PERFORMANCE ANALYSIS OF DATA MINING ALGORITHMS FOR PREDICTION OF HEART DISEASES

[1]Er. Manpreet Kaur, [2]Er. Shailja
[1]Research Scholar, [2]Assistant Professor
[1]Computer Science and Engineering,
[1]Guru Nanak Dev Engineering College, Ludhiana, India.

Abstract : Data mining is a technique that turn around examine and producing out important data for large volume of dataset to making the beneficial decisions. Data mining in medical area plays a vital role based on classification. There are various problems that can be analyzed early and can be treated at the first stage. Heart diseases the classical way of analyzing heart diseases depend on medical profile and many medical tests. The main objective of this research on heart disease prediction by acquiring the data and information from the medical profile of patients. The prediction of heart disease using different techniques like support vector machine (SVM), K Nearest Neighbors (KNN) and logistic regression. These techniques produce a result with the different accuracy level but KNN algorithms gives more accurate result as compared to SVM and logistic regression. The system predicted the disease with good accuracy.

IndexTerms **- Data mining, classification, Predictive modelling, logistic regression, support vector machine, KNN (k nearest neighbors.**

I. INTRODUCTION

Data mining generates the knowledgeable information from the large aggregate of data in distinctive research zone. In healthcare zone, data mining represents a very important act. Basically, it helps to sort the patient data for producing the incurious decisions. The main research work concentration on predicting the heart disease from patient's data and compare with the different data mining algorithms. In 2017, 365914 people are died by heart disease. Heart problem arise in body due to consume large amount of tobacco, Smoking, hyper stress, blood pressure, chest pain, high cholesterol, drinking alcohol etc. When a nerve becomes blocked by blood clot and the tissues of heart losses due to blood clot then heart attack occurs in body. In hospitals, all the patient's data are stored in the database and apply the data mining techniques on the patient's data to predict the disease. Two factors define the data mining – supervised learning and unsupervised learning.

Supervised Learning

Supervised learning includes the training set to train the attributes for dataset. It contains two approaches one is classification and other is regression. Classification summarize the category of variable that differentiate the data classes. It helps to identify the problems by given the training set of data. Some classifiers are used in data mining such as decision tree, neural network, SVM, KNN. Decision tree generates the results in the form of tree for classification and regression techniques. It separates the dataset into various stages and depicts the Boolean value 'yes' or 'no'. These values help to formulate the decisions that can easily fulfill the objectives. It can easily map from parent node to leaf node. Neural network is mathematical that made up of interlinked group of artificial nerve cell. In learning mode, neural network is a robust system that change its design depends upon internal and external data that passes through the network area. Regression summarize that values which are known of variables and that values to be predicted. Linear regression and logistic regression are most popular techniques used to predictive analysis. Regression contains one or more independent values.

Unsupervised Learning

Unsupervised learning is an algorithm used to complete the complex task. It helps to find the various types of new patterns in documents or data. It contains two approaches one is clustering and other is association rule. Clustering is a technique of dividing the data into the same unit. The process of dividing the data into slots known as cluster(Rishab Sexana,2019). A cluster contains the objects of data with greater similarity. Clustering also known as data segmentation. Various types of clustering are – HM (hierarchical method), PM (partitions method), DBM (density-based method) and GBM (grid-based method). Association rule helps to find the interesting organization and relationships for large volume of data set. This rule indicates the how constant itemset appears in a data. The expression of association rule in the form X->Y where X and Y are any two itemset from data.

II. LITERATURE REVIEW

Vembandasamy et al., diagnosed heart disease using the Naive Bayes algorithm. Bayes' theorem is used in Naive Bayes. Naïve Bayes is a classifier that based on the probability method and apply with consideration between the features. The data collected are from diabetes research institutes in Chennai. The data set consists of 500 patients. WEKA is used as a tool and performs classification using 70 percent of the Percentage Split. Naive Bayes offers 86.419 percent accuracy(Deepa,2015).

Sushmita Manikandan, proposed heart attack prediction system using machine learning repository of UCI to develop model. Rapid Miner tool and Anaconda v2.7 software is used to construct the classifier and purifying the dataset. Naive Bayesians theorem is used to predict the heart diseases. Different tools are used for analysis and prediction can be successfully completed using data mining(Sushmita Manikandan,2017).

