## **Project Title: Heart Disease Prediction using Data Mining Techniques**

Contents

[**Project Title: Heart Disease Prediction using Data Mining Techniques** 1](#_Toc119542039)

[**ABSTRACT:** 2](#_Toc119542040)

[**INTRODUCTION:** 2](#_Toc119542041)

[**RELATED WORK:** 3](#_Toc119542042)

[**METHODOLOGY:** 6](#_Toc119542044)

[**DATA SOURCE:** 7](#_Toc119542045)

[**Features of Data source:** 7](#_Toc119542046)

[**Steps** 7](#_Toc119542047)

[1. Exploratory Data Analysis 7](#_Toc119542048)

[3. Machine learning algorithm 7](#_Toc119542049)

[**Exploratory Data Analysis (EDA):** 9](#_Toc119542050)

[**Data Visualization:** 9](#_Toc119542051)

[**Machine Learning algorithm used:** 11](#_Toc119542052)

[**Decision Tree** 12](#_Toc119542053)

[**Naive Bayes** 13](#_Toc119542054)

[**Logistic Regression:** 15](#_Toc119542055)

[**Conclusion:** 16](#_Toc119542056)

[**References:** 17](#_Toc119542057)

## **ABSTRACT:**

Heart disease increases day by day so it’s essential to know at what stage it becomes serious or dangerous for the patient. As many machines and highly expensive doctors are available in the market but our aim is to make a cost-free and highly accurate machine learning model to predict with few attributes and faster efficiency the risk of having heart disease. We are using an open-source dataset from Kaggle. Today, heart disease is the condition that poses the greatest risk to life. Heart disease identification is a significant problem. Although there are tools that can forecast heart attacks, they are either expensive or ineffective when used to estimate the likelihood of heart attacks in humans.

## **INTRODUCTION:**

The quantity of electronic health information that healthcare facilities are gathering has advanced dramatically. When something comes to patient care, accuracy is especially crucial, and computerizing this vast volume of data raises the standard of the entire system. But how do medical professionals efficiently sort through all the data? This is an area that data mining has been proven to be quite successful. In order to uncover hidden patterns and relationships from huge databases, data mining integrates statistical analysis, machine learning, and database technology [1]. Data mining known as an inter-disciplinary computer science field and statistics with the overarching objective of extracting useful information within a dataset and structuring it for later use. As a result, the goal of this research project is to anticipate outcomes by using data mining techniques on health care data.

Cardiovascular diseases (CVDs) are now the main cause of death. More than 80% of deaths from CVD are caused by cardiovascular diseases [2]. The provision of high-quality services at affordable prices is one of the biggest challenges facing healthcare companies. Correct patient diagnosis and good therapy delivery are indicators of high-quality care. Instead of using the knowledge-rich data that is hidden in the database, doctors frequently base their clinical decisions on their perception and practice. This technique highlights unwelcome prejudices, errors, and exorbitant medical costs that have an impact on the standard of care provided to patients [3]. According to the WHO, heart disease is the leading cause of mortality worldwide, killing around 17.9 million people year, or about 31% of all fatalities (World Health Organization). Heart disease, often known as cardiovascular disease, is a term used to describe a variety of illnesses affecting the blood arteries and heart, such as cerebrovascular disease, rheumatic heart disease, and other disorders. Strokes and heart attacks account for four out of every five heart disease fatalities. The mortality rate and total consequences can be reduced by early identification of heart attacks.

Unsupervised learning discovers intrinsic structures or hidden patterns in input data, while supervised learning develops a model using current inputs and outputs from dataset to predict future outputs. This study aims to forecast heart problems using supervised machine learning techniques. Determine the relationship between input attributes and a target attribute using supervised algorithms. A structure known as a model serves as a representation of the relationship that is disclosed.

The two primary algorithms in supervised learning are the classification model and the regression model. This work focuses on a categorization model in this instance. Rather than evaluating continuous numbers, classification focuses with categorizing observations. This study compares the performance of various classification algorithms for heart disease prediction, including SVM, Nave Bayes, and KNN. The patient data are continuously categorized and forecasted in all techniques. The patient's activity is continuously observed, and if any changes take place, the patient and physician are advised of the disease risk level. Thanks to machine learning techniques and computer technologies, doctors can diagnose cardiac illnesses earlier than before.

