IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Classification of HARP dataset using Machine Learning Algorithms

**D. Priyanka[1], K.Sai venkat[2], G.Sirisha[3], K.Akshitha patnaik[4], S.Dillewar rao[5]**

**[1]Assistant Professor, B.Tech Students[2,3,4,5]**

**Dept. Of Computer Science and Engineering**

**Aditya Institute of Technology and Management, Tekkali**

**Srikakulam, Andhra Pradesh, 532201, India**

**Abstract:**Heart diseases are one of the deadly but are silent killers for humans, which results in the increase in death rate of sufferers every year. The World Health Organization (WHO), in the year 2021, reported that 20.5 million deaths that occur worldwide per year are a result of heart disease. Early prediction of heart disease helps in improving the health of patients with corrective measures. Machine learning (ML) based data-driven approaches are found to be viable alternatives. However, such machine learning algorithms suffer from data redundancy and the presence of irrelevant features. This has led to the deterioration of performance in disease prediction when ML approaches are used Decision tree,Naive-Bayes,Logistic regression, SVM, Bagging, Gradient-Boosting, KNN. An empirical study is made with the Cleveland dataset collected from KAGGLE.

**Keywords:**Decision tree, Naive-Bayes, Logistic regression, SVM, Bagging, Gradient-Boosting, KNN.

## 1. INTRODUCTION

Heart disease is the major cause of morbidity and mortality globally it accounts for more deaths annually than any other cause. Over three quarters of these deaths took place in low- and middle-income countries. Machine learning algorithms are widely used to solve problems by learning from data. Data is one of the world's most valuable resource. It is one of the most viable resource to acquisition of new knowledge, and is the key to information gathering. Data has helped in breakthroughs in science and technology. It is a veritable study tool in the field of medicine, law, engineering, etc. People try to gather as much data as possible in their quest to learn new ideas and increase knowledge. There are lots of digital data (Big Data) in sectors like health, agriculture, business, science, and technology to mention just a few. These data are in a raw form which are either unstructured or structured. Huge amount of data (big data) is necessary for the extraction of useful information using a data mining technology or technique. In the healthcare sector, there are a huge amount of medical data of patients that are not mined. These medical records or data of the patients have patterns that are hidden, which is necessary and requires performing data

analysis on them. Health, which deals with the physical, mental, and social wellbeing of persons, is one of the most important sectors that requires much focus on. This is because there are diseases that are very much harmful but silent in their pattern of attack to humans. Heart disease is one such deadly but silent killer resulting in increasing death rate of sufferers yearly. The heart and blood vessel diseases are referred to as cardiovascular disease (CVD) or coronary heart disease.

It serves as an engine room for the body. It pumps blood, and its inability to work properly can cause insufficient circulation of blood to the brain resulting in its deprivation of oxygen and subsequent seizure. This condition, according to medical experts, can lead to death within a few minutes. Heart disease and attack is a major health problem and affects the health of an individual involved due to different types of complications arising from it. Some risk factors, however, have been identified to be associated with heart diseases (depending on the type of heart disease). These include age, gender, family history, high blood pressure is also known as hypertension, diabetes, high level of cholesterol, smoking, alcohol intake, chest pain, stress, poor diet and hygiene, obesity, ethnicity. The complications include but not limited to heart failure, Especially supervised learning methods like Support Vector Machine (SVM) are used in heart disease prediction modes. There is a number of factors contributing to heart disease. These factors when analysed, it is possible to predict heart disease early. The factors include smoking, cholesterol, family history, physical inactivity, obesity, and so on. Many researchers contributed towards defining heart disease prediction models based on ML Techniques. In all the machine learning models, a common problem is that the algorithm performance goes down when there is no feature selection method. Therefore, feature selection is given importance in this paper. With the proposed algorithm and the underlying framework, the feature is reduced prior to the training phase. When a feature is contributing to the class label prediction, it is used as part of the feature selection mechanism.

The proposed feature selection method is known as Information Gain based Feature Selection (IGFS). It makes use of entropy and gains measures in order to compute the utility of a feature in the prediction of class labels. The main contribution of this work is to improve the accuracy of heart disease risk prediction with the feature selection algorithm is known as Information Gain based Feature Selection (IGFS).

### 1.2 Motivation of the project

Today, cardiovascular diseases are the leading cause of death worldwide with 20.5 million deaths annually, as per the World Health Organization reports. Various unhealthy activities are the reason for the increase in the risk of heart disease like high cholesterol, obesity, increase in triglycerides levels, hypertension, etc. There are certain signs which the American Heart Association lists like the persons having sleep issues, a certain increase and decrease in heart rate (irregular heartbeat), swollen legs, and in some cases weight gain occurring quite fast; it can be 1-2 kg daily. All these symptoms resemble different diseases also like it occurs in the aging persons, so it becomes a difficult task to get a correct diagnosis, which results in fatality in near future.

But as time is passing, a lot of research data and patients records of hospitals are available. There are many open sources for accessing the patient's records and researches can be conducted so that various computer technologies could be used for doing the correct diagnosis of the patients and detect this disease to stop it from becoming fatal. Nowadays it is well known that machine learning and artificial intelligence are playing a huge role in the medical industry. We can use different machine learning and deep learning models to diagnose the disease and classify or predict the results. A complete genomic data analysis can easily be done using machine learning models. Models can be trained for knowledge pandemic predictions and also medical records can be transformed and analyzed more deeply for better predictions.

