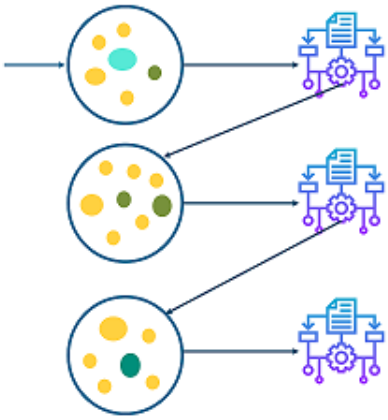


Figure 5.1: Boosting mechanism [13]

Boosting integrates weak learners (base learners) to form a strong learner or rule. When the base learning algorithm's implementation is done, a recently developed weak prediction rule is created which is many times an iterative process. The journey from these weak rules to bexome a strong prediction rule are affiliated by boosting algorithms after many iterations. There are some steps which are given below [2]:
1st Step: All the distributions are drawn by the weak or base learner and each observation is distributed with equal weight or attention.

2nd Step: If the first base learning algorithm causes any prediction error, then we focus on observations or related studies having prediction error. After that, the next base learning algorithm is applied for further calculations.

3rd Step: The 2nd step is iterated till the higher accuracy is attained or the maximum of base learning algorithm is reached.

$$\begin{array}{rl} \text{1st re-weighted data} & \xrightarrow{baseprocedure} \quad \hat{g}^{[1]}(.) \\ \text{2nd re-weighted data} & \xrightarrow{baseprocedure} \quad \hat{g}^{[2]}(.) \\ ..... & ..... \\ ..... & ..... \\ \text{Mth re-weighted data} & \xrightarrow{baseprocedure} \quad \hat{g}^{[M]}(.) \end{array}$$

$$aggregation : \hat{\int} A(.) = \sum_{m=1}^{M} \alpha_m \hat{g}^{[m]}(.). \tag{5.1}$$

Where, g(.) is the function with R values established on few data $(X1, Y1), ..., (Xn, Yn)$.

$$(X_1, Y_1), ..., (X_n, Y_n) \xrightarrow{baseprocedure} \hat{g}(.). \tag{5.2}$$

Finally, by amalgamating the outputs from the weak learners, it generates a strong learner that at least makes better prediction capacity of the system or model. Boosting pays a good amount of attention to those examples which are basically not in the right class or have excessive amount of errors by conducting weak rules.

5.1.1 Adaboost

The full form of AdaBoost is Adaptive boosting which is performed by combining several weak learners into a single strong learner. The couple of steps involved here that this adaptive boosting algorithm follows. Now, adaptive boosting starts with the assignment of an equal weight edge to all of the data points. It also starts with a decision stump (if a decision tree having one depth with two or more leaf nodes) for a single input feature that is drawn out.

The next step is the outputs that we have gotten from the first decision stump. It will be analyzed and if any observations are found misclassified, the higher weights will be assigned using the boosting concept. After that a new decision stump is drawn by considering the observations with the higher weights. That means missclassified data points are given a higher weight.

In the next step, another decision stump will be drawn and it will try to classify the data points by giving more importance to the data points with better weight. Again, if there are any observations that are misclassified then they are given higher weight and this procedure will be persistent. It will keep looping until all the observations get the right class.

Some weights are going to be initialized for separate sample points [2]: $w_i^{[0]} = 1/n$ for $i = 1, ..., n$.
Set $M = 0$.
Increasing M by 1. and the base procedure needs to be fitted for the weighted data, i.e., the weights are being used by a weighted fitting is done $w_i^[m-1]$, by yielding the classifier $\hat{g}^{[m]}(.)$
Now the weighted needs to be measured in-sample miss-classification rate

$$err^{[m]} = \sum_{i=1}^{n} W_i^{[m-1]} I(Y_i \neq \hat{g}^{[m]}(X_i)) / \sum_{i=1}^{n} w_i^{[m-1]}, \tag{5.3}$$

$$\alpha^{[m]} = \log(\frac{1 - err^{[m]}}{err^{[m]}}) \tag{5.4}$$

and up-date the weights

$$\hat{W}_i = W_i^[m-1] exp(\alpha^m I(Y_i \neq \hat{g}^m(X_i))). \tag{5.5}$$

$w_i^m = \hat{w}_i / \sum_{j=1}^{n} \hat{w}j$.Steps 2 and 3 need to be iterated until $m = m_{stop}$ and build the aggregated classifier by weighted majority voting:

$$\hat{\int} AdaBoost(x) = argmin_{y\epsilon(0.1)} \sum_{m=1}^{m_{stop}} \alpha^{[m]} I(\hat{y}^{[m]}(x) = y). \tag{5.6}$$

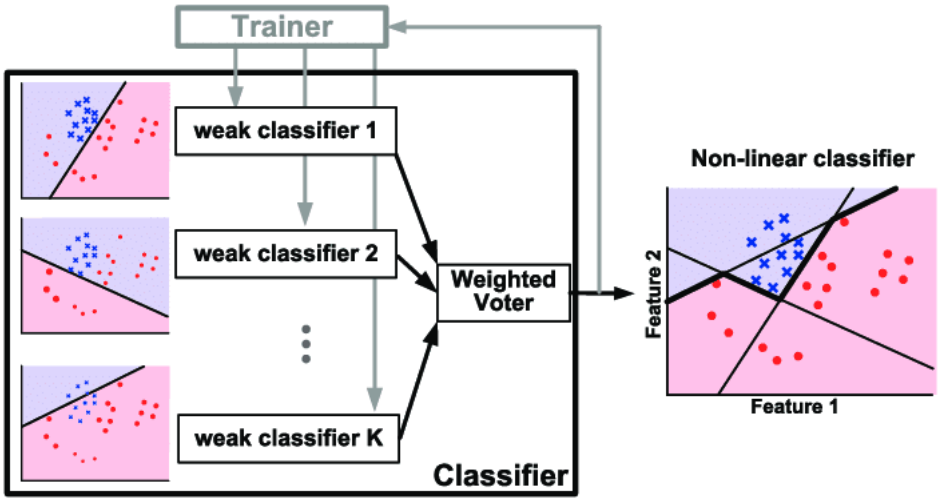


Figure 5.2: Adaboost working mechanism [19]

Therefore, the ultimate goal here is to ensure that each of the data points are classified into the exact classes. Adaptive boosting or adaboost can also be used for both classification and regression but it is more commonly seen in classification problems.

5.1.2 Gradient Boosting

The sequential model supports gradient boosting and it is supported by the symbol learning model as well. The base learners are generated consecutively in such a way where the previous one is usually less competent than present based learners. Basically the improvement of the general model sequentially happens with each and every iteration. Tn this type of boosting, there is a difference which is the outcomes' weights that are not classified do not add up. Here, we attempt to optimize the previous learners' loss function by including a new adaptive model. This new model adds weak learners so as to scale back the loss function.

Now the primary notion here is to beat the errors in the previous learner's prediction so that the next output gets better. There are three major components in this type of boosting (It has something known as the loss function). Loss function optimization is needed to reduce the error. The other type is that the weak learners are essential for calculating the predictions and then generating strong learners. After that, an additional model is important to make. The reason is it will regularize the loss function. As it will try to do so, the loss will be almost fix or the previous week learner will be formed by the errors. Input taken as [16]

training set $(x_i, y_i)_{i=1}^n$, an ability of differentiate loss function L(y,F(x)), number of iteration M. Then,
The model will be initialized with a constant value:

$$F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \tag{5.7}$$

For m=1 to M¡

1. compute so called pseudo-residuals:

$$r_{in} = -[\frac{\delta L(y_i, F(x_i))}{\delta F(x_i)}]_{F(x)=F_{m-1}(x)} for i = 1, ..., n. \tag{5.8}$$

   for i=1,....,n.

2. A base learner needs to be fitted (or weak learner, e.g tree) $h_m(x)$ to pseudo residuals, i.e. now it needs to be trained by using the training set $(x_i, y_i)_{i=1}^n$

3. Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma m = argmin_\gamma \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i + \gamma h_m(x_i))). \tag{5.9}$$

4. Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_x h_m(x) \tag{5.10}$$

   .
   Output $F_m(x)$.

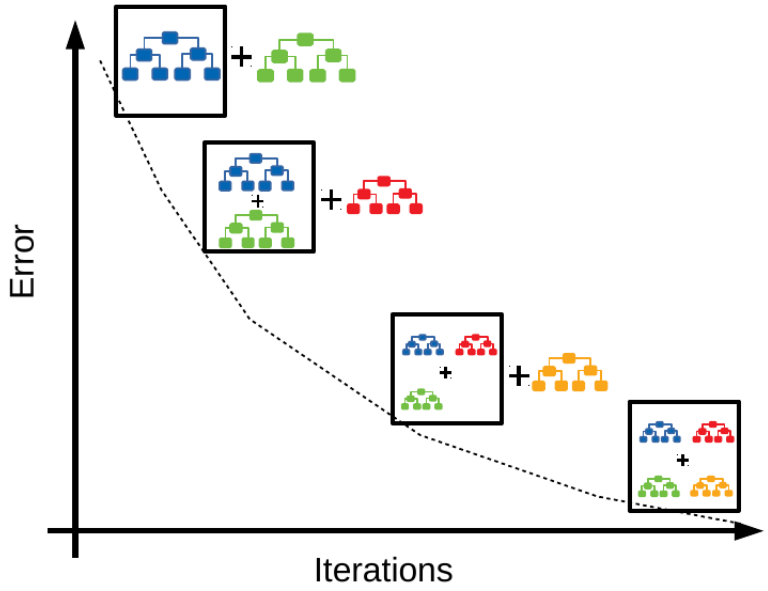


Figure 5.3: Gradient Boosting working mechanism [21]

Therefore, in distinction to the previous learner, we keep adding a model so that the loss function can be regulated. This is also used for both classification and regression problems.

5.1.3 XGBoost

XGBoost is a shorter form of Extreme Gradient Boosting which is an advanced version of gradient boosting. XGBoosting actually is a type of distributed machine learning community. To increase the speed and the efficiency of the competitions of the model performance are the main aim in this algorithm. The XGBoost was introduced for a few reasons. The output which is generated by the Gradient boosting algorithm at a very slow rate because the data set has a sequential analysis that takes a longer or more time. It basically boosts or extremely boosts the performance of the model.

Thus XGBoost is mainly going to focus on the speed and the model efficiency. In order to do this, it creates decision trees parallely means It supports parallelization. There is no non-parallel (sequential) modeling in this and it implements something known as distributed computing methods so that any complex and large modules can be evaluated. Moreover, core computing is used here in order to do analysis on a varied and huge dataset. Cache optimization is implemented as well in order to make best use of the hardware and of the resources in general.