In the modern world, we have a lot of data, so we can use a variety of machine-learning algorithms to examine the data and look for hidden patterns. Now that these datasets are available, analysis and information extraction are possible. We may thus predict heart illness in its early stages by using machine learning algorithms on this enormous amount of data to discover traits that we will extract from datasets. A number of machine learning techniques, such as logistic regression, support vector machines, and Naive Bayes. We may use them to categorize whether a person has cardiovascular disease or not after using them to extract features from datasets in order to predict heart disease. We may compare them to find the most effective algorithm for properly forecasting heart disease. Various algorithms will produce results with different levels of accuracy.

## **RELATED WORK:**

Heart disease diagnosis has been the subject of numerous scientific studies. They tried a variety of data-mining methods for diagnosis and got diverse outcomes using various approaches.

Authors in [4], compared a number of ML algorithms, including SVM, Naive Bayes, and MLP, as well as selected specific features from the datasets for which they obtained varying levels of accuracy when selecting various features. For example, when they selected all 14 features from the datasets, they obtained the highest accuracy using naive bayes (81%), whereas when they selected only specific features, they obtained the highest accuracy using SVM. They concluded from their research that the most crucial stage in increasing the precision of machine learning algorithms is feature selection.

In [5], authors discusses the dataset that contains 14 distinct features, but only 11 of them were employed in the Nave Bayes and Decision Tree machine learning algorithms to predict heart disease. In their research, they used the WEKA tool, which allowed them to obtain 82% accuracy for Nave Bayes and 84% accuracy for decision trees. In [6], authors employed Nave Bayes, Decision trees, and neural networks. The heart disease is predicted most accurately by the neural network algorithm out of the three.

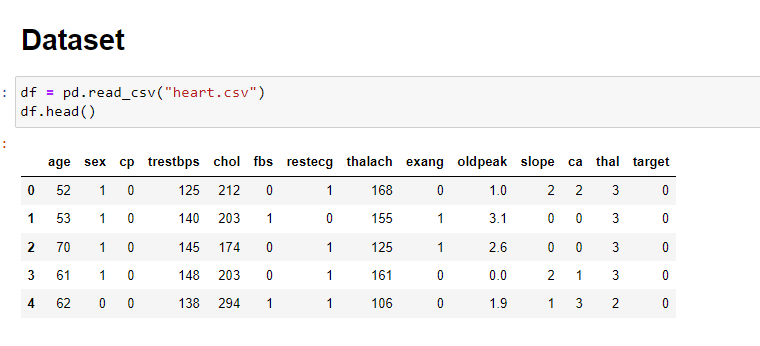
Decision Trees, Naive Bayes, and neural networks are the three techniques used by authors in [7]. Their experimental findings demonstrate that each methodology has a special advantage in achieving the specified mining aims. A neural network-based prediction system for assessing heart risk is presented by researchers in [8]. The results obtained have demonstrated that the suggested diagnostic system is capable of accurately predicting the risk of cardiac illnesses.

The accuracy of the system for predicting heart disease that the authors in [9] reported employed K-mean clustering and the MAFIA algorithm was 89%. Mirpouya Mirmozaffari et al. [10] demonstrated that among the greatest performance algorithms, the random tree approach provides the highest reliability and lowest errors. The authors of [11] intended to use weighted association Apriori algorithm, MAFIA algorithm, and decision trees as part of their data mining classification techniques. In [12], authors introduced a data mining model that was created using a Random Forest classifier to improve forecast accuracy and look at various CHD-related events.

Utilizing conditional independence, the Naive Bayes classification algorithm assumes that a parameter for a given class is distinct from the values of other variables. A Nave Bayes algorithm for high risk heart disease prediction was suggested in [13] for web-based health care detection. Data that has been previously processed has been used as a training set. In that paper, categorization and prediction were two phases that were covered. In the categorization phase, preprocessing takes place. Data cleaning, standardization, and reduction are all part of the preprocessing. A training set is created depending on the disease type, and a test set is created based on the questions, both of which happen during the prediction phase. The doctor receives the anticipated outcomes.

ANN is a computational model or mathematical model that is utilized for biological purposes. It is frequently referred to as a "neural network." It mimics the biological neurological system, in other words. A neural network-based prediction technique for heart disease has been put out in [14]. The input, hidden, and output layer make up its three primary layers. The input layer receives the input, while the output layer receives the output. The production that was produced vs what was anticipated is then compared. To identify the fault and modify the weight between the output and earlier hidden layers, back propagated has been used. The forward procedure is initiated after back propagation is finished, and it is continue till the error is minimized.