In machine learning, a common problem is the high dimensionality of the data; the datasets which we use contain huge data and sometimes we cannot view that data even in 3D, which is also called the curse of dimensionality. So, when we perform operations on this data, we require a huge amount of memory, and sometimes the data can also grow exponentially and overfitting can happen. The weighting features can be used, so the redundancy in the dataset can be decreased which in turn also helps in decreasing the processing time of the execution . For decreasing the dimensionality of the dataset, there are various feature engineering and feature selection techniques which can be used to remove that data not having that much importance in the dataset.

We collected the data from KAGGLE website .And collected the papers from last ten years in Google Scholar website. We used total seven algorithms namely they are Decision Tree, Support Vector Machine**,** Gradient Boosting, K-Near Neighbour, Naive Baye's, Logistic Regression, Bagging.

## 1.3 ALGORITHMS

### 1.3.1 Decision Tree

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

### 1.3.2 Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1 Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

### 1.3.3 Naive bayes

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

### 1.3.4 K-Nearest Neighbours

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

### 1.3.5 Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. SVM are classified as two types namely they are linear and non-linear.

### 1.3.6 Gradient Boosting

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels. Machine learning is one of the most popular technologies to build predictive models for various complex regression and classification tasks. Gradient Boosting Machine (GBM) is considered one of the most powerful boosting algorithms.Although, there are so many algorithms used in machine learning, boosting algorithms has become mainstream in the machine learning community across the world. Boosting technique follows the concept of ensemble learning, and hence it combines multiple simple models (weak learners or base estimators) to generate the final output.

### 1.3.7 Bagging

Bagging is used when our objective is to reduce the variance of a decision tree. Here the concept is to create a few subsets of data from the training sample, which is chosen randomly with replacement. Now each collection of subset data is used to prepare their decision trees thus, we end up with an ensemble of various models. The average of all

the assumptions from numerous trees is used, which is more powerful than a single decision tree. Machine Learning uses several techniques to build models and improve their performance. Ensemble learning methods help improve the accuracy of classification and regression models. This article will discuss one of the most popular ensemble learning algorithms, i.e., Bagging in Machine Learning.

## 2. Literature Survey

G.Ramesh, et.al[1]IGFS is the feature selection algorithm proposed to enhance the performance various heart disease prediction models made up of machine learning techniques namely Naive Bayes, SVM, Bagging etc., are the prediction models evaluated with and without feature selection. The performance of the prediction models increased in terms of accuracy when feature selection algorithm IGFS is employed. Python data science platform known as. SVM and Bagging showed highest performance with accuracy rate 0.99. In future we improve the prediction of heart disease with deep learning models. Another direction for future work is to develop a Decision Support System that not only predicts heart diseases but also heart diseases occurred due to diabetes.

M.Rahian et.al[2] found in-depth analysis including as many metrics and instances as possible offers more options flexible in trade-off among the algorithms. They provide constructive selection approaches that are the uttermost inimitable approach. Considering all the facts, we have determined Bagging as the best algorithms to predict ACS in view of evaluation metrics and performance parameters respectively which have been rationalized by the receiver operating characteristic curves (ROC). We can use this Bagging technique for any system and will get the accurate result in anticipating the heart attack risk.

N.Komal Kumar et.al[3] The machine learning classifiers such as Bagging, Decision Tree, Logistic Regression, Support vector machine (SVM), K-nearest neighbours (KNN), Gradient Boosting, Naive bayes were used in the prediction of Cardio Vascular Disease (CVD). The proposed method using a Bagging machine learning classifier has achieved a greater accuracy of 99.3% with a ROC AUC score of 0.9903 which outperformed all the classifiers under analysis in classifying patients with Cardio Vascular Disease.

T Obasi et.al[4] Observed that out performed Logistic Regression and Naive Bayes Classifier both in accuracy and precision. An accuracy of 0.7296 was obtained for Naive Bayes Classifier while Logistic Regression has the lowest accuracy of 0.5970 with precisions of 0.7020 and 0.6060 for Naive Bayes and Logistic Regression a proposed model that has been built using existing machine learning approaches, to detect and predict heart diseases and heart attacks in humans using the existing medical records of patients as the dataset used in training and testing the model. which is a supervised machine learning used for classification purposes. It was compared with Logistic Regression, and Naive Bayes Classifier used to build the second and third models, respectively, in terms of efficiency and accuracy of our model. New patterns were discovered from the analysis performed, such as the importance of the variable. After execution, it was found that Bagging performed better than Logistic Regression Naive Bayes Classifier with accuracy of 99.3%, 82.7%, and 61.96%, respectively.

More attributes (risk factors) were captured in our model compared to other existing works to enhance and expand the system by detecting and predicting heart diseases in individuals who have one of the risk factors that were not included in other researches. This system can be used in the health sector to assist medical professionals/practitioners in detecting and predicting heart disease/attack in patients to minimize the mortality rate it causes yearly.