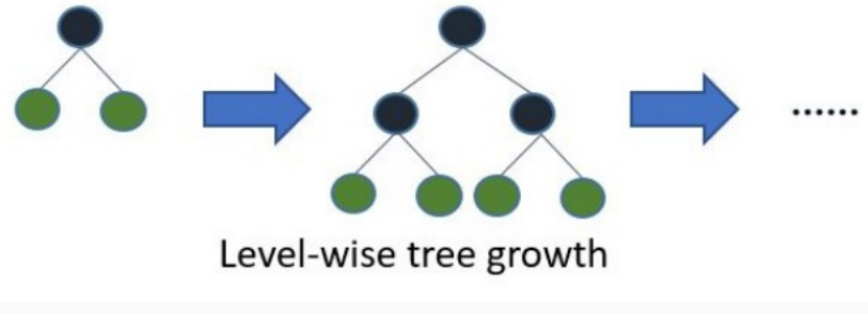


Figure 5.4: Level wise tree growth in XGBoost [28]

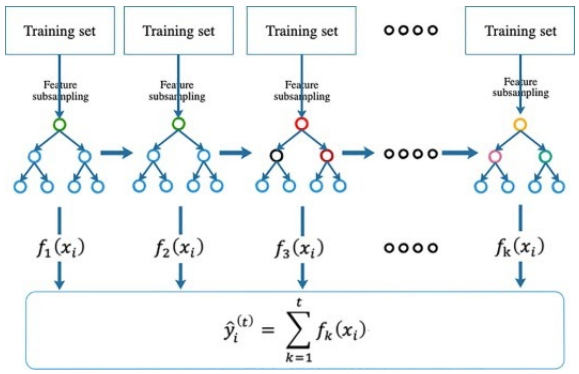


Figure 5.5: XGboost working mechanism [22]

Since XGBoost is an advanced option of Gradient Boosting, they almost share the same mathematical approach.

5.1.4 LightGBM

LightGBM is a shorter form of Light Gradient Boosting Machine which is an efficient gradient boosting framework that uses three based learning in a gradient boosting framework. The trees are built one after the other unlike in a random forest where the tree where a tree is created for each sample. One of the advantages of using lightGBM is it is fast. Training speed has advantages including its better accuracy, ability to handle missing values and the fact that it can be used both for regression and classification problems. The framework uses an algorithm which is leaf wise tree growth and it is an unbalanced tree unlike a level wise tree.

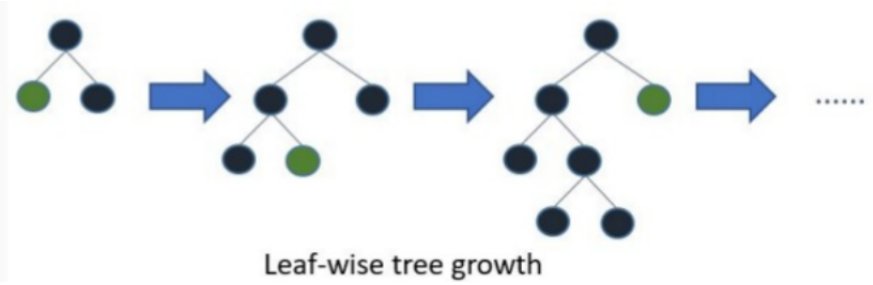


Figure 5.6: Leaf wise tree growth in LightGBM [28]

Since this also follows a gradient boosting algorithm, they almost share the same mathematical approach.

5.1.5 Catboost

The CatBoost is a short form of Categorical Boosting algorithm which actually makes use of decision trees and for each iteration it finds out the error. Basically, the algorithm looks for a path with the least error and passes to the next iteration. Initially, weight for every feature is set to be equal with some error 'e0'. Then the weights are adjusted and the same model will run on the new tree and then again it will get a new error which will be sent to its next iteration and then the weight will be again adjusted according to the error it has just gotten from the previous iteration that means with each iteration, the previous error is carried forward and the weight is adjusted such that the new found error is lower, such that e1¡e0 and at every iteration, it tries to reduce the error further until the error no longer decreases significantly. At each iteration the decision tree is built upon gradient fitting (chosen statistical model). This is how it will go until it finds out the best fit iteration. Once the best iteration is found, we do not attempt too many further iterations to deter over fit. Another interesting trait of this algorithm is that Catboost does not consider the part from which it shows over fitting unless it is not defined to take the entire iteration with over fitting.

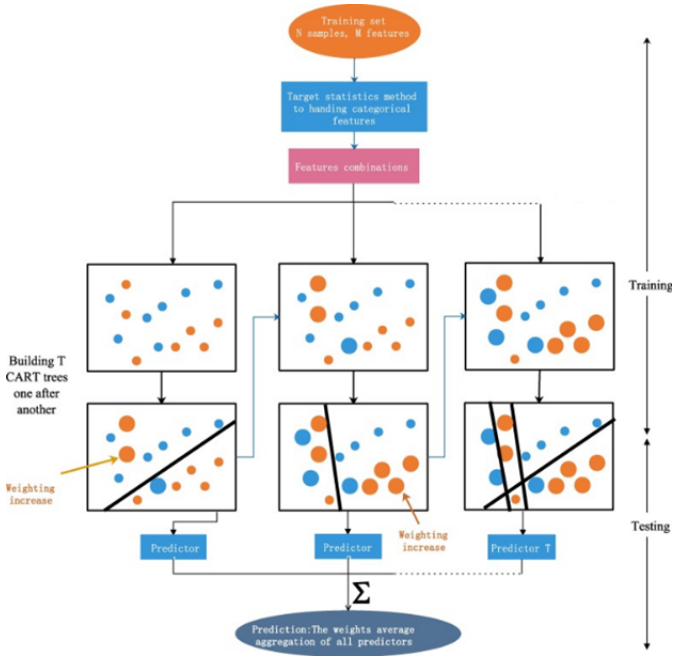


Figure 5.7: Catboost working mechanism [24]

As this follows a gradient boosting algorithm as well, they almost share the same mathematical approach which is already shown above.

5.2 Decision Tree

Decision Trees is a Machine Learning algorithm that can perform both classification and regression , and even multi output tasks. It is a white-box type ML algorithm. White box models are the type of models that share decision making logic, how they predict, what are the influencing variables. On the other hand with black-box type ML, like deep neural network, boosting etc., we can not know how it determines the output from the inputs.

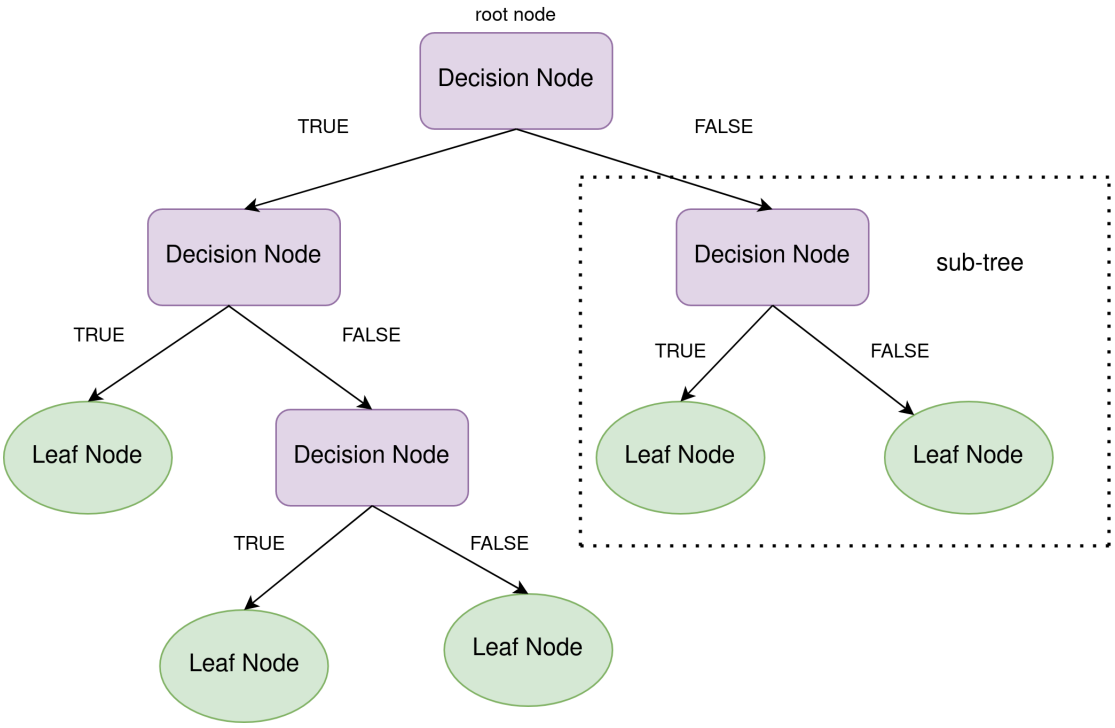


Figure 5.8: Basic structure of Decision Tree

A decision tree is a flowchart like tree structure where each internal node represents a condition based on an attribute called decision node, the branch represents the result of the condition, whether it is True or False. Finally each leaf node represents the output. Every decision tree first selects the best attribute using Attribute Selection Measures (ASM) to split the whole data set into two subsets. Making the attributes a decision node. The algorithm repeats this process recursively for each subset until all tuples belong to the same attribute value or there are no more remaining attributes. Attribute selection measure is a heuristic for selecting the splitting criterion that "best" separates a given data partition, D, of a class-labeled training tuple into individual classes [5]. It determines how the tuples at a given node are to be split.

We have used the CART training Algorithm to create our decision tree to predict heart disease. Classification and Regression Tree (CART) algorithm works by first splitting the data into two subsets. The splitting is done using a single feature k and threshold t_k (e.g. chest pain type 3.5). It chooses t_k and k by searching a pair (k,t_k) that produces the purest subsets.

$$J(k, t_k) = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right} \tag{5.11}$$

where, $G_{left/right}$ measures the impurity of left/right subset, it is the gini index

and $m_{left/right}$ is the number of instances in the left/right subset.
CART uses the gini index to measure impurity.

$$G_i = 1 - \sum_{k=1}^{n} P_i, k^2 \tag{5.12}$$

where, $P_{i,k}$ is the ratio of class k instances among the training instances in the i^{th} node [14].

5.3 Random forest

Random forest can be used for regression and classification. Random forest operates by constructing a multitude of decision trees. So a random forest is an assembly of decision trees. Each decision tree predicts a class for a data point and the class with highest frequency is assigned to the data point. While growing the trees, random forest adds additional randomness to the model. It searched for the best feature among a random subset of features, unlike the decision tree. In the case of a decision tree it searches for the most important feature on the whole training data set.

