HEALTH INVESTIGATION SYSTEM

[1]Siddharth Mundra, [2]Kiran Manjrekar, [3]Nimit Lalwani, [4]Nilesh Rathod

[1,2,3]UG Student, [4]Assistant Professor, Rajiv Gandhi Institute of Technology, Mumbai, India,

[1]siddharthmundra1@gmail.com, [2]manjrekarkiran7@gmail.com, [3]nimit.lalwani@gmail.com, [4]nilesh.rathod@mctrgit.ac.in

Abstract: **Heart diseases have a copious pact of attention in medical research due to its impact on human health. Heart diseases are amongst the nation's leading cause of death. Data mining is the process which converts a collection of data into knowledge. Data mining has developed as a crucial approach for computing applications in medical area. Numerous algorithms connected with data mining have greatly helped to acknowledge medical data more evidently. In this work, supervised machine learning algorithms namely Random Forest, KNN and Naive Bayes are used to prediction of heart disease. The machine learning algorithms are implemented using Python programming language and UCI repository's data set is used which has 303 instances and 14 attributes. The performances of the algorithms are measured in terms of their accuracy. The goal of this study is to extract hidden patterns by applying data mining techniques, which are noteworthy to cardiovascular diseases and to predict the presence of heart disease in patients where this presence is valued from no presence to likely presence.**

Keywords — Data mining, cardiovascular disease, classification, naïve bayes, KNN, Random Forest

I. INTRODUCTION

Heart disease is one of the prevalent diseases that can lead to reducing the lifespan of human beings nowadays. Each yearly 18.5 million people are dying due to heart disease. Cardiovascular disease is a disease that affects the function of the heart. An estimate of a person's risk for heart disease is important for many aspects of health promotion and clinical medicine. A risk prediction model may be obtained through a multivariate regression analysis of a longitudinal study. Due to digital technologies are rapidly growing, healthcare centers store huge amount of data in their database that is very complex and challenging to analysis. Data mining techniques and machine learning algorithms play vital roles in the analysis of different data in medical centers. The techniques and algorithms can be directly used on a data set for creating some models or to draw vital conclusions, and inferences from the data set.

Data Mining is the study of substantial data sets to extricate the hidden and formerly unidentified patterns, relationships, and knowledge that are hard to investigate with conventional measurements. Data mining methods are the result of a long method of research and product improvement. Data Mining is divided into two assignments such as Predictive Tasks and Descriptive Tasks. Predictive Tasks forecast the estimation of an explicit attribute based on other attributes. Categorization, Regression come under Predictive Tasks. Descriptive Tasks design that outlines the connection between the data. Clustering, Association Rule Mining, and Pattern Discovery are future under Descriptive Tasks. Data Mining consists of few steps from raw data collection to some form of new knowledge. The iterative process comprises of following stages like Data Integration, Data cleaning, Data transformation, Data Selection, Data Mining, Knowledge Representation, and Pattern Evaluation.

Common attributes used for heart disease are Resting ECG (test that measures the electrical activity of the heart),Number of major vessels colored