Sushmita Manikandan et al.[5] discovered a prototype of a system that identifies an individual based on his risk factor. The dataset was downloaded and pre-processed from the KAGGLE machine learning repository. There are 14 predictor variables or features in the final dataset, as well as one response variable labelled num. If the number is 0, it means that less than 50% of the blood vessels are narrowing, indicating that the person is a 'low risk individual.' If the number is 1, it means that more than 50% of the blood arteries are narrowing, indicating that the person is a 'high risk individual.' The Gaussian Nave Bayes technique was used to classify the data since it followed a normal distribution. Among the most up-to-date classifications.

K. Srinivasa et.al[6] have presented automated and effective heart attack prediction methods using data mining techniques. Firstly, we have provided an efficient approach for the extraction of significant patterns from the heart disease data warehouses for the efficient prediction of heart attack Based on the calculated significant weightage, the frequent patterns having value greater than a predefined threshold were chosen for the valuable prediction of heart attack. Three mining goals are defined based on data exploration. All these models could answer complex queries in predicting heart attack. This can further enhanced and expanded. For predicting heart attack significantly 15 attributes are listed in medical literature. Besides this list, we have to incorporate other attributes which will effect on results such as financial status, stress, pollution and previous medical history. Other data mining techniques, Time Series, Clustering and Association Rules are also can be used to analyze patients behaviour.

Paul T. Williams et.al[7] absence of cross-border direction and technology integration demands standards to enable interoperability amid the elements of the big data value chain. Big data proposes vast promises for detecting interactions and nonlinearities in relationships among variables. Mobile devices, such as smart phones and tablets, and sensors, will continue to be the most indispensable tools available to deliver heart attack prediction and telecardiology services over wireless networks to reduce cardiovascular disease morbidity and mortality. He deployment of cloud computing has inexpensively facilitated the collaborative application of telecardiology between hospitals and has expanded services from regional to global. He most important factor, however, in the development and application of big data, telecardiology, sensor use, mobile phone or tablet use and landline use is patient privacy and to safeguard the patient's ability to direct and discover the use of his or her health care information.

| S.NO | TITLE | METHODOLOGY USED | PERFORMANCE MATRIX | REFERENCE NUMBER |
|---|---|---|---|---|
| 1 | Improving the accuracy of heart attack risk prediction based on information gain feature selection technique | Bagging | 99% | [8] |
| 2 | A Comprehensive Analysis on Risk Prediction of Acute Coronary Syndrome using Machine Learning | Gradient Boosting | 96.09% | [9] |
| 3 | Analysis and Prediction of Cardio Vascular Disease using Machine Learning | Decision Tree | 90.24% | [10] |
| 4 | Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases | Logistic Regression | 86.34% | [11] |
| 5 | Heart Attack Prediction System | Naive Bayes | 81.46% | [12] |
| 6 | Analysis of Coronary Heart Disease and Prediction of Heart Attack | SVM | 74.63% | [13] |
| 7 | Big Data Analytics in Heart Attack Prediction | KNN | 74.63% | [14] |

**TABLE 2.1:** LITERATURE SURVEY

### 3.1 PROPOSED METHODOLOGY

Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. Heart disease was the major cause of casualties in the different countries including India. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases.

### 3.2 Existing System:

The World Health Organization (WHO) has estimated that 17.9 million deaths occur worldwide, every year due to the Heart diseases. About 30% deaths in the age group of 25-69 year occur because of heart diseases. In urban areas, 32.8%. Deaths occur because of heart ailments, while this percentage in rural areas is 22.9. Over 80% of deaths in world are because of Heart disease. Here they have predicted the accuracy of heart diseases using Bagging algorithm. Therefore, in our project we decide to spend some time working on better algorithms to detect better accuracy compared to those algorithms.

### 3.3 Proposed system:

In this system we are implementing effective heart attack prediction system using Naive Bayes, Decision tree, Logistic Regression, SVM, KNN, Gradient Boosting algorithm. We can give the input as in XLS file or manual entry to the system. After taking inputapply the algorithms on that input**.** After accessing data set the operation is performed and effective heart attack level is produced.

The proposed system will add some more parameters significant to heart attack with their weight,age and the priority levels are by consulting expertise doctors and the medical experts. The heart attack prediction system designed to help the identify different risk levels of heart attack like normal, low or high and also giving the prescription details with related to the predicted result.

### 3.4 Flow chart

Below Figure represents the step by step procedure of the proposed system. It clearly explains us about complete view of proposed system that is where we have taken the data and what methods and techniques applied etc. Here we are selecting the data from KAGGLE then applying data pre-processing techniques for data cleaning and transformation. After that evaluating the results based on machine learning classification algorithms.