5.4 Support Vector Machines (SVM)

A support vector machine can perform linear or nonlinear classification and regression. SVM are well suited for classification of complex small or medium sized data set. Linear SVM Classification : The process of a SVM linear classification can be explained using figures.

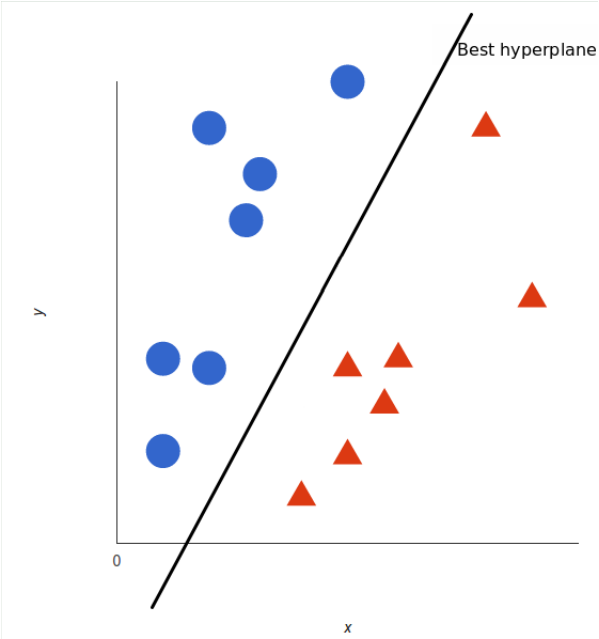


Figure 5.9: Best hyper plane [10]

If the data sets of the data are plotted on a plane and labled the SVM takes the data points and outputs a hyper plane that best separates the tags. The hyper plane separates the data sets belonging to different classes. There can be many hyper planes that divide the data sets into different classes. The best hyper plane that divides the classes in the one having a large distance from both classes. That is the main goal of SVM to find such best hyper planes.

5.5 K-Nearest Neighbour

KNN algorithms can be used for classification and regression. While training, the algorithm represents the data points as vectors in a multidimensional feature space, each with a class label. In the training phase of the algorithm the feature vectors and the class labels are sorted. Here the k is a user defined constant. On the classification part a test data point which is an unlabeled vector is classified by assigning the label which is most frequent among the k training data points nearest to the query point. The best choice of k depends upon the nature of a data set.

5.6 Logistic Regression

In machine learning there is one of the most popular and necessary classification algorithms called Logistic Regression. Logistic Regression is a classification algorithm, applied when the value of the target variable is unconditional in nature. Statisticians developed the Sigmoid function which is another naming of logistic function which was elaborate features of people rising in the environment, growing fast and maxing out at the bearing receptivity of the environment. Any actual-valued number can be accepted by this S-shaped curve and into a value between 0 and 1 it can be designed, but those limits are not exact. By not fitting a straight line or hyper plane,

the logistic function is used by logistic regression model to strain a linear equation which has the output between 0 and 1. The logistic function is defined as [26]:

$$logistic(\eta) = \frac{1}{1 + exp(-\eta)} \tag{5.13}$$

using weights Input values (x) are added linearly or coefficient value to calculate an output value (y). There is a main difference from linear regression which is the output value which is modeled instead of a numeric value there is a binary value (0 or 1).

$$L = \frac{e^{r1+r2*x}}{1 + e^{r1+r2*x}} \tag{5.14}$$

In this equation, L is the proposed output, r1 is the intercept term and for the single input value (x) r2 is the Coefficient. There is a connected r coefficient (a constant real value) in every column of the input data which has to be learned from the training data.

There is a probabilistic framework called maximum likelihood estimation which is used to calculate the parameters in a logistic regression model. Various machine learning algorithms use maximum likelihood estimation which is a popular knowledgeable algorithm, the assumptions about the distribution of the data is generally taken by this. A value very close to 1 would be calculated by the Best coefficient results in a model for the default class and for the other class there is a value very close to 0. For logistic regression the sense for maximum likelihood is that values for the coefficients are asked by a search system that reduce the mistakes in the probabilities estimated by the model to those in the data.

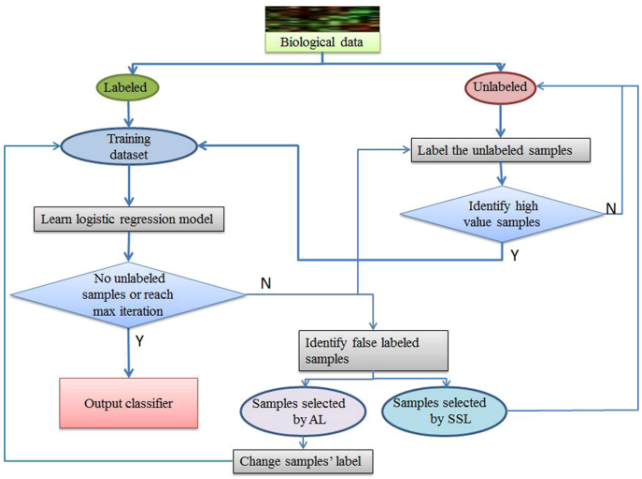


Figure 5.10: Workflow of Logistic Regression [11]

5.7 Linear Regression

Linear regression has a vast application in machine learning and statistical problems and also very easy to understand. Linear regression pretend to model the relationship between two variables by fitting a linear equation to executes data. In this algorithm, one variable is considered to be an independent variable, and the other is considered to be a dependent variable. This algorithm is a observed algorithm of machine learning and the predicted output of this algorithm has a continuous slope and is continuous. Linear regression is used for the prediction for both real and numerical variables such as price, salary, age etc. However, the linear regression algorithm represents the linear relationships among the variables. In other words, it represents how the dependent variable changes with respect to independent variables.We can represent a linear regression equation as [25]:

$$y = a + bX \tag{5.15}$$

In the above equation X represents independent variables and Y represents dependent variables, b is the linear coefficient and a is the interception line.
Generally Linear regression model provides a straight line which has a slope representing the changes of dependent variable over independent variable.
Below is the linear regression graph:

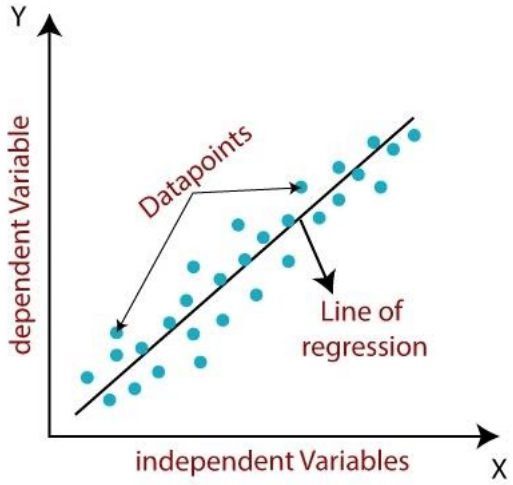


Figure 5.11: Relationship between variable in Linear Regression [25]

There are two types of linear regression.

Simple linear regression: For predicting the value of a dependent variable if there uses only one independent variable, then this is called simple linear regression.

Multiple linear regression : For predicting the value of a dependent variable if there uses more than one independent variable, then this is called multiple linear regression.

5.8 Naive Bayes

Naive Bayes is a machine learning algorithm that depends and predicts based on probability which depends on the Bayes theorem, utilized in a wide assortment of order errands. Regular applications of Naive Bayes incorporate separating spam, arranging records, estimation expectation, etc. The name Naive is utilized as it accepts the highlights in the models that is free of one another. The highlights that is changing the estimation of one element, does not impact on the estimation of any of the other highlights utilized on the calculation. This algorithm works with Bayes theorem which can be illustrated as [14]:

$$P(A|B) = P(B|A).P(B)/P(A) \tag{5.16}$$

Here, in the above equation, P(A—B) represents posterior probability, P(B—A) represents the likelihood of the event, P(B) represents the class prior probability and P(A) represents the predictor prior probability. A is a dependent feature vector (of size n) where:

$$A = (a1, a2, a3.....an) \tag{5.17}$$

The diverse NB classifiers vary fundamentally by the suppositions which are made in regards to the dissemination of the conditional probability P(Ai—B) [18]. There is a huge scope to Naive Bayes. Since it is a probabilistic model, the calculation can be coded up effectively and the predictions can be made in real-time very quickly. Moreover, it is effectively adaptable, fast learner and is traditionally used for the calculation of decisions in real-world applications, that are required to react to a client's solicitations quickly.

Chapter 6

Result Analysis

We have used a total of 12 algorithms in our observations. Classifiers are most commonly used here. For the first observation, our target column was target (Heart disease is present or not) which falls under binary classification. Here are the results of the algorithms used.

Table 6.1: Table of Results for heart disease

Algorithms	**Accuracy**	**Classification error**	**Precision**	**Recall**	**F1 score**
CatBoost	99.12%	0.88%	99%	99%	99%
lightGMB	99.12%	0.88%	99%	99%	99%
XGBoost	99.12%	0.88%	99%	99%	99%
Gradient	98.23%	1.77%	98%	98%	98%
AdaBoost	91.15%	8.85%	91%	91%	91%
Decesion Tree	100%	0%	100%	100%	100%
Random Forest	100%	0%	100%	100%	100%
SVM	85.21%	14.79%	86%	86%	85%
KNN	100%	0%	100%	100%	100%
Naive Bayes	84.44%	15.56%	85%	84%	84%
Logistic Regression	87.16%	12.84%	88%	87%	87%
Linear Regression	84.79%	15.21%	85%	84%	84%

Based on Table 6.1, some graphical analysis on the results, the comparison among the used algorithms are shown. Accuracy, classification error and F1 score are taken to compare. We did not take Precision and Recall individually because F1 score is the average of these two and the F1 score would almost give similar results, if Precision or Recall was represented in a graph.

6.1 Accuracy, Classification Error and F1 Score according to the Algorithms

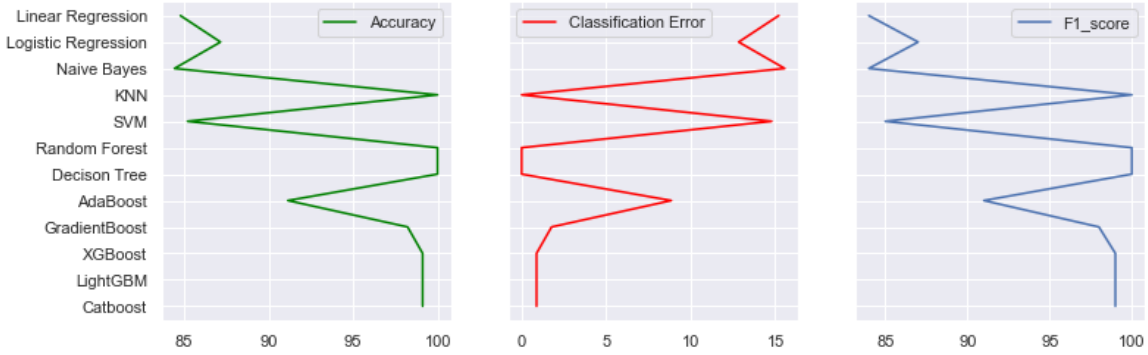


Figure 6.1: Accuracy, Classification and F1_Score of every algorithms

From Figure 6.1, Decision tree along with Random Forest and KNN are giving 100% accuracy. Naive Bayes is showing the least accuracy here.