KNN is a not a parametric technique used for regression and classification. KNN is the simplest machine learning algorithm when compared to other algorithms. In the feature space, this method uses K-closet training examples. The user-defined constant K is used in this algorithm. By selecting the most chronic constant value from the K-training samples closest to the point, the test data are categorized. The KNN has a strong consistency outcome, according to the literature. A decision tree uses a tree-like structure to construct categorization models. It separates the data set into smaller subsets while also developing the decision tree linked with it in stages.

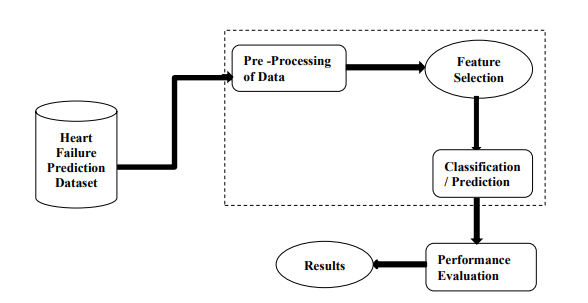


### **Figure 1.0: Dataset**

The decision tree employs a top-down methodology. The data set serves as the decision tree's root, and each leaf represents a subset of the dataset. In [15], a hybrid method for predicting heart disease risk level was put forward. KNN technique and ID3 are combined to create a hybrid algorithm. For the prediction of cardiac disease, these algorithms are employed. The KNN algorithm, often known as a preprocessed algorithm, is used to preprocess the data. After the preprocessing, data is used as a training set, and a tree structure has been created to classify the data. The classifier uses the ID3 algorithm to predict cardiac disease. KNN Algorithm is used to classify the wrong values.

## **METHODOLOGY:**

Data gathering and the selection of critical qualities are the first steps in this project. The relevant data is then preprocessed into the format needed. After that, the data is split into training and testing data. The methods are used, and the training data is used to train the model. Using test data, the system is tested to determine its correctness. We use the Python Pandas package to convert some category variables in the data to numerical values. Prior to doing the data cleaning process, we use the isnull() method to see if the dataset had any missing values.



**Figure 1.1: System Overview**

## **DATA SOURCE:**

It is possible to predict a patient's risk of developing a cardiac condition using medical characteristics such as age, sex, blood pressure, and blood sugar. It makes it possible to learn important things, such as patterns, and links between medical parameters that are associated with heart disease. Various heart disorders can be identified using publicly accessible heart disease data.

## **Features of Data source:**

1. Age: Years of age

2. sex: Sex (Male = 1; Female = 0)

3. cp: kind of chest pain

Value 4: asymptomatic resting blood pressure (trestbps) (in mm Hg on admission to the hospital)

5. chol: Serum cholestoral in mg/dl

6. fbs: (1 = true; 0 = false) (fasting blood sugar > 120 mg/dl)

7. restECG: findings of the resting electrocardiogram.

8. thalach: reached maximal heart rate

9. exang: exacerbation of angina due to activity (1 = yes; 0 = no)

10. oldpeak = ST depression brought on by activity compared to rest and other too.