Senthilkumar Mohan, propose a method that applying machine learning techniques resulting in improving the accuracy in the prediction of heart disease. Various techniques in data mining like K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Logistic regression, and Naive Bayes (NB) and Support vector machine (SVM). The dataset is collected from UCI repository. The prediction model is introduced for heart disease with the hybrid random forest with a linear model and produce the accuracy level of 88.7 percent through the prediction model(Senthilkumar Mohan,2019).

G. Purusothaman and P. Krishnakumari, proposed data mining techniques for prediction of heart disease. Various methods are applied to evaluate heart disease risk prediction. Two prediction models used for data analysis. Applying single model to heart data is first one and second is for implementing combined model to get hybrid model. The hybrid model gives 96% accuracy for prediction(G. Purusothaman,2015).

Vincy Cherian et al., proposed prediction using Laplace smoothing and Naive Bayesian algorithms. Early diagnosing patient with heart disease both time and money can be saved. This database contains 14 attributes. The system indicates whether a patient have heart disease or not. Laplace smoothing technique is more accurate than the naïve Bayesian and it gives 86% accuracy to predict the heart diseases(Vincy Cherian,2017).

Otoom et al., presented a system for analysis and prediction. Cleveland Heart data are taken from the UCI. This dataset consists of 303 cases and 76 attributes/features. 13 features are used out of 76 features. Two tests with three algorithms: Bayes Naive, Support vector machine, and Functional Trees are performed for detection purposes. By applying the test to the 7 best features selected, Bayes Naive achieved 84.5 percent accuracy, SVM provides 85.1 accuracy and FT classifies 84.5 percent correctly(Kilan,2015).

Theresa Princy. R et al., conducts a survey related to various classification techniques that can predict risk factors related to every individual considering the factors such as gender, age, BP, cholesterol, pulse rate. By means of various data mining classification techniques like Naive Bayes, Decision Tree Algorithm, KNN and Neural Network etc., patients risk level can be classified. Since a lot of attributes are taken into account, high accuracy is achieved for the risk level(Theresa Princy. R,2016).

K.S.Thirunavukkarsu, recommends an effective disease diagnosis model and the accuracy with the disease prediction were compromised. In order to obtain higher classification accuracy for heart and stroke disease diagnosis, a Fast Correlation Filter Classifier (FCF) technique is introduced. The main objective of the FCF technique is effectively performs disease. Fast Correlation Filtering (FCF) algorithm is used to select the most relevant attributes for disease diagnosis and filter out the irrelevant attributes in dataset(K.S.Thirunavukkarsu,2017).

III. METHODOLOGY

The dataset of patients with heart diseases is taken from UCI repository. in this dataset, each heart patients record contains 13 attributes. totally, 300 instances are designed for prediction. every patients attribute is being amassed such as population based, life history and lab features. attribute can be numeric or float valued are in given below table1.

Table 1 Data Description

S.No.	Attribute Name	Data Type
1	Age	Numeric
2	Sex	Numeric
3	Chest pain type (CP)	Numeric
4	Trestbps	Numeric
5	Cholesterol	Numeric
6	Fasting blood sugar (FBS)	Numeric
7	Resting ECG	Numeric
8	Thalach	Numeric
9	Exang	Numeric
10	Oldpeak	Float
11	Slope	Numeric
12	Thal	Numeric
13	Ca	Numeric

Various algorithms used for the prediction of heart diseases. In this work, Performance of different classification techniques like Support vector machine, Logistic regression and KNN (k-nearest neighbor) algorithm is analyzed for prediction of heart diseases. These steps are defined as follows:

- To import the data for heart diseases from UCI repository
- Apply pre-processing on dataset to check the missing values in data.
- Apply Scaling on all the features in dataset using min and max scalar for data division.
- Then dataset is divided into two parts one is training dataset and other is testing dataset.
- Apply Data mining algorithms on the dataset for the prediction.
- Analyze the performance of work on the basic of various attributes if the performance
- is degraded then we optimized the accuracy to dataset.

- At last we calculate performance comparison of these algorithms with their accuracy.