Again, these three are giving 100% F1 scores. Naive Bayes is showing the least F1 scores here.

No classification error was given by them as well whereas Naive Bayes is showing the highest.

6.2 Comparison between two results

Here, we have found almost similar work in a paper where [17] 9 algorithms in a dataset. 7 out of 9 algorithms are matched with our work. That is why those 7 algorithms with their result are given below in Table 6.2 for further comparison.

Table 6.2: Table of results of the compared work [17]

Algorithms	**Accuracy**	**Classification error**	**Precision**	**F1 score**
Naive Bayes	75.8%	24.2%	90.5%	84.5%
Linear Regression	85.1%	14.9%	88.8%	91.6%
Logistic Regression	82.9%	12.6%	90.7%	92.6%
Decision Tree	85%	15%	86%	91.8%
Random Forest	86.1%	13.9%	87.1%	92.4%
Gradient Boost	78.3%	21.7%	94.1%	86.8%
SVM	86.1%	13.9%	90.2%	84.4%

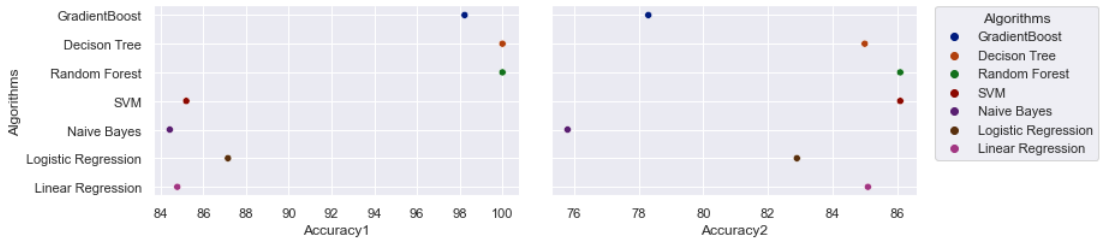


Figure 6.2: Comparison between both of the Accuracy

Accuracy1 is implemented by us and Accuracy2 is taken from the paper [17]. From Figure 6.2, except SVM and Linear Regression, all of our algorithms have given better accuracy. The difference of SVM and Linear Regression is in fraction.

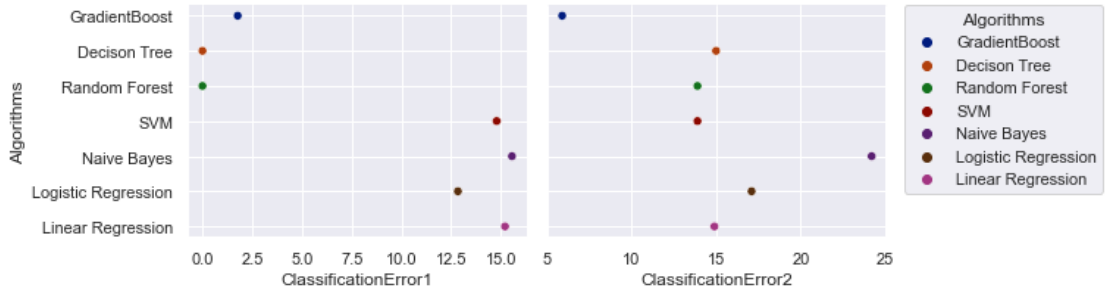


Figure 6.3: Comparison between both of the Classification Error

Again, ClassificationError1 is implemented by us and ClassificationError2 is taken from the paper [17]. From Figure 6.3, except SVM and Linear Regression, all of our algorithms have given less classification error. The difference of SVM and Linear Regression is in fraction.

AdaBoost, XGBoost, LightGBM, CatBoost, KNN are not present in their paper [17]. Although the data they used is different and less than us, we share the same attributes with the same quantity. However, we have 3 algorithms (Decision tree, Random Forest, KNN) which give a 100% accuracy rate. So this can be our standard and other algorithms can be compared with any of these three algorithms.

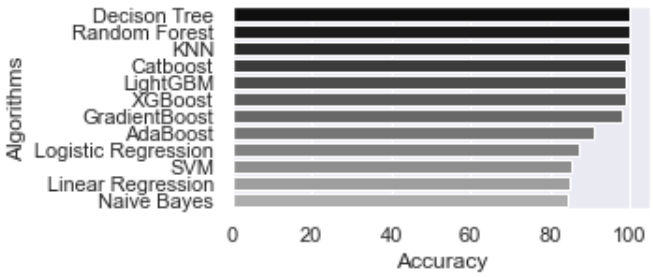


Figure 6.4: Sorted Algorithms based on the performance of predicting heart disease

We can see from the Figure 6.4, boosting algorithms are doing better. If we observe a bit more, we can see that CatBoost is doing better among other boosting algorithms. Naive Bayes's accuracy is the worst in this dataset.

Again, in the feature importance, Catboost predicted the exact features like 'ca', 'cp' which were also present in the algorithms which have given 100% accuracy.

6.3 Feature importance for heart disease

Importance of Features | Less risk of Heart Disease | More risk of Heart Disease | Flow of featured values

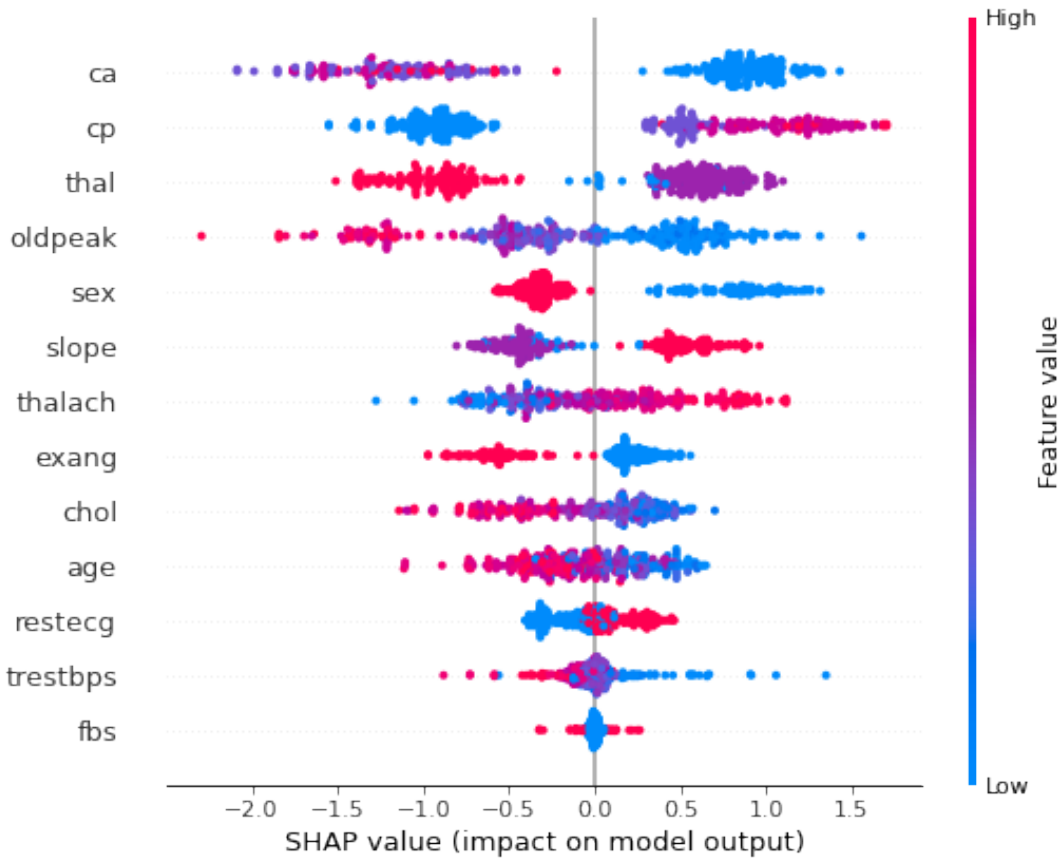


Figure 6.5: Feature Importance for Heart Disease

In figure 6.5, the first column is for the importance of features in the dataset according to this algorithm. For instance, according to the graph, ca has the greatest impact on making a prediction, then cp, after that thal and it continues till fbs. So, it is visible that fbs results have the least impact on the prediction.
From the graph, it is possible to say which data are discrete and which are continuous. For instance, binary value has only two colors in the representation, whereas continuous values have different shades from red to blue.
Moreover, each feature has individual impacts which can be found from the graph. For example, from Sex feature, 0(low) is representing females and 1(High) is representing male and low values(0) are placed on the left which means females suffer less than males, as higher values(1) are placed on the right side of the middle line(0.0).

According to the graph, if it is going toward left means less impact on causing or having heart Disease and going to the right means more impact.
From every algorithm used in our work, feature importances are shown here in a tabular manner in Table 6.3. We can see, almost every algorithm gives different different feature importance for the heart disease dataset, We took the top three feature importance of every algorithm.

Table 6.3: Feature Importance Table

Algorithms	**1st**	**2nd**	**3rd**
CatBoost	ca	cp	thal
lightGBM	age	chol	thalach
XGBoost	age	chol	thalach
Gradient	cp	thal	ca
AdaBoost	age	chol	threstbps
Decesion Tree	cp	ca	chol
Random Forest	cp	ca	oldpeak
SVM	cp	restecg	slope
Logistic Regression	cp	restecg	slope
Linear Regression	cp	restecg	slope

From the Table 6.3, ca(number of major coronary artery vessels colored by fluoroscopy), age, cp(chest pain type), these are the most repeated features in the first position means for causing a heart disease, these are the most prominent features.

6.4 Target Column : ca

For this particular observation, ca became the target column. The reason behind choosing 'ca' is our motive is to find out what factors or features are there to cause blockage of the coronary arteries. After running all the algorithms, the results we got are given below.