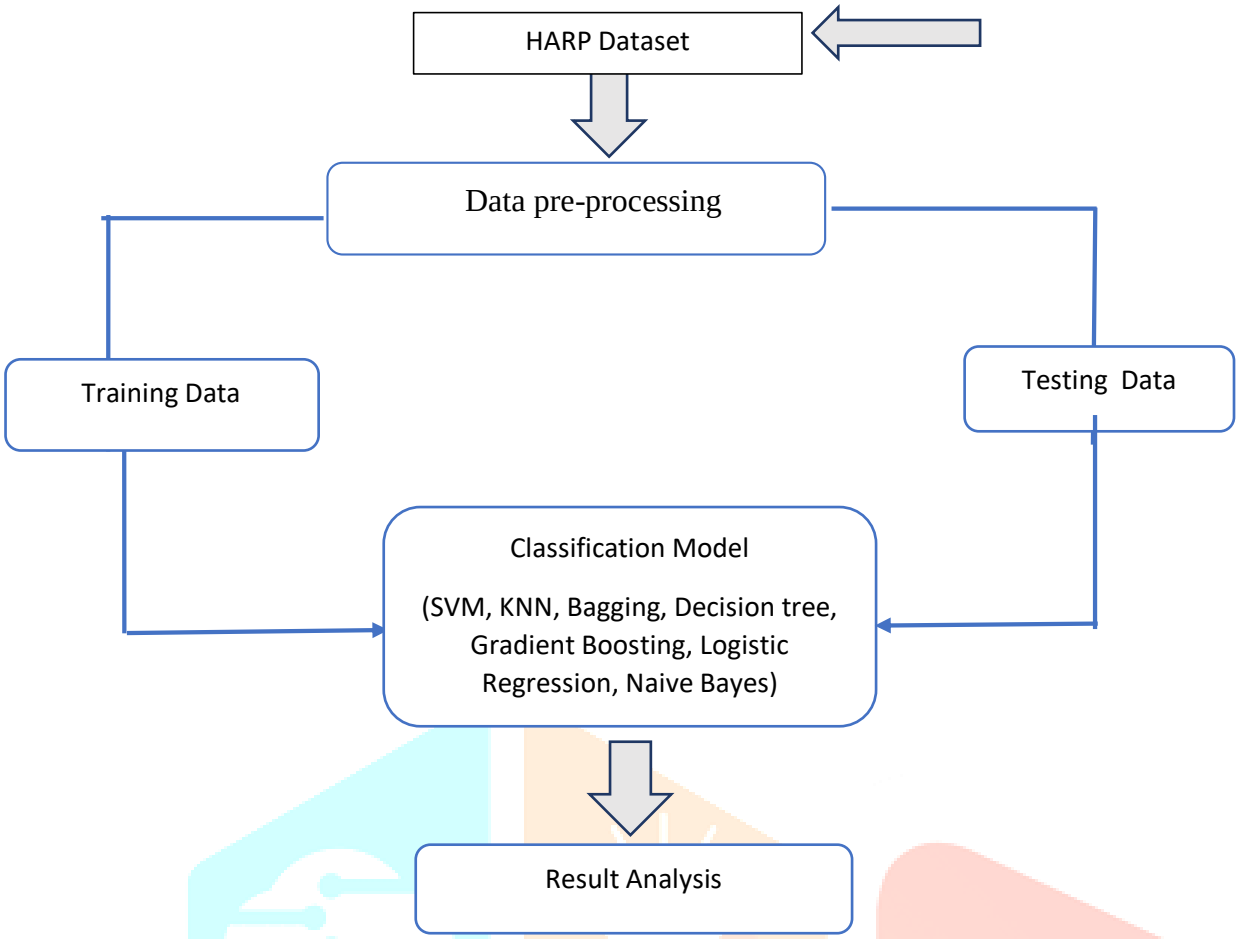



**Fig.3.4:** Architecture for the proposed system

### 3.5 Methods used

In this, we have used seven different classification algorithms; they are SVM, Logistic Regression, KNN, Decision tree, Gradient boosting, Naive bayes, Bagging. By using these algorithms, we have computed the analytical results based on various statistical parameters like Accuracy, Precision and Recall.

#### 3.5.1 Support Vector Machines

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane.

#### 3.5.2Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and

1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

### 3.5.3 K-Nearest Neighbour (KNN)

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

### 3.5.4 Decision Tree

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

### 3.5.5 Gradient Boosting

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels. There is a technique called the Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees). The AdaBoost Algorithm begins by training a decision tree in which each observation is

assigned an equal weight. After evaluating the first tree, we increase the weights of those observations that are difficult to classify and lower the weights for those that are easy to classify. The second tree is therefore grown on this weighted data. Here, the idea is to improve upon the predictions of the first tree. Our new model is therefore Tree 1 + Tree 2. We then compute the classification error from this new 2-tree ensemble model and grow a third tree to predict the revised residuals. We repeat this process for a specified number of iterations. Subsequent trees help us to classify observations that are not well classified by the previous trees. Predictions of the final ensemble model is therefore the weighted sum of the predictions made by the previous tree models.

### 3.5.6 Naive Baye's

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

$$P(c|x)=P(x1|c)*P(x2|c)*.....*P(xn|cx)*P(c)$$

### 4.2 Dataset

Experiments are made with the application developed using a Python data science platform known as Spyder. The

### 3.5.7 Bagging

Bagging is used when our objective is to reduce the variance of a decision tree. Here the concept is to create a few subsets of data from the training sample, which is chosen randomly with replacement. Now each collection of subset data is used to prepare their decision trees thus, we end up with an ensemble of various models. The average of all the assumptions from numerous trees is used, which is more powerful than a single decision tree. Machine Learning uses several techniques to build models and improve their performance. Ensemble learning methods help improve the accuracy of classification and regression models. This article will discuss one of the most popular ensemble learning algorithms, i.e., Bagging in Machine Learning.