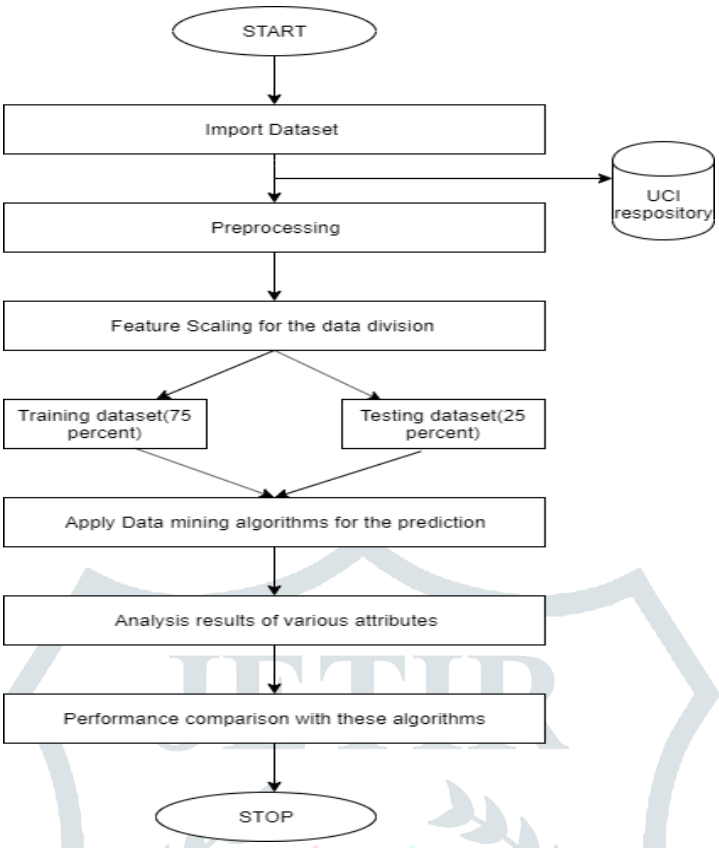

Figure 1 Overview of System

The technique, which is used for the prediction is as follows:

SVM (Support Vector Machine)

Support vector machine which separates the data into two categories to identify a maximum distance in hyperplane. Support vector machine algorithms apply with kernel that convert an input data into specific term. Support vector machine can be executed with classification, regression and outlier detection. It utilizes the quadratic optimization problem and easily extensible. It also helps to decrease the errors in data.

- Linear kernel: Dot product used by linear kernel between the two any considerations.
- Polynomial kernel: It is established form of linear kernel and characterize rounded and nonlinear input zone.
- Radial basis function (RBF) kernel: It is used to mapping the input slot in infinite volume space. In radial basis function, gamma range start from zero to one. Mostly Radial basis function used the value of gamma is 0.1 in support vector machine classification.

KNN (K Nearest Neighbors)

KNN algorithm is used for both predictive and descriptive methods. K Nearest Neighbors basically collect all the available data and distinguish the new data that depends upon distance function that is Euclidean distance. Euclidean distance is the separation between two slots that occur in graph and measure the distance between two slots. It also helps to calculate the K distinct number of neighbors. We can understand its working with the help of following steps:

1. For implementing any algorithm, we need dataset. So, during the first step of KNN, we must load the training as well as test data.
2. Next, we need to choose the value of K that is the nearest data points. K can be any integer.
3. For each point in the test data. Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean. The most commonly used method to calculate distance is Euclidean.
4. Now, based on the distance value, sort them in ascending order. Next, it will choose the top K rows from the sorted array. Now, it will assign a class to the test point based on most frequent class of these rows.

LOGISTIC REGRESSION

Logistic regression is also a sigmoid function that helps to predict the values. Logistic regression is a discriminative classification technique that works on real-valued input vector. The calculations of input vector to be categorized is known as features or predictors. Logistic regression can also be applied with various class classification. It is a method of fitting the best

line to the attributes present. It is used to predict other attributes one can use every attribute in the dataset. Logistic regression is an alternative method to use other than the simpler Linear Regression. It can also remove the improper data and predict the probability for a particular class. In this work, logistic regression can be used to predict the disease whether a person is diagnosed with disease or not. ROC (Receiver operating characteristics) is used in logistic regression. ROC curve basically represent the data in graphical form for every parameter and diagnosis the diseases.

PERFORMANCE METRICS

Various performance metrics are used to classify the heart disease such as accuracy, precision and recall. These parameters are help to measure the result for prediction and some terms are involve in parameters.