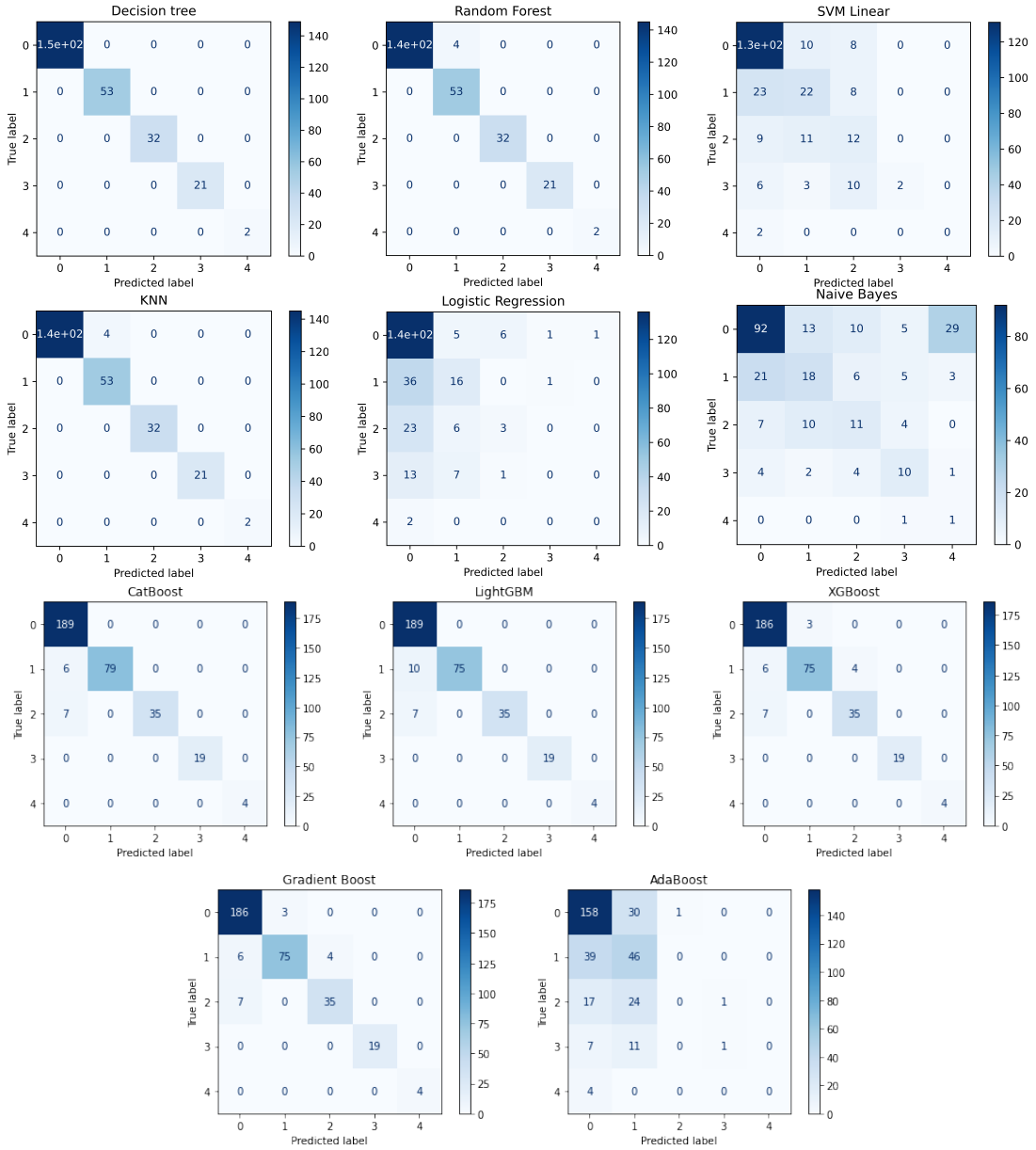


Figure 6.6: Confusion Matrix of every algorithms

In figure 6.6, the confusion matrices are shown where

	Actually positive(1)	Actually negative(0)
predicted positive (1)	True positives	False positives
predicted negative (0)	False negatives	True negatives

Table 6.4: Table of Results

Algorithms	Accuracy	Classification error	Precision	Recall	F1 score
CatBoost	96.17%	3.83%	99%	95%	97%
lightGMB	94.99%	5.01%	98%	94%	96%
XGBoost	94.10%	5.90%	96%	94%	95%
Gradient	94.10%	5.90%	96%	94%	95%
AdaBoost	60.47%	39.53%	32%	29%	27%
Decesion Tree	100%	0%	100%	100%	100%
Random Forest	98.44%	1.56%	99%	98%	98%
SVM	64.98%	35.02%	66%	65%	62%
KNN	98.44%	1.56%	99%	98%	98%
Naive Bayes	51.36%	48.64%	59%	51%	55%
Logistic Regression	60.31%	39.69%	51%	60%	53%
Linear Regression	77.78%	22.22%	68%	77%	70%

6.5 Accuracy according to the Algorithms

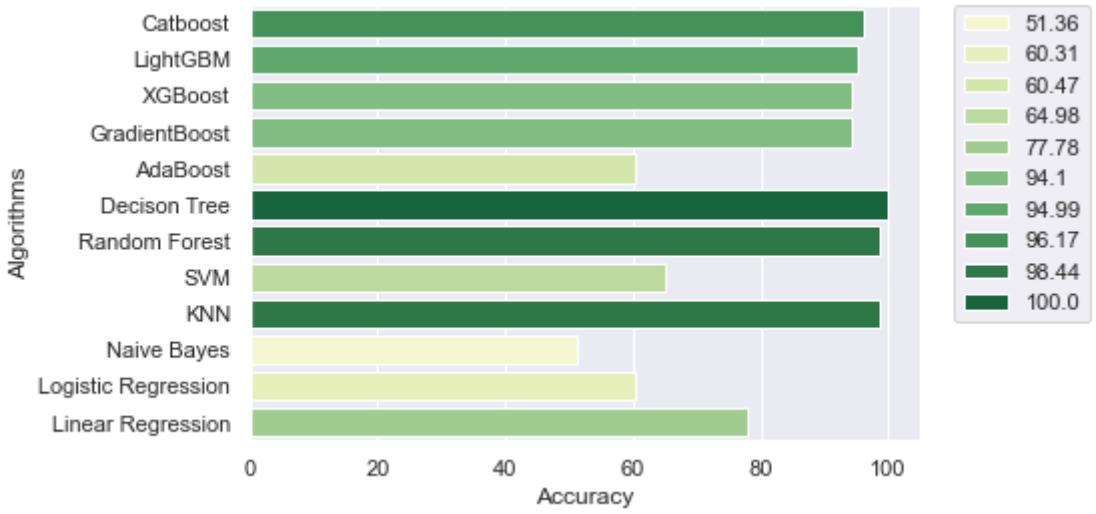


Figure 6.7: Accuracy based on the Algorithms' performance

From Figure 6.7, only the Decision Tree has given 100% accuracy and Naive Bayes has given the least accuracy.

6.6 Classification Error according to the Algorithms

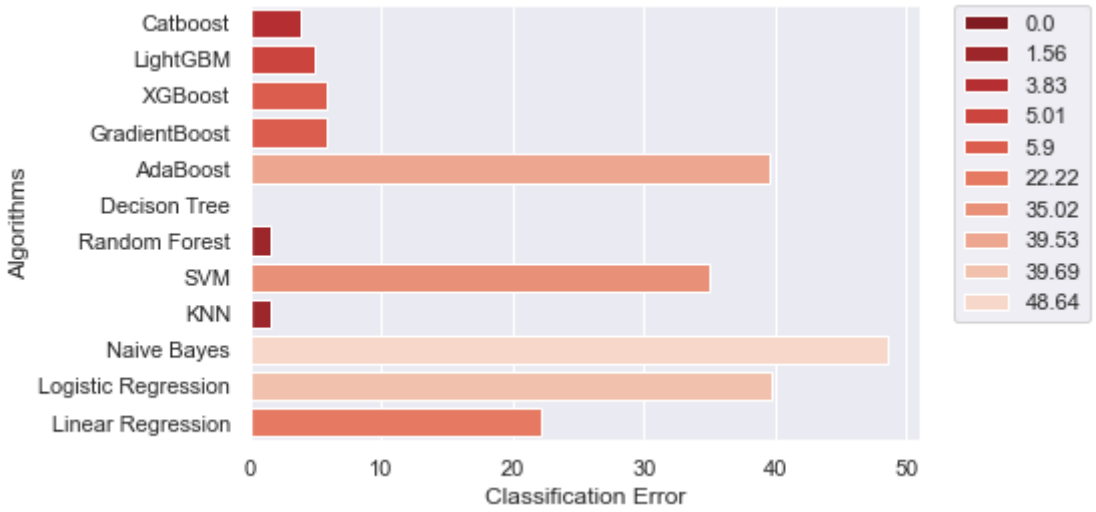


Figure 6.8: Classification Error based on the Algorithms' performance

From Figure 6.8, we can see there is not classification error in Decision tree and the highest classification error is shown in Naive Bayes.

6.7 F1 scores according to the Algorithms

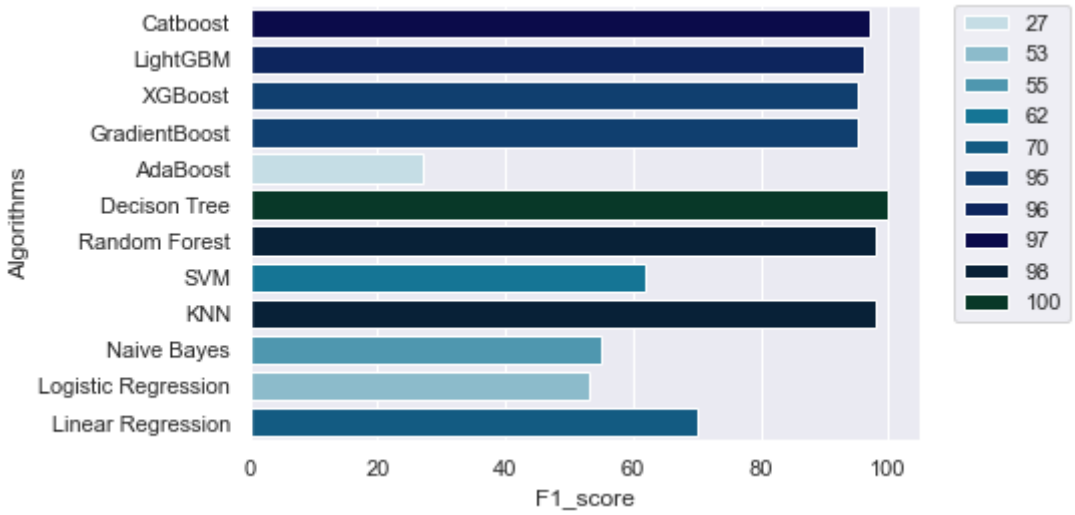


Figure 6.9: $F1_{s}cores based on the Algorithms' performance$

Again, in Figure 6.9, only the Decision Tree has given 100% accuracy and Naive Bayes has given the least.
Here, if we sort the accuracy, it can be seen how well the algorithms performed for the 'ca' target column.