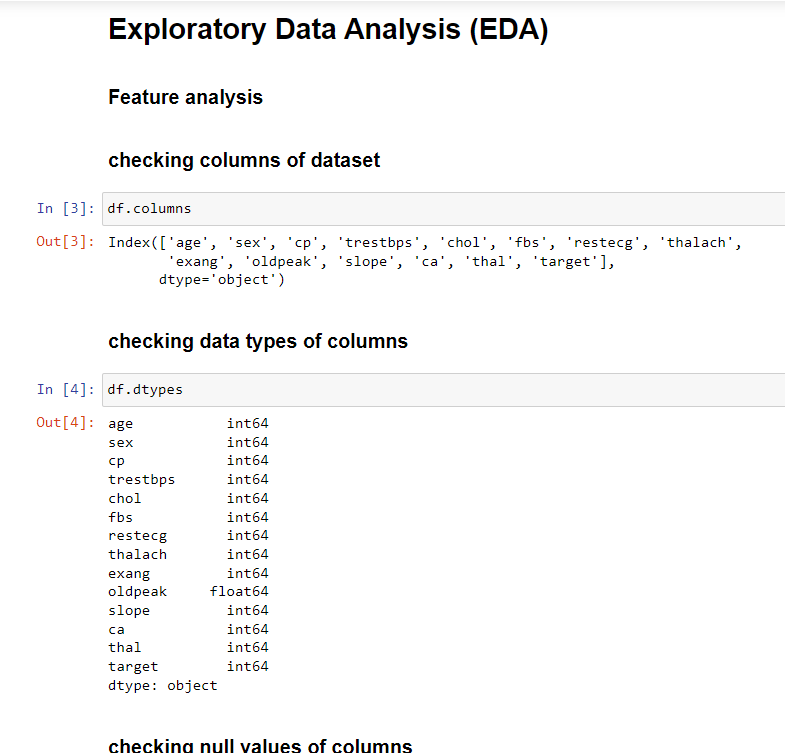
## **Steps**

# Exploratory Data Analysis

1. Data Visualization.

# Machine learning algorithm

1. Conclusion



**Figure 2: EDA**

## **Exploratory Data Analysis (EDA):**

EDA is a method for analyzing data sets to highlight their key features, frequently utilizing graphs or charts and other types of data visualization. EDA differs from traditional hypothesis testing in that it is primarily used to explore what the information can tell us besides the formal modelling. A statistical model can be utilized or not. EDA's major goal is to encourage data analysis before coming to any conclusions. It can assist in finding glaring errors, better understanding data patterns, spotting outliers or unusual occurrences, and discovering intriguing relationships between the variables.

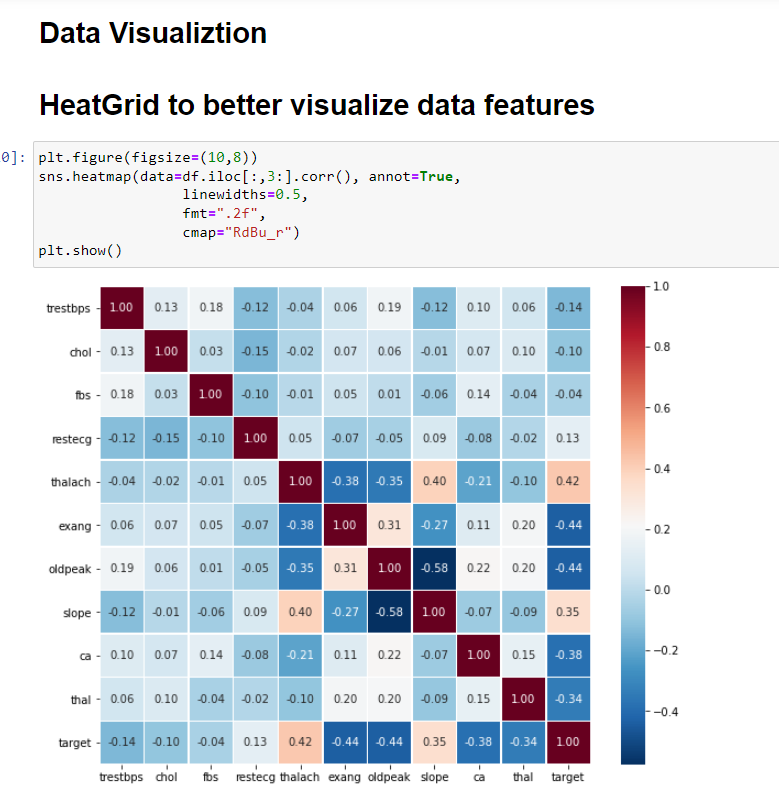
To make sure the findings they create are reliable and relevant to any targeted business objectives and goals, data scientists can employ exploratory analysis. EDA aids stakeholders by assuring them that they are posing the proper questions. Categorical variables, standard deviations, and confidence intervals are all topics that EDA may aid with. EDA's features can then be employed for more complex data analysis or modelling, including machine learning, when it is finished and conclusions have been formed from it.

Using pandas we are performing EDA on our dataset. We are checking null values, data imbalancing, data types, and correlations between the target and other feathers. After that, we are plotting different charts to get similar to datasets so it can help in making an ML model. In figure 1 we can clearly see heart attacks are more common in men as compared to women. In figure 2 with the help of a histogram, we can clearly say that heart attack chances increase with age and unfortunately people with age 55-60 are at high-risk.