TP (True Positive): - True Positive is the result in which model predicts the positive values that are actually true.
TN (True Negative): - True Negative is the result in which model predicts the negative values that are actually true.
FP (False Positive): - False Positive is the result in which model predicts the positive values that are actually false.
FN (False Negative): - False Negative is the result in which model predicts the negative values that are actually false.

Three parameters are used for prediction in this work are as follows:
Accuracy: - It is the ratio of sum of true negative and true positive to the total samples in dataset.

$$\text{Accuracy} = (TP+TN)/\text{total} \quad \ldots\ldots \text{Eq (1)}$$

Precision: - It is the ratio of true positive to the sum of true positive and false positive.

$$\text{Precision} = TP/(TP+FP) \ldots\ldots \text{Eq (2)}$$

Recall: - It is the ratio of true positive to the sum of true positive and false negative.

$$\text{Recall} = TP/(TP+FN) \ldots\ldots \text{Eq (3)}$$

IV. RESULTS AND IMPLEMENTATIONS

The dataset is divided into two parts one is training part and other is testing part for implementations. All work is done in python and firstly import the dataset. After apply preprocessing to fill the missing values and check the null values then apply the various data mining techniques such as support vector machine, KNN and logistic regression. A confusion matrix is generated to measure the performance matrices for each technique and it is represented in two-dimensional form.

Table 2 Confusion Matrix

	1(has heart disease)	0(no heart disease)
1(has heart disease)	TP	FN
0(no heart disease)	FP	TN

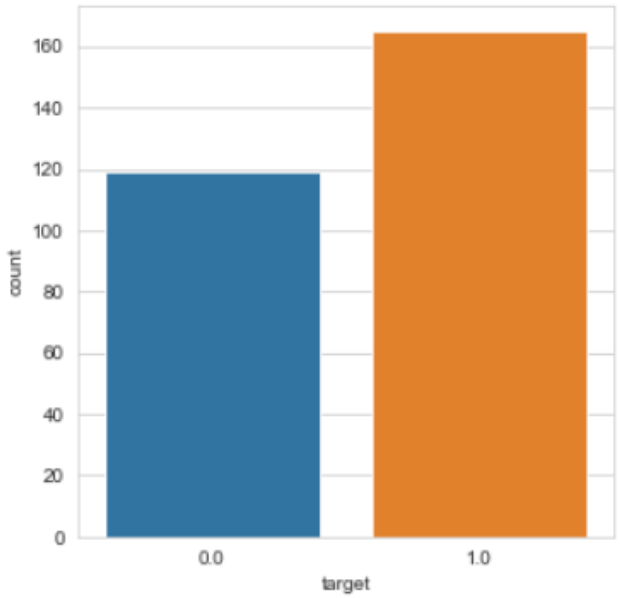

Figure 2 Target about disease

Above figure 2 represents the target value for heart disease in which 0 value follow how many patients are diagnosed with heart disease and 1 value follow how many patients are not diagnosed with heart disease. In this section, figure 3 describes the comparison between three algorithms with performance metrics and show the accuracy, precision and recall for three algorithms in bar chart. KNN algorithm depicts the good performance metrics as compared to SVM and logistic regression.