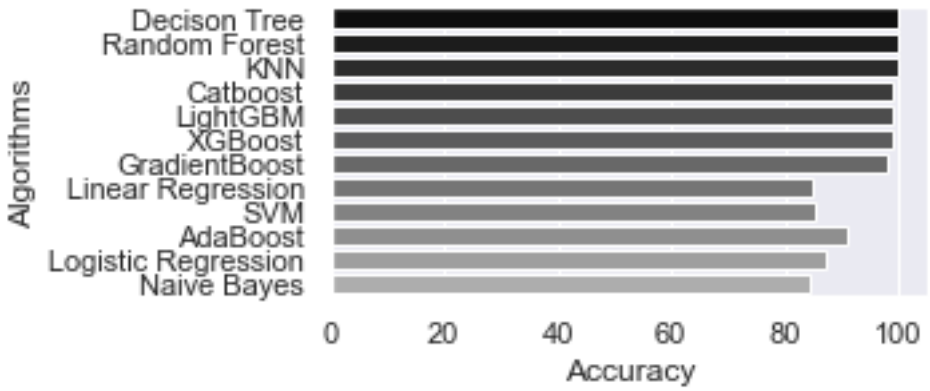


Figure 6.10: Sorted Algorithms based on the performance

Still, after sorting the algorithm based on their performance showm in Figure 6.10, Decision Tree gave the 100% accuracy and the boosting algorithms performed better than non boosting algorithms here and among all boosting algorithms, CatBoost still gave finer results.
Since only the Decision Tree algorithm gave us 100% accuracy, we can make it as a standard. We can observe what features are important for other algorithms and can match them with the standard algorithm.

6.8 Feature importance when the target column is ca

Table 6.5: Feature Importance

Algorithms	**1st**	**2nd**	**3rd**
CatBoost	age	target	thalach
lightGBM	chol	thalach	age
XGBoost	chol	thalach	age
Gradient	age	chol	target
AdaBoost	chol	age	target
Decesion Tree	age	chol	target
Random Forest	age	chol	thalach
SVM	target	restecg	thal
Logistic Regression	target	cp	restecg
Linear Regression	fbs	slope	sex

6.9 Features that came multiple times

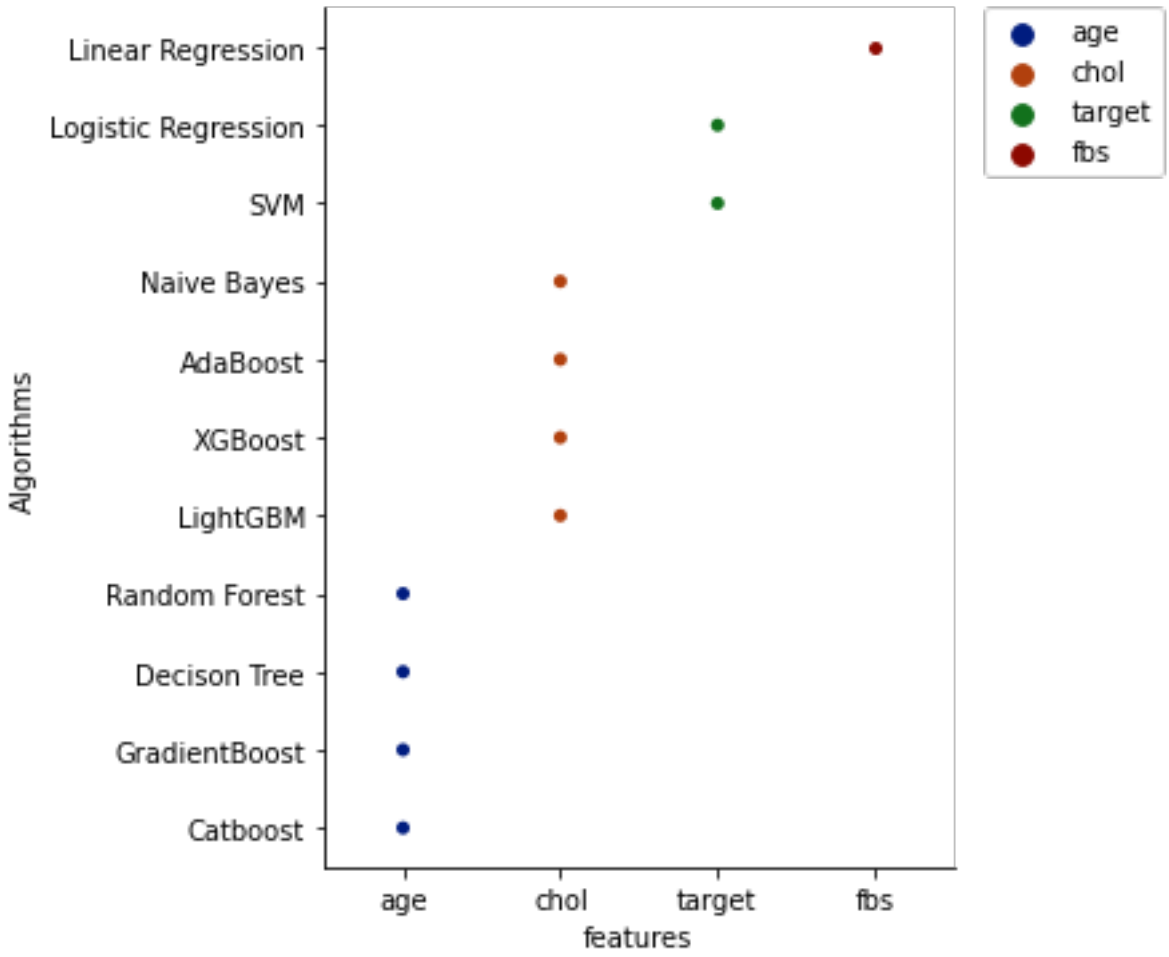


Figure 6.11: First Feature Importance of every Algorithms

We can see from the graph in Figure 6.11, age and chol (cholesterol) has a similar and higher effect on the model. Whereas the target (heart disease is there or not) has less and the fbs(fasting blood sugar) has the least effect.
Now, those people who have 0 in their ca column, those rows are omitted because our motive is to figure out for those who have major vessels blocked in their heart, how related those ca with the first important features and how these features behave with different number of vessels blockage in their heart.

Therefore, most important features that occurred multiple times are age and chol. The data density among them is given for further analysis.

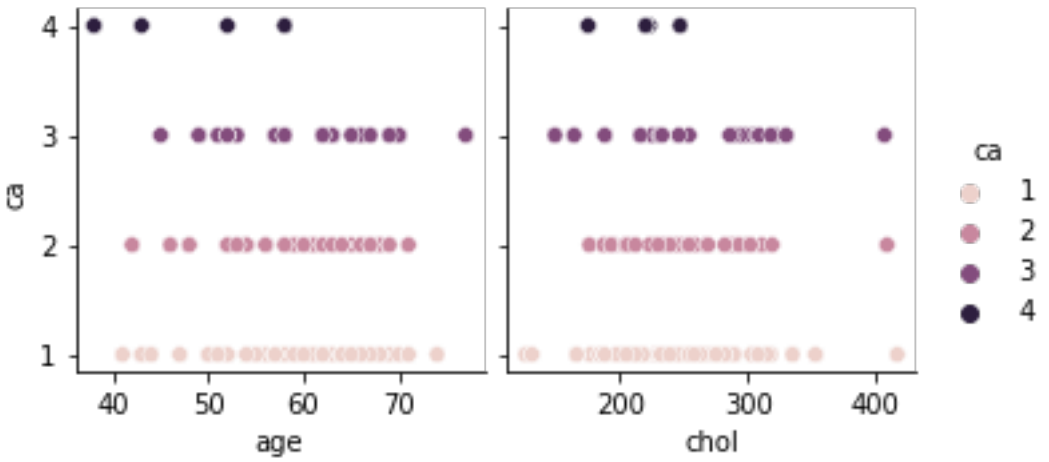


Figure 6.12: Most appeared Features' data density

From the graph shown in Figure 6.12, when ca is 1 means one of the coronary arteries becomes completely blocked, has more people in age, chol. When ca is 2 means two blockages in the heart, has also a lot of patients but compared to ca=1, it has less both in age, chol. Same thing happened in ca=3,4; ca=3 has less than the previous and ca=4 has the least people.
Therefore, we can observe that the flow of age and cholesterol has a decreasing mode from 1 to 4 for the ca feature.

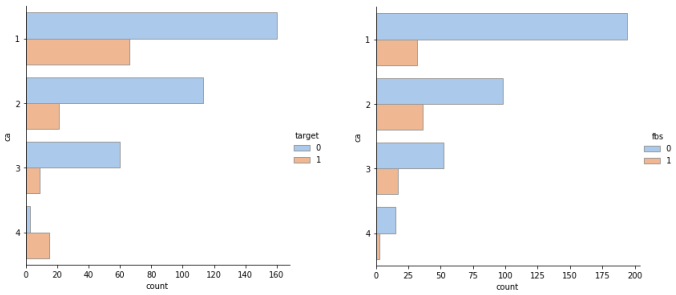


Figure 6.13: Data count or patients according to target & fbps Feature

From Figure 6.13, the ca vs. target graph, we can see that having 4 blocks of the coronary arteries has less target 0 and high target 1 that means it has the highest tendency to cause heart disease. However, in ca=1,2,3, the target 0 is higher than target 1, which means it can also cause heart disease as target 1 is present in the graph but we can say they have less risk than ca=4.
In ca vs. fbs and according to the dataset, there is not a very strong relationship between them because even for less than 120 mg/dl people have blockage(s) in their heart.

6.10 Age and Cholesterol patient count based on Coronary Artery blockage

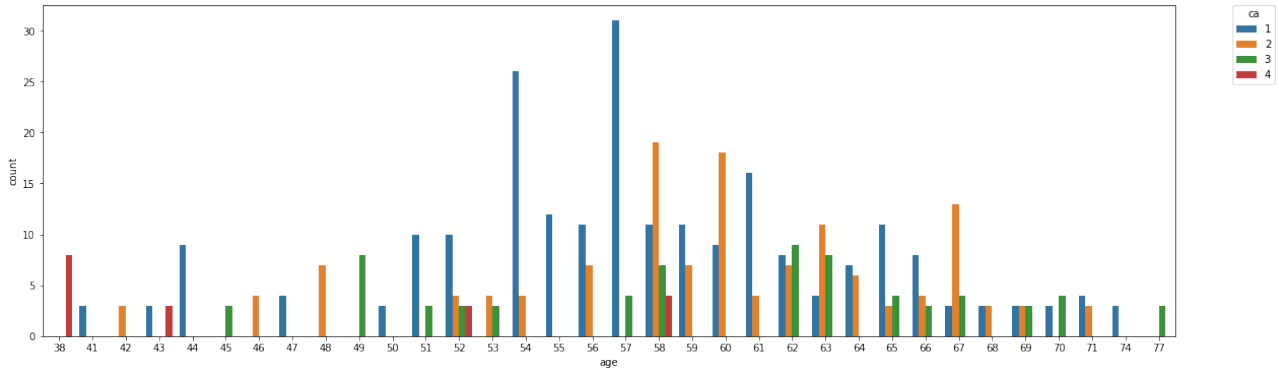


Figure 6.14: age vs ca count

from figure 6.14, most of the patients are from 50 to 70 years. however, according to the data 39 years old patients have maximum heart blocks.