## 4. EXPERIMENTAL SETUP

### 4.1 Simulative Environment

The developing environment for the proposed method is Spyder(python 3.9) version on a system with Intel i3or i5 or further versions or AMD x86-64 processor, 1.8 GHz, 8 GB RAM and Microsoft Windows with Family. In Experimental setup we are explaining about the Data Set i.e, Heart attack risk prediction data set and python environment called Spyder.

Dataset is collected from KAGGLE and the attributes description of dataset is as shown below(fig 3.6). The dataset has 14 attributes and the last Attribute is diagnosis feature or class label. In other words, the 14th feature is

known as dependent variable while 13 features are used as the independent variables. However, feature selection reduces the dimensions to improve prediction performance. In our data set chest pain(cp) attribute further classified into four coloumns such as value1,value2,value3 and value4,to know the type of pain.In ourdataset there are 1026 instances and 14 attributes. And another attribute called Resting Electrocardiographic attribute has three coloumns for specific details and the three coloums named as value1,value2 and value3.The data was collected from the four following locations:

1. Cleveland Clinic Foundation (cleveland.data)

- cp : Chest Pain type chest pain type
    - Value 1: typical angina
    - Value 2: atypical angina
    - Value 3: non-anginal pain
    - Value 4: asymptomatic
- trtbps : resting blood pressure (in mm Hg)
- chol : cholestoral in mg/dl fetched via BMI sensor
- fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- rest_ecg : resting electrocardiographic results

2. Hungarian Institute of Cardiology, Budapest (hungarian.data)

3. V.A. Medical Center, Long Beach, CA (long-beach-va.data)

4. University Hospital, Zurich, Switzerland (switzerland.data)

Using this data set we have analyzed different factors which are mainly leading to heart attack data by applying various machine learning algorithms

About this dataset

- Age : Age of the patient
- Sex : Sex of the patient
- exang: exercise induced angina (1 = yes; 0 = no)
- ca: number of major vessels (0-3)
    - Value 0: normal
    - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- thalach : maximum heart rate achieved
- target : 0= less chance of heart attack 1= more chance of heart attack

## 5. RESULT ANALYSIS

### 5.1 Results of Classification

By Using Different Classification Algorithms we are getting different values for each algorithm in different cases such as

© 2022 IJCRT | Volume 10, Issue 6 June 2022 |

f1score, recall, accuracy and precision. Based on these parameters we can know which algorithm is best for our system.

### 5.2 Confusion Matrix

The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing. To find out the accuracy of error rate algorithms confusion matrix plays major role. The Accuracy is one metric for evaluating classification models.

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

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

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

**Precision**

Precision is the number of correct positive results divided by the number of positive results predicted by the classifier**.**

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

**Recall**

Recall is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

$$\text{Recall} = \frac{True\ Positive(TP)}{True\ Positive(TP)+False\ Negative(FN)}$$

**F1 Score**

F1 Score is used to measure a test's accuracy. F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1].

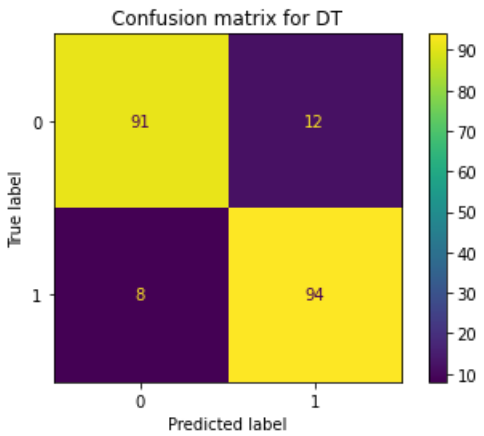
$$F1 = 2 * \frac{1}{\frac{1}{Precision}+\frac{1}{Recall}}$$

confusion matrix displays the total number of observations in each cell. The rows of the confusion matrix correspond to the true class, and the columns correspond to the predicted class. Diagonal and off diagonal cells correspond to correctly and incorrectly classified observations, respectively.

**Accuracy**

### 5.3 Confusion matrix for Decision tree:

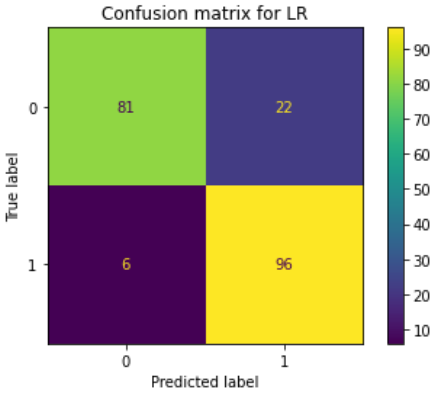
Decision Tree has predicted 94 True Positives and 91 True Negatives and 12 False Positives and 8 False Negatives.



© 2022 IJCRT | Volume 10, Issue 6 June 2022 |

**Fig.5.3:**Confusion matrix for DT

**5.4**




**Confusion matrix for Logistic Regression:**

Logistic Regression has predicted 96 True Positives and 81 True Negatives and 22 False Positives and 6 False Negatives.