## **Data Visualization:**

The graphic display of data and information is known as data visualization. Data visualization tools offer an easy approach to observe and analyze trends, anomalies, and data patterns by utilizing visual elements like graphs, charts, and maps. Additionally, it offers a great tool for staff members or business owners to clearly deliver data to non-technical audiences. To analyses vast volumes of data and make data-driven decisions, data visualization technologies and tools are crucial in the world of big data.

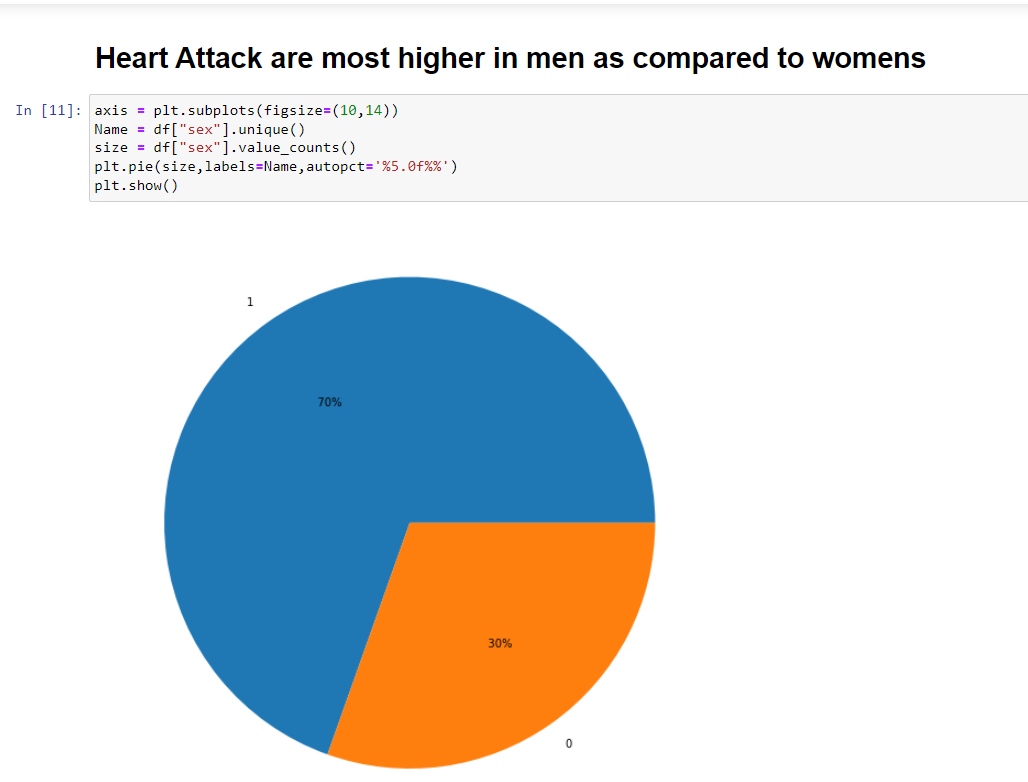
Correlation reveals if the traits are correlated with one another or with the goal variable. If one value rises, the value of the objective variable rises, and vice versa is called a positive correlation (if increase in one value, the value of the target variable decreases). This heatmap shows how closely associated the target variable is with the 'cp' chest pain. We may state that chest discomfort adds the most to the prediction of the presence of heart disease when compared to the relationship between the other two factors.

****

**Figure 3.0: HeatGrid**

## **Machine Learning algorithm used:**

Classification is a supervised learning technique used in statistics and machine learning, where a computer program learns from data and creates new observations or categories. The act of classifying involves grouping a collection of data into various categories. Both structured and unstructured data can use it. The initial stage in the process is to predict the type of data points. The words for the classes typically used are target, label, and categories. The issue with classification predictive modelling is estimating the transformation matrix from discrete input parameters to discrete output variables. Identifying which group or category the new data falls in is the main goal.