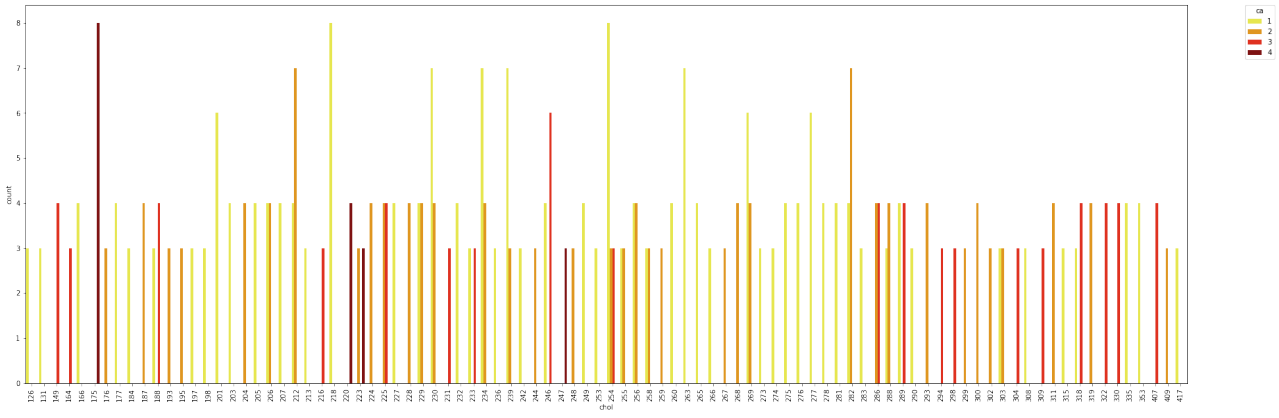


Figure 6.15: chol vs ca count

from figure 6.15, most of the patients are in danger when they have more than 200 mg/dl. however, according to the data patients who have 175 mg/dl recorded the most heart blocks.

According to the result we can say when a person at his/her middle age needs to control the cholesterol the most because the stressful lifestyle can be damaging to the health of our heart.

6.11 Interrelationship among ca, age and cholesterol

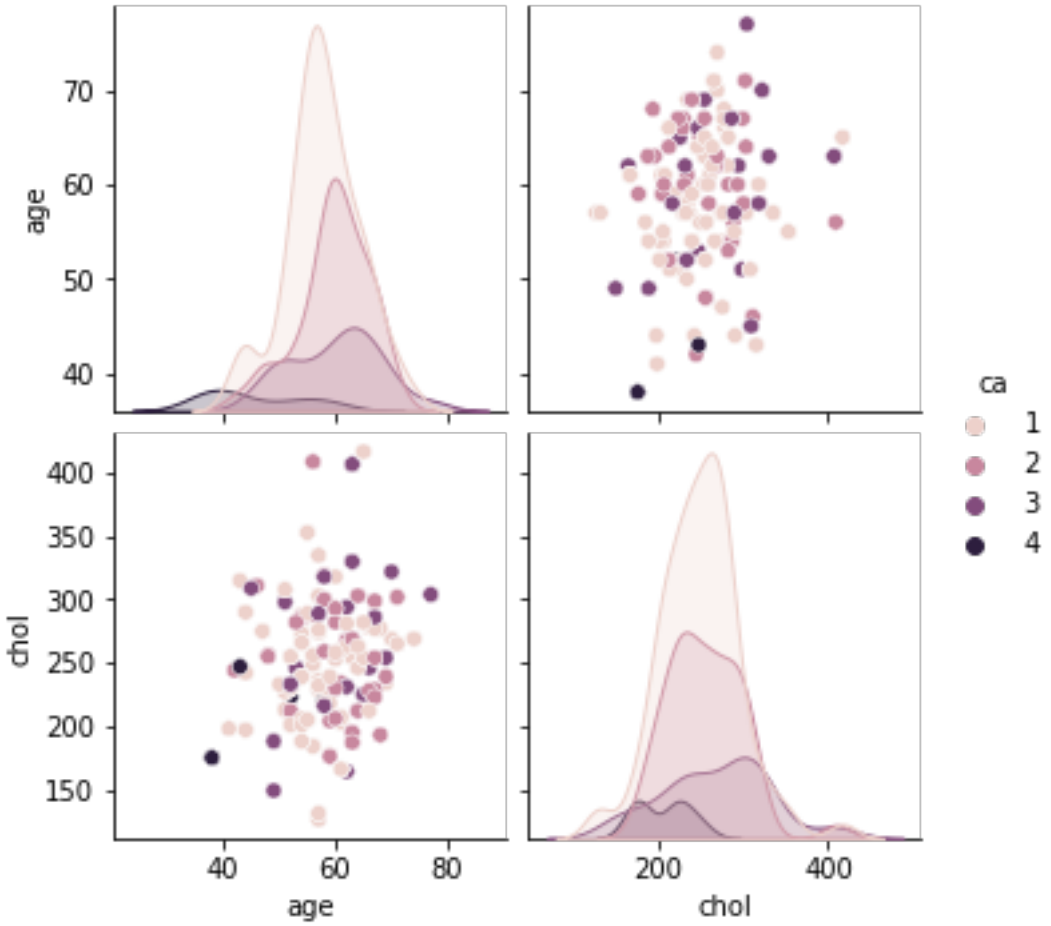


Figure 6.16: Relationship among ca, age and chol Feature

As it is already observed that the relation among ca, age and cholesterol is stronger, we are now going to observe the interrelationship among them. Here in Figure 6.16, we can see that from the age of 50 to 70 years and the cholesterol level from 200 to 300 mg/dl, has the tendency to get at least one heart vessel blocked (ca=1). Then for the age from 60 to 70 years and the cholesterol level again from 200 to 300 mg/dl, it is showing the density of dots more for ca=2. Now for the age again from 60 to 70 years and having the cholesterol level from 300 to 350 mg/dl is showing more tendency to have at least three heart vessels blocked (ca=3). Lastly, when ca=4, is less and also very mystifying, it shows people at the age of their 40's with the cholesterol level around 200 or more have 4 blocks in their heart. The number is less because many people can not survive till 4 blockages, some of them may die before reaching that many blocks in their heart.

6.12 Further analysis on ca, age, chol with cp

In coronary artery disease, the vessels of heart get blocked due to many reasons, increase of cholesterol level is one of them. Also, scarring of the heart tissue as people get older is another significant reason for heart blocks. No matter what cause it is, most of the coronary artery patients have to go through the chest pain
Now,due to the heart blockages, reduced blood flow to the heart causes chest pain which is known as angina. so we can say it is a symptom of coronary artery disease. when a person feel tightness, pressure, heaviness or pain in his/her chest, it is often expressed as angina.
Shortness of breath, dizziness, nausea, fatigue, sweating are some symptoms of angina.
There are three possible types of chest pain

- Atypical Angina: chest pain feels like angina without angina symptoms.
- typical Angina: chest pain with angina symptoms.
- Non-typical Angina: other type of chest pain.

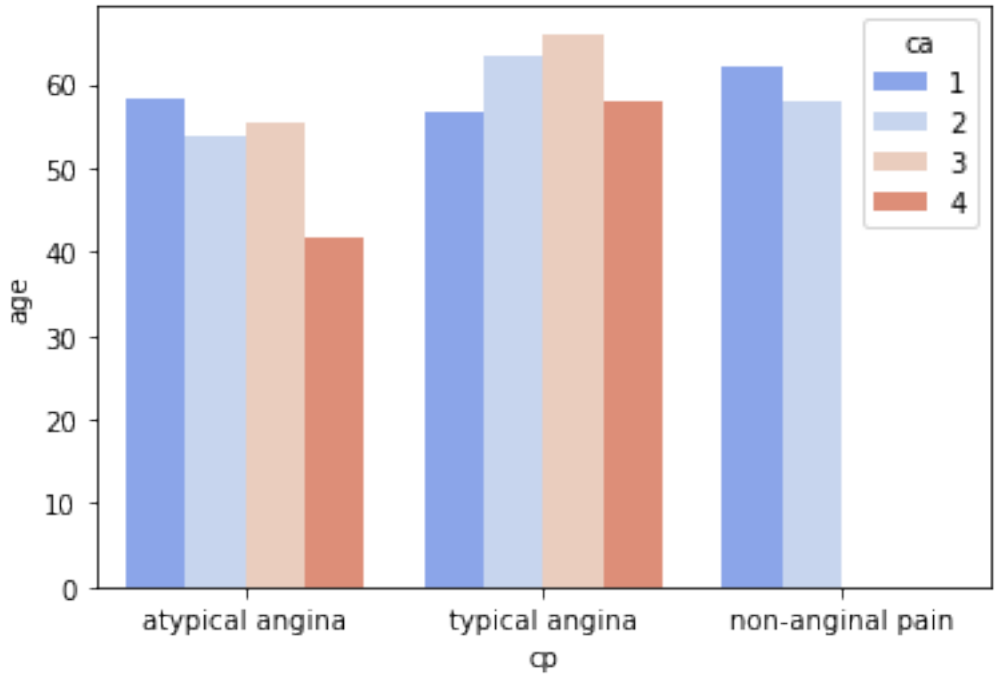


Figure 6.17: Analysis on ca, age and cp

From figure 6.17, it is observed that most of the patients aged from 50 to 60 years feel typical angina chest pain due to heart blocks than other type of chest pain. From the graph, it is also observed that higher ca has the highest tendency to have typical angina type chest pain.