**Fig.5.4:**Confusion matrix for LR

**5.5 Confusion matrix for SVM:**

SVM has predicted 80 True Positives and 73 True Negatives and 30 False Positives and 22 False Negatives.

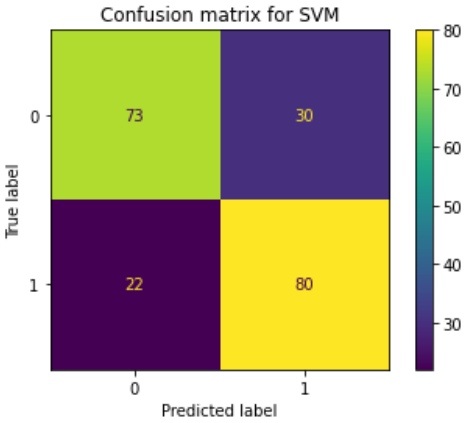



**Fig.5.5:**Confusion matrix for SVM

**5.6 Confusion matrix for KNN:**

KNN has predicted 72 True Positives and 81 True Negatives and 22 False Positives and 30 False Negatives.

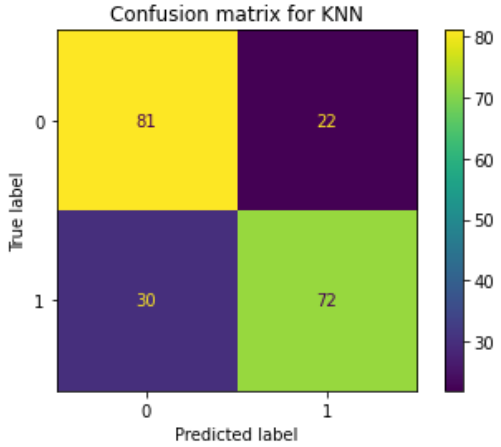



**Fig.5.6:**Confusion matrix for KNN

**5.7 Confusion matrix for Bagging:**

Bagging has predicted 100 True Positives and 102 True Negatives and 1 False Positives and 2 False Negatives.

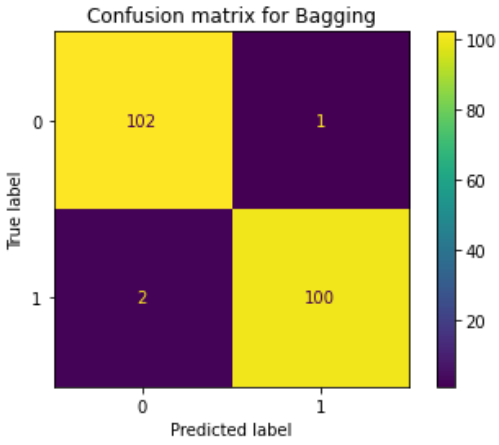



**Fig.5.7:**Confusion matrix for Bagging

© 2022 IJCRT | Volume 10, Issue 6 June 2022 |

### 5.8 Confusion matrix for NB:

NB has predicted 85 True Positives and 82 True Negatives and 21 False Positives and 17 False Negatives.

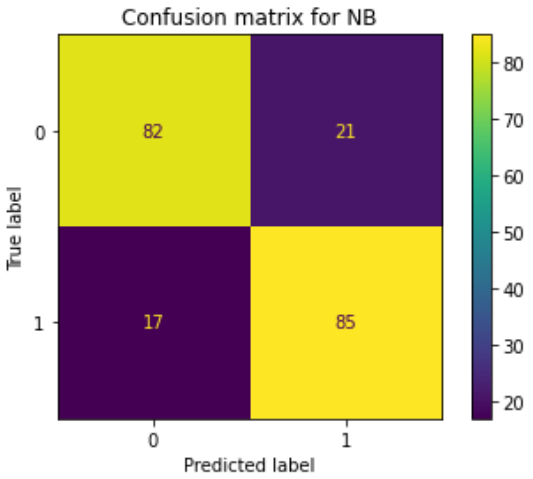



**Fig.5.8:**Confusion matrix for NB

### 5.9 Confusion matrix for GBC:

KNN has predicted 99 True Positives and 98 True Negatives and 5 False Positives and 3 False Negatives.

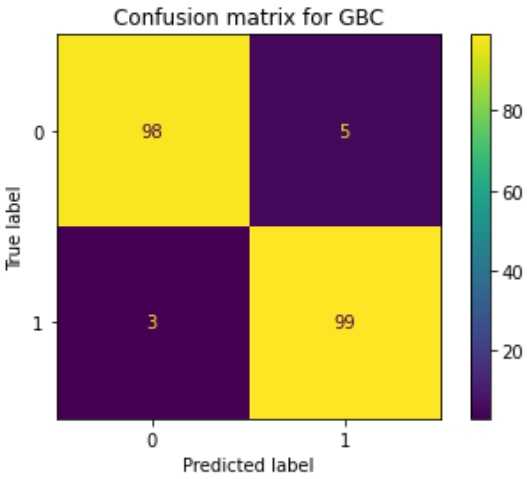



**Fig.5.9:**Confusion matrix for GBC

### 5.10 Bar plots for classifiers

#### 5.10.1 Test accuracy

The test accuracy for different classifiers like DT, LR, SVM, KNN, Bagging, NB, GBC is shown below. Among those classification algorithms Bagging gives the highest accuracy i.e is 98%.