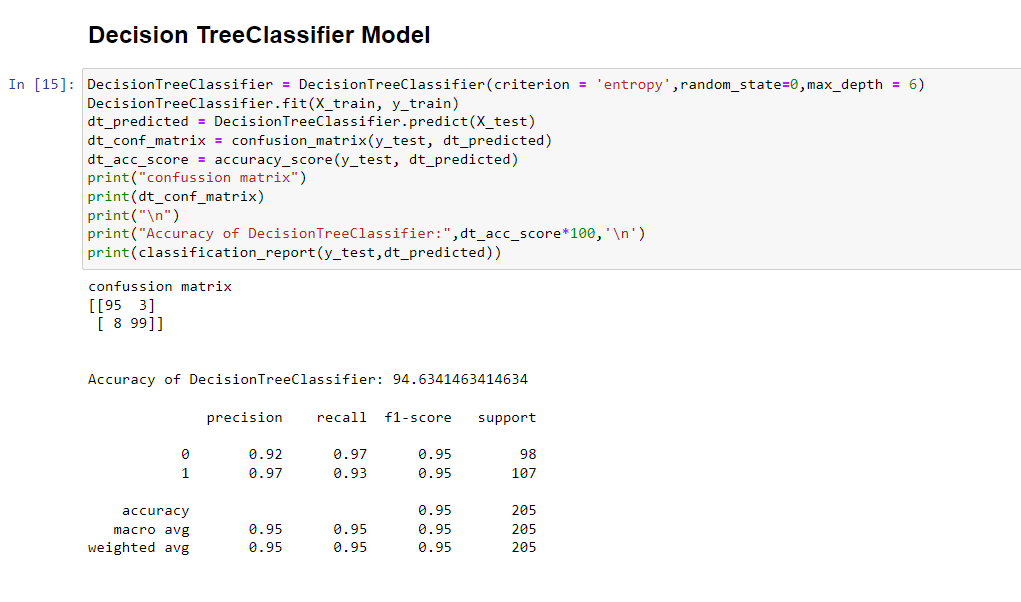
****

**Figure 3.1: Piechart**

I used the below binary classification machine learning algorithms and saw what is working best for our dataset:

## **Decision Tree**

The most well-known decision tree algorithm employs a pruning strategy to create a strong decision tree. Pruning is a technique used to try to get rid of the overly relevant data that overfits the model and makes bad predictions. Finally, a tree is constructed to offer balance, flexibility, and accuracy. For classification issues, the decision tree method is more effective. This method consists of two steps: creating a tree and using the tree to analyze the dataset. The pruning procedure is used by the decision tree algorithm to form a tree. In order to minimize the size of the tree and improve prediction accuracy, pruning removes overfitting data from the model. The classification process is repeated until the data has been as accurately classified as feasible. The training data are most accurately rendered using this technique. The goal is to create a tree with a balance between accuracy & flexibility.



**Figure 4: Decision Tree**

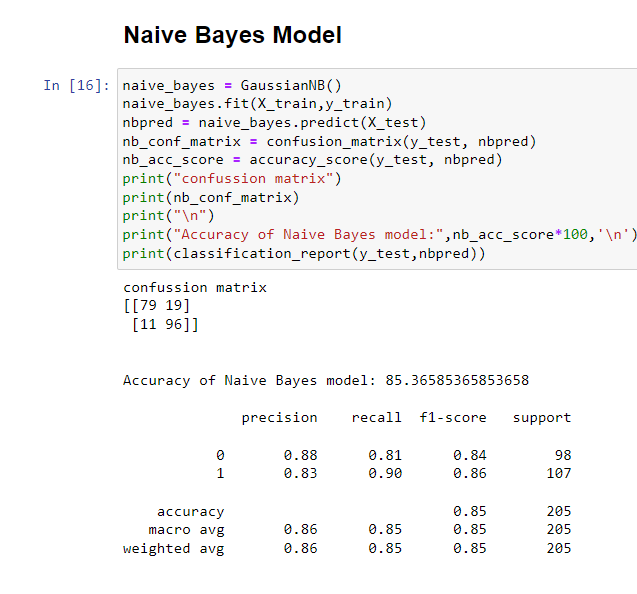
When visualized, the decision tree is a tree-like model that uses decision rules to allocate data to its true target classes in a top-down method. It is a particular type of supervised learning rule that is frequently applied to classification problems. The functions for both continuous and categorical dependent variables. In this rule, we often divide the population into two or more uniform sets.

## **Naive Bayes**

On the Bayes theorem, the Naive Bayes classifier is based. By using conditional independence, this classification technique assumes that the value of an attribute on a specific class is separate from the values of other attributes. As it is based on the Bayes theorem, this classifier uses assumption of independence, which implies that such an attribute value for a given class is distinct from the values of many other attributes.

A statistical classifier called Naive Bayes makes no assumptions about the relationship between attributes. This classifier algorithm employs conditional independence, which asserts that a given class's attribute values are unrelated to those of other attributes. The benefit of utilising naive bayes is that anyone can deal with it without applying any Bayesian techniques.