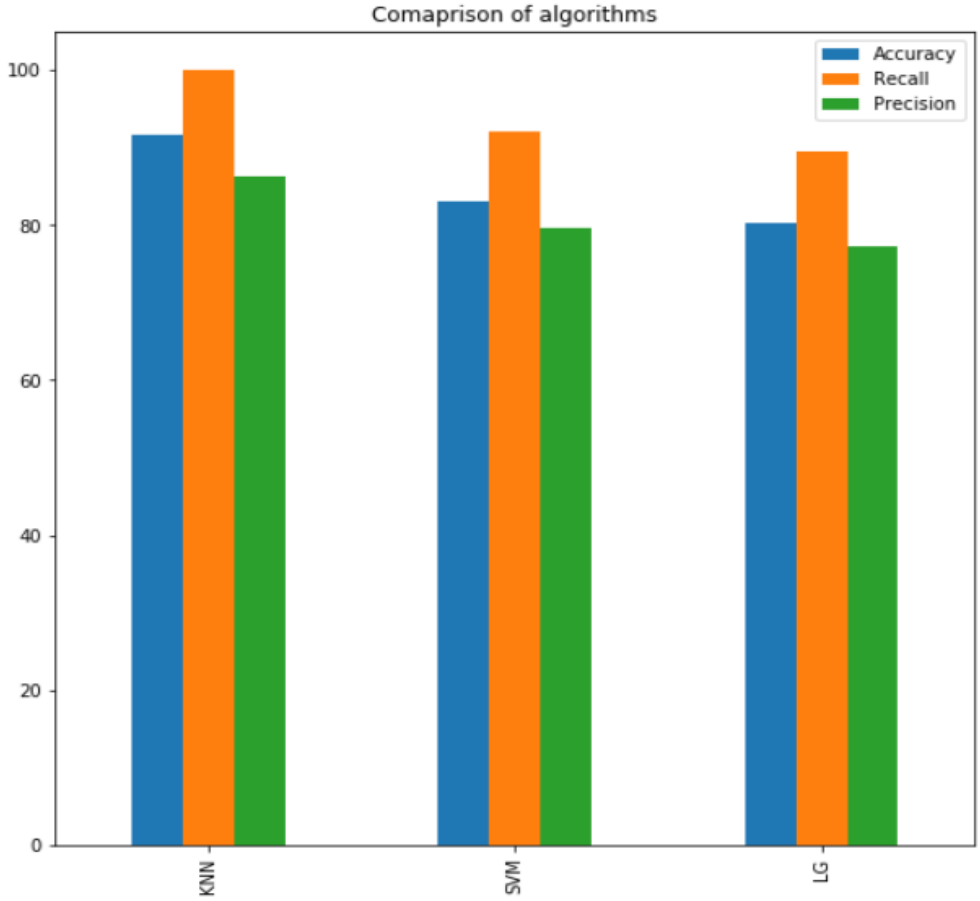

Figure 3 Representation of comparison

Table 3 Comparison of techniques

S.NO.	TECHNIQUES	ACCURACY (%)	PRECISION (%)	RECALL (%)
1.	KNN (k nearest neighbors)	91.54	82.23	99.02
2.	SVM (Support vector machine)	83.09	80.23	90.04
3.	Logistic regression	80.21	78.03	82.63

V. CONCLUSION

This work aims to predict the heart disease using classification techniques. It is difficult to manually the predict of heart disease based on risk factors and attributes. Data mining techniques are helpful to predict the output from existing data. Various data mining techniques that can be used to predict the disease and the good accuracy is obtained. Three algorithms are trained and tested with maximum scores as K Neighbors Classifier, Support Vector Classifier and Logistic regression Classifier. Out of these three algorithms, KNN obtained with goo accuracy and prediction the heart diseases with parameters.

REFERENCES

[1] R. Saxena, A. Johri, V. Deep, and P. Sharma, "Heart diseases prediction system using chc-tss evolutionary, knn, and decision tree classification algorithm," 2019.

[2] V. KSasipriya and Deepa, "Heart diseases detection using naive bayes algorithm," in IJISET-International Journal of Innovative Science, Engineering & Technology, 2, 441- 444., 2015.

[3] S. Manikandan, "Heart attack prediction system," in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pp. 817–820, Aug 2017.

[4] S. K. Mohan, "Effective heart disease prediction using hybrid machine learning techniques," in special section on smart caching, communications, computing and cybersecurity for information-centric internet of things(ieee), 2019.

[5] G. Purusothaman and P. Krishnakumari, "A survey of data mining techniques on heart disease prediction," in Indian Journal of Science and Technology, 2015.

[6] V. Cherian and B. M.S, "Heart disease prediction using naive bayes algorithm and laplace smoothing technique," in International Journal of Computer Science Trends and Technology (IJCST), 2017.

[7] O. Ahmed, A. Emad, and Kilan, "Effective diagnosis and monitoring of heart disease," International Journal of Software Engineering and its Applications,pages 143-156, vol. 9, pp. 143–156, 01 2015.

[8] J. Thomas and R. T. Princy, "2016 international conference on circuit, power and computing technologies (iccpct)," in Human heart disease prediction system using data mining techniques, 2016.

[9] K.S.Thirunavukkarsu, "Heart disease prediction using data mining techniques," International Journal of Advanced Research in Computer Science, 2017.