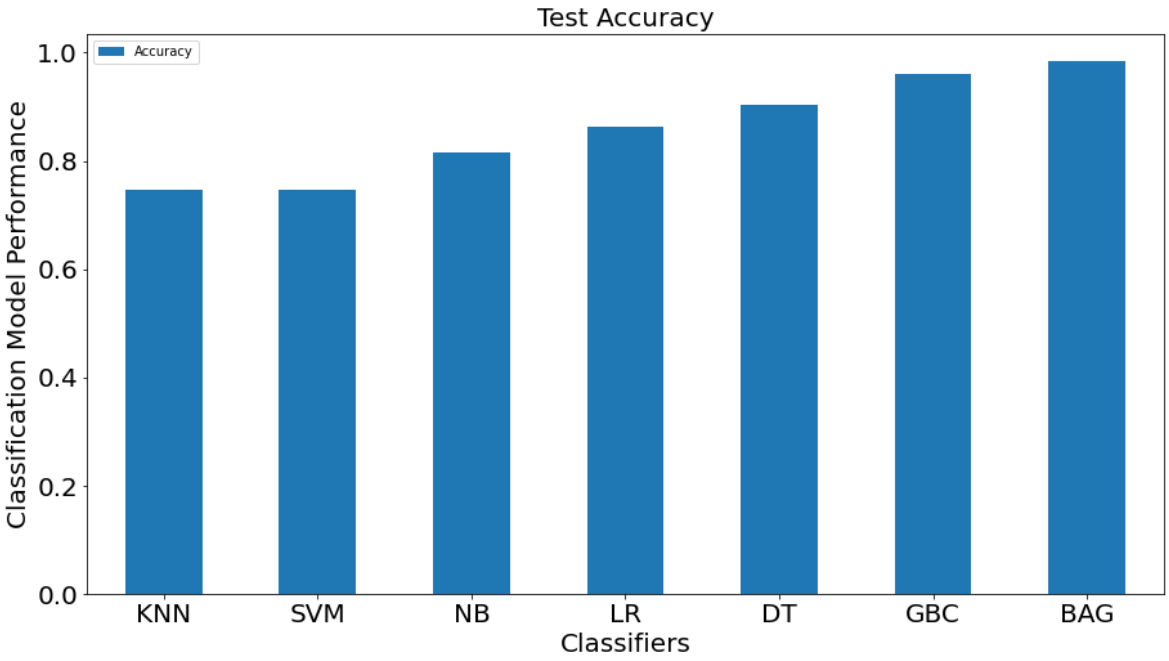



**Fig 5.10.1:**Test accuracy for classifiers

### 5.10.2 Comparison of various performance metrices

The below figure shows the comparison between the performance metrics (Accuracy, Precision, Recall, F1_Score ) of heart attack risk prediction. As shown in the Figure blue, orange, green, red colours indicate the performance metrics Accuracy, Precision, Recall, F1_ Score. The performance metrices of bagging is best compared to other classification algorithms. Hence we take bagging as our best evaluation model for heart attack risk prediction.