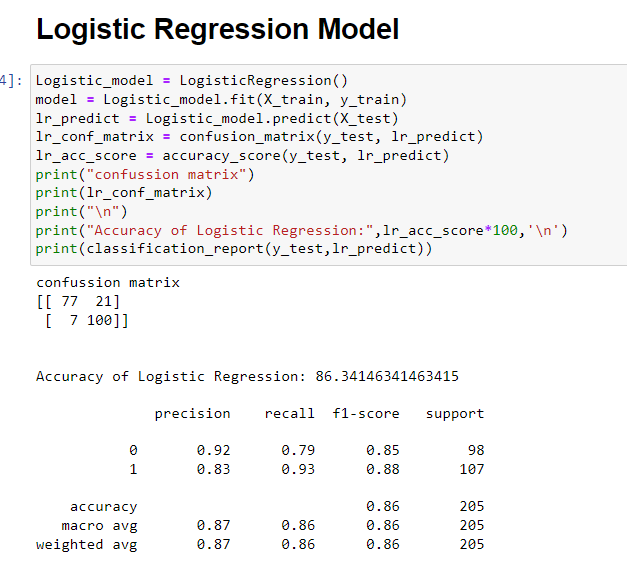
Bayes Theorem, which is derived using conditional probability, serves as the classifier's primary foundation. Given a collection of input attribute value values, this classifier calculates the likelihood that the output feature will occur. The likelihood of each value in relation to the feature will be determined internally. We take into consideration the value of the output attribute that has the highest probability after computing the probabilities of each value. For big datasets, the naive Bayes model is incredibly quick and simple, and it performs well for binary classification.



**Figure 5: Navies bayes**

## **Logistic Regression:**

When a dependent variable is dichotomous, the proper regression analysis to use is logistic regression (binary). The logistic regression is an assumption based analysis, just like all regression analyses. To describe data and analyze the relationship between the dependent variable and one or more independent nominal, ordinal, interval, or ratio-level variables, we employ logistic regression.



**Figure 6: Logistic Regression**

The model fit is yet another important factor to take into account when choosing the framework utilized in the logistic regression analysis. The amount of variation described in the log odds (usually represented as R2) of a model of logistic regression will always rise with the addition of independent variables. The generalization of the system outside of the data with which it is fitted decreases if the model is over fitted, which can happen when there are too many variables included in the model.

Using a collection of independent explanatory variables, logistic regression is often employed in machine learning to forecast the likelihood of response attributes. The target characteristic is used when another dependent variable, such as true or false, also serves as a dependent variable. It's widely used to deal with classification problems. It falls under the category of guided computer learning. It successfully fixes issues with binary and linear classification. This algorithm is one of the most widely used and user-friendly ones. It is a statistical approach to binary class forecasting.

## **Conclusion:**

We used three algorithms Decision Tree, Naïve Bayes and Logistic regression, and implemented them on the Heart Failure dataset and, recorded the results in this project. But we are comparing the results in Figures 4 and 5, we were able to analyze Naive Bayes and Decision Tree in-depth and record the best accuracy results for the dataset under evaluation, which were 85.0% and 94.0%, respectively. With the help of ML techniques, earlier problem recognition can help save lives and prompt the taking of prevention methods for heart disease. This project's main objective is to provide knowledge of prediction models for the Heart Failure Dataset under consideration, which will help in the medical field by employing Decision Tree and Naive Bayes. We come to the conclusion that the Decision Tree algorithm, with an accuracy of 94.0%, performs well on the Heart Failure Dataset. Naive Bayes has an accuracy of 85%, a precision of 86%, and a recall of 85%. The Decision Tree has a 94% accuracy rate, 95% recall rate, and 94% precision rate.

## **References:**

[1] Thuraisingham, B., “A Primer for Understanding and Applying Data Mining”, IT Professional, 28-31, 2000.

[2] Prabhakaran et al, “Cardiovascular Disease in India”, Circulation, Vol 133, No.16, pg.no: 1605 – 1620, 2016.