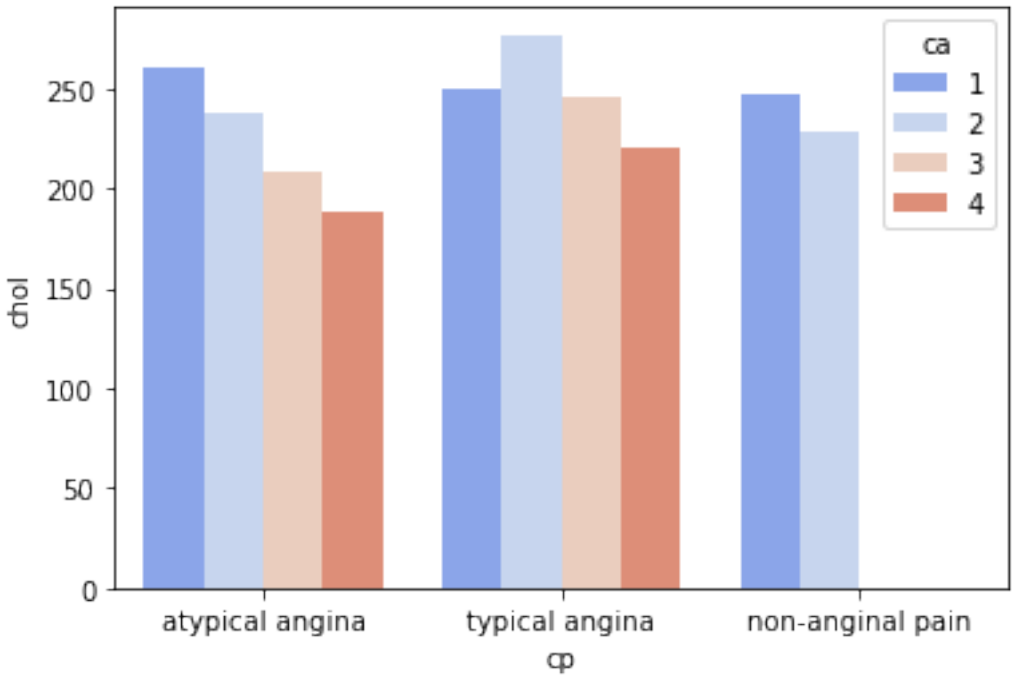


Figure 6.18: Analysis on ca, chol and cp

From figure 6.18, it is again observed that more than 200 mg/dl cholesterol level has patients with typical angina type chest pain due to heart blockages.

Therefore, we can say that most of the time, the presence of the coronary artery disease can show more angina type of pain when it can be with or without the symptoms of angina.

Presence of coronary artery disease:
typical angina chest pain > atypical angina chest pain > non-typical angina chest pain

To sum up, age and cholesterol both play a significant role to cause heart blocks which is also known as coronary artery disease. In addition, a person with blocks in his/her heart feel angina type of pain customarily. Therefore, it is clear that heart related problems cause more angina chest pain with or without symptoms.

Chapter 7

Conclusion

Our heart is one among the foremost vital organs keeping us alive. It is important to have a healthy lifestyle to keep our heart in a good shape. Cardiovascular diseases are the amount one explanation for death within the world. Cardiovascular diseases include blood vessels such as coronary heart diseases, disorders of the heart, rheumatic heart disease, cerebrovascular disease, blocks of coronary arteries in heart and other conditions. One of the most important uses of machine learning in the healthcare field is diagnosis of diseases. Use of machine learning in heart disease prediction can help provide suitable treatment to the patients in time. To utilize machine learning tools in this field hospitals must continue collecting good quality data related to heart disease. Machine learning is a growing field of data science. We have new machine learning algorithms being introduced to the world quite often. It is important to keep testing these algorithms to achieve high accuracy in predicting heart disease.

In this paper we have compared various algorithms to identify algorithms with high accuracy for heart disease prediction and heart vessel blockage prediction. Furthermore, we have focused on the features contributing to heart vessel blockage. From our study we see that patients of middle to old age are at risk of having heart vessel blockage. Moreover, high cholesterol contributes to heart vessel blockage. Therefore, monitoring patients' cholesterol level will decrease the risk of blockage. Using machine learning we can predict heart vessel blockage. Through getting the result patients can be alert before having a heart issue which will eventually save valuable time for both the doctors and patients. Use of machine learning on diagnosing will also decrease medical expenses. Therefore, research on machine learning and heart disease prediction is appreciated.

Bibliography

[1] M. R. Islam, M. U. Chowdhury, *et al.*, "Spam filtering using ml algorithms," in *Proceedings of the IADIS international conference WWW/Internet 2005*, IADIS Press, 2005, pp. 419–426.

[2] (Oct. 17, 2007), [Online]. Available: https://web.stanford.edu/~hastie/Papers/buehlmann.pdf (visited on 05/31/2021).

[3] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *2008 IEEE/ACS International Conference on Computer Systems and Applications*, 2008, pp. 108–115. DOI: 10.1109/AICCSA.2008.4493524.

[4] M. N. Banu and B. Gomathy, "Disease predicting system using data mining techniques," *International Journal of Technical Research and Applications*, vol. 1, no. 5, pp. 41–45, 2013.

[5] R. Devi and K. Nirmala, "Construction of decision tree: Attribute selection measures," *International Journal of Advancements in Research & Technology*, vol. 2, no. 4, pp. 343–347, 2013.

[6] M. A. Khaleel, S. K. Pradham, G. Dash, *et al.*, "A survey of data mining techniques on medical data for finding locally frequent diseases," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 8, pp. 149–153, 2013.

[7] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, "Hybrid intelligent modeling schemes for heart disease classification," *Applied Soft Computing*, vol. 14, pp. 47–52, 2014.

[8] N. Alshurafa, C. Sideris, M. Pourhomayoun, H. Kalantarian, M. Sarrafzadeh, and J.-A. Eastwood, "Remote health monitoring outcome success prediction using baseline and first month intervention data," *IEEE journal of biomedical and health informatics*, vol. 21, no. 2, pp. 507–514, 2016.

[9] Wiharto, H. Kusnanto, and Herianto, "Intelligence system for diagnosis level of coronary heart disease with k-star algorithm," *Healthcare informatics research*, vol. 22, no. 1, p. 30, 2016.

[10] (Jun. 22, 2017). "An introduction to support vector machines (svm)," [Online]. Available: https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/ (visited on 05/30/2021).

[11] (Aug. 29, 2018). "A novel logistic regression model combining semi-supervised learning and active learning for disease classification — scientific reports," [Online]. Available: https://www.nature.com/articles/s41598-018-31395-5?fbclid=IwAR1u9o01CnEm2vH6aIGtXC4C_Oi87iULOdwdXGHGHho2WkFjAs_2blasufI (visited on 05/30/2021).

[12] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of heart disease using machine learning," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, 2018, pp. 1275–1278.

[13] (Jun. 28, 2019). "A beginners guide to boosting machine learning algorithms — edureka," [Online]. Available: https://www.edureka.co/blog/boosting-machine-learning/ (visited on 05/30/2021).

[14] A. Géron, "Hands-on machine learning with scikit-learn, keras tensorflow," 1005 Gravenstein Highway North, Sebastopol, CA 95472.: O'Reilly Media, Inc., 2019.

[15] S. Krishnan and S. Geetha, "Prediction of heart disease using machine learning algorithms.," in *2019 1st international conference on innovations in information and communication technology (ICIICT)*, IEEE, 2019, pp. 1–5.

[16] K. K. Mahto. (Feb. 25, 2019). "Demystifying maths of gradient boosting — by krishna kumar mahto — towards data science," [Online]. Available: https://towardsdatascience.com/demystifying-maths-of-gradient-boosting-bd5715e82b7c (visited on 05/31/2021).

[17] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81 542–81 554, 2019.

[18] D. M. van Rijmenam. (Aug. 14, 2019). "7 steps to machine learning: How to prepare for an automated future — by dr mark van rijmenam — dataseries — medium," [Online]. Available: https://medium.com/dataseries/7-steps-to-machine-learning-how-to-prepare-for-an-automated-future-78c7918cb35d (visited on 05/30/2021).

[19] (Sep. 29, 2020). "Illustration of adaboost algorithm for creating a strong classifier... — download scientific diagram," [Online]. Available: https://www.researchgate.net/figure/Illustration-of-AdaBoost-algorithm-for-creating-a-strong-classifier-based-on-multiple_fig9_288699540 (visited on 05/30/2021).

[20] (May 30, 2021). "Applications of machine learning - javatpoint," [Online]. Available: https://www.javatpoint.com/applications-of-machine-learning (visited on 05/30/2021).

[21] (May 20, 2021). "Block-distributed gradient boosted trees," [Online]. Available: http://tvas.me/articles/2019/08/26/Block-Distributed-Gradient-Boosted-Trees.html (visited on 05/30/2021).

[22] (May 30, 2021). "Building safety monitoring based on extreme gradient boosting in distributed optical fiber sensing - sciencedirect," [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1068520019307564 (visited on 05/30/2021).

[23] (Jun. 8, 2021). "Cardiovascular diseases," [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases/ (visited on 06/08/2021).

[24] (May 30, 2021). "Evaluation of catboost method for prediction of reference evapotranspiration in humid regions - sciencedirect," [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0022169419304251 (visited on 05/30/2021).

[25] (May 30, 2021). "Linear regression in machine learning - javatpoint," [Online]. Available: https://www.javatpoint.com/linear-regression-in-machine-learning?fbclid=IwAR0Rl3qAQnFKnNeW-k_ibHgzl6XOFIhPtEOC_kegICjRfsRIO0xmIaetZaA (visited on 05/30/2021).

[26] C. Molnar. (May 30, 2021). "4.2 logistic regression — interpretable machine learning," [Online]. Available: https://christophm.github.io/interpretable-ml-book/logistic.html?fbclid=IwAR33CRylxmUAlH3maL0QpOnF9AbfywK0zLkwwjqzv5wnfethV (visited on 05/31/2021).

[27] (Mar. 12, 2021). "Supervised vs. unsupervised learning: What's the difference? — ibm," [Online]. Available: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning (visited on 05/30/2021).

[28] (May 30, 2021). "Xgboost - decision trees: Leaf-wise (best-first) and level-wise tree traverse - data science stack exchange," [Online]. Available: https://datascience.stackexchange.com/questions/26699/decision-trees-leaf-wise-best-first-and-level-wise-tree-traverse (visited on 05/30/2021).

[29] N. Rajpal, "Decision support system for heart di decision support system for heart disease diagnosis using neural network sease diagnosis using neural network niti guru* anil dahiya,"

Comments From Panel

There were 4 panel members.

1. Mr. Rubayat Ahmed Khan
2. Ms. Ipshita Bonhi Upoma
3. Mr. Tanvir Rahman
4. Ms. Ahanaf Hassan Rodoshi

Mr. Tanvir Rahman & Mr. Rubayat Ahmed Khan: Age and Cholesterol high indicate hear disease which can be told normally. This does not need any machine learning techniques.

Ms. Ahanaf Hassan Rodoshi & Ms. Ipshita Bonhi Upoma: Your workings are fine but you can also detect other diseases from these so that it can be more meaningful.

In a summery, Almost everyone said this work is too mainstream. However, this work would be perfect, if we could add new features like occupation, obesity, symptoms, etc.