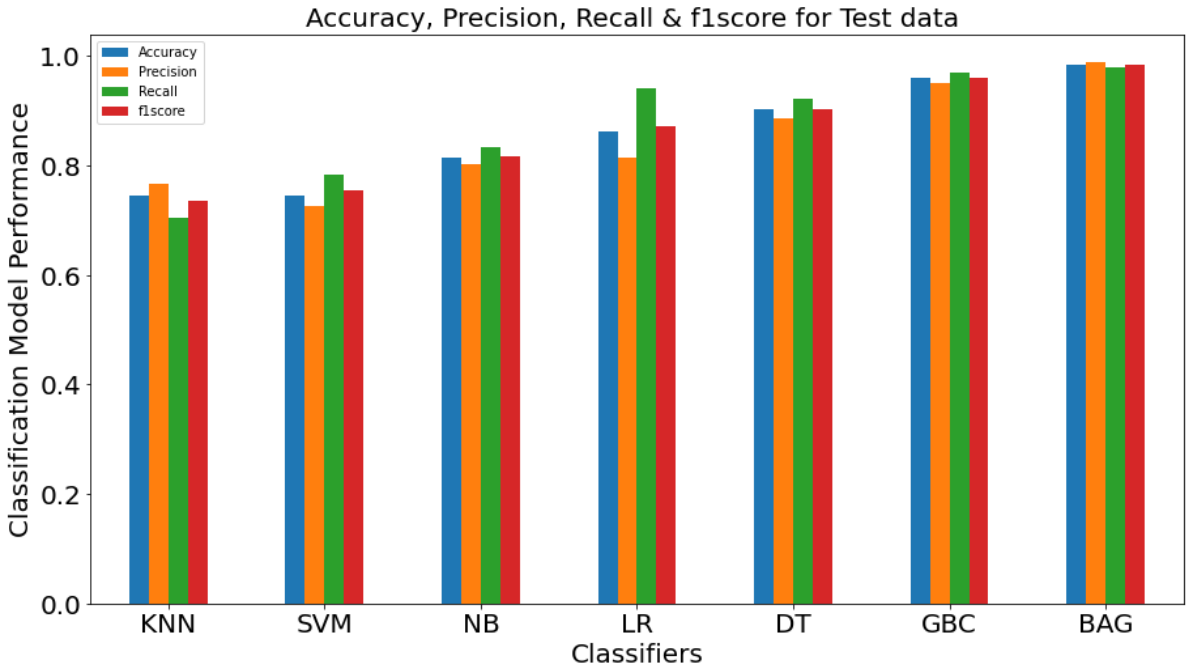



**Fig :5.10.2** Performance metrices of different classifiers

The below table represents the four performance measures (Accuracy, Precision, Recall, F1_Score) for all classification algorithms that are applied to heart attack risk prediction dataset. This shows that Bagging outperforms in all the performance parameters and provides the best results. The accuracy of 98% has been achieved on the heart attack risk prediction data set by bagging is the highest accuracy.

**Table 5.10.2** Accuracy,Precision,Recall,F1-Score

| Classifiers | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| **DT** | 0.90 | 0.88 | 0.92 | 0.90 |
| **LR** | 0.86 | 0.81 | 0.94 | 0.87 |
| **NB** | 0.81 | 0.80 | 0.83 | 0.81 |
| **SVM** | 0.74 | 0.72 | 0.78 | 0.75 |
| **KNN** | 0.74 | 0.76 | 0.70 | 0.73 |
| **GBC** | 0.96 | 0.95 | 0.97 | 0.96 |
| **Bagging** | 0.98 | 0.99 | 0.98 | 0.98 |

## 6. CONCLUSION AND FUTURE WORK

### 6.1 Conclusion

In present scenario, Heart attack risk prediction is the systematic process to obtained and evaluating objectively evidences about the correspondence between information, situations or procedures and established criteria. It is to ensure that appropriate attention is devoted to important areas, potential problems are promptly identified, and work is completed expeditiously and also properly coordinated. In existing system they have taken Heart attack risk prediction data set and applied

several algorithms are implemented to understanding the complete risk prediction process and the researchers got 98.5% of accuracy by using Bagging algorithm. Means while we have triggered out the best algorithm among K-Nearest Neighbour, Support Vector Machine, Bagging, Gradient Boosting Classifier, naive bayes classifier and Logistic Regression by analyzing these seven algorithms and we got the result that Bagging is best among all. From result analysis, by considering the recall value, the accuracy, precision of Bagging is more when compared with other algorithms. And also find which algorithms is best performance classifier.

**6.2 Future work**

For future works, we are targeting to improve the performance of the classifiers by the ensemble machine learning approach; effort can be made for obtaining more accuracy and better execution time for prediction.

**REFERENCES**


1. Williams, Paul T., and Paul D. Thompson. "Increased cardiovascular disease mortality associated with excessive exercise in heart attack survivors." Mayo Clinic Proceedings. Vol. 89. No. 9. Elsevier, 2014.
2. Anooj, P. K. "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules." Journal of King Saud University-Computer and Information Sciences 24.1 (2012): 27-40.
3. Manikandan, Sushmita. "Heart attack prediction system." 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS). IEEE, 2017.
4. M. Raihan, S. Mondal, A. More, M. Sagor, G. Sikder, M. Arab Majumder, M. Al Manjur and K. Ghosh,"Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and datamining approaches, a prototype design", in 2016 19th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2016, pp. 299-303.
5. Obasi, Thankgod, and M. Omair Shafiq. "Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases." 2019 IEEE international conference on big data (big data). IEEE, 2019.
6. Kumar, N. Komal, et al. "Analysis and prediction of cardio vascular disease using machine learning classifiers." 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, 2020.
7. Ramesh, G., et al. "Improving the accuracy of heart attack risk prediction based on information gain feature selection technique." Materials Today: Proceedings (2021).
8. Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart disease prediction using machine learning. International Journal of Research and Technology, 9(04), 659-662.
9. Srinivas, K., G. Raghavendra Rao, and A. Govardhan. "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques." *2010 5th* International

Conference on Computer Science & Education. IEEE, 2010.

10. Manikandan, S. (2017, August). Heart attack prediction system. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 817-820). IEEE.
11. Obasi, T., & Shafiq, M. O. (2019, December). Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases. In 2019 IEEE international conference on big data (big data) (pp. 2393-2402). IEEE.
12. Williams, P. T., & Thompson, P. D. (2014, September). Increased cardiovascular disease mortality associated with excessive exercise in heart attack survivors. In *Mayo Clinic Proceedings* (Vol. 89, No. 9, pp. 1187-1194). Elsevier.
13. Gour, Sanjay, et al. "A Machine Learning Approach for Heart Attack Prediction." *Intelligent Sustainable Systems*. Springer, Singapore, 2022. 741-747.
14. Krishnan, Santhana, and S. Geetha. "Prediction of Heart Disease Using Machine Learning Algorithms." *2019* 1st international conference on innovations in information and communication technology (ICIICT). IEEE, 2019.