[3] G.Subbalakshmi et al., “Decision Support in Heart Disease Prediction System using Naïve Bayes”, Indian Journal of Computer Science and Engineering (IJCSE), Vol.2 , No.2,pg.no:170-176, 2011. [4] Chaurasia, Vikas and Pal, Saurabh, Data Mining Approach to Detect Heart Diseases (January 9, 2014). International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol. 2, No. 4, 2013, Page: 56-66, Available at SSRN: <https://ssrn.com/abstract=2376653>

[5] Hossam Meshref, “Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach” International Journal of Advanced Computer Science and Applications (IJACSA), 10(12), 2019. <http://dx.doi.org/10.14569/IJACSA.2019.0101236>

[6] Chaitrali S. Dangare et.al, “Improved Study of Heart Disease Prediction System using Data

Mining Classification Techniques”, International Journal of Computer

Applications,Vol.47,No.10,pg.no:44 – 48, 2012.

[7] Sellappan Palaniappan et.al, “Intelligent Heart Disease Prediction System Using Data Mining

Techniques”, IJCSNS International Journal of Computer Science and Network Security,

Vol.8, No.8,pg.no: 343 – 350, 2008.

[8] Poornima Singh et.al, “Effective heart disease prediction system using data mining techniques”, International Journal of Nano medicine, pg.no:121- 124, 2018.

[9] Era Singh Kajal and Nishika, “Prediction of Heart Disease using Data Mining Techniques”,

International Journal of Advance Research, Ideas and Innovations in Technology, Vol.2,

No.3, pg.no: 1 – 7, 2016.

[10] Mirpouya Mirmozaffari et.al, “Data Mining Classification Algorithms for Heart Disease Prediction”, International Journal of Computing, Communications & Instrumentation Engg.

(IJCCIE), Vol. 4, No.1, pg.no: 11-15, 2017.

[11] Aditya Methaila et al, “Early Heart Disease Prediction Using Data Mining Techniques”, Computer Science and Information Technology, pg.no:53 – 59, 2014.

[12] A. Sheik Abdullah et al , “ A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier”, International Conference on Recent Trends in Computational Methods, Communication and Controls, pg.no:22 – 25, 2012.

[13] Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, “A Data mining approach for prediction of heart disease using neural network’s”, International Journal of Computer Engineering & Technology(IJCET)), Volume 3, Issue 3, October – December (2012), pp. 30-40.

[14] S. U. Amin, K. Agarwal, and R. Beg, “Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors,” ,IEEE Conference on Information and Communication Technologies (ICT 2013), 2013.

[15] Beant Kaur h, Williamjeet Singh, “Review on Heart Disease Prediction System using Data Mining Techniques”, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 10, pp.3003-08, October 2014.

[16] G.V. Nadiammai, M. Hemalatha "E\_ective approach toward Intrusion Detection System using data mining techniques" (2014) 15, 371750.

[17] Robert Mitchell, Ing-Ray Chen "A survey of intrusion detection in wireless network applications" 42 (2014) 11723.

[18] Panos Louvieris n, NatalieClewley,XiaohuiLiu "E\_ectsbased feature identi\_cation for network intrusion detection" (2013)26517273.

[19] Kusrini, Sri Hartati "Implementation of c4.5 algorithm to evaluate the cancellation possibility of new student applicants at stmik amikom yogyakarta" 17-19, 2007.

[20] Bernhard Pfahringer "Winning the KDD99 Classi\_cation Cup: Bagged Boosting" (2000).

[21] Chaitrali S. Dangare, Sulabha S. Apte "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques".

[22] Douglas Burdick, Manuel Calimlim, Johanne Gehrke,“MAFIA: A Maximal Frequent Item set Algorithm For Transactional Databases”, Proceedings of the 17th International Conference on Data Engineering .

[23] Jiawei Han, Micheline Kamber & Jian Pei-Data Mining: Concepts and Techniques; 3rd ed; 2011.

[24] Sellappan Palaniappan, Rafiah Awang, “Intelligent Heart Disease Prediction System Using a Data Mining Techniques”, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008.

[25] Ms. Ishtake S.H , Prof. Sanap S.A., “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, International J. of Healthcare & Biomedical Research, Volume: 1, Issue: 3, April 2013, Pages 94-101.

[26] R. chitra, v. seenivasagam, “review of heart disease prediction system using data mining and hybrid intelligent techniques”, ICTACT journal on soft computing, july 2013, volume: 03, issue: 04.

[27] G. M. Nasira , N. Hemageetha, “Vegetable Price Prediction Using Data Mining Classification Technique”, Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, March 21-23, 2012.

[28] Liao Shaowen, Qiu Decheng, Chen Yong, “Application of decision tree in Data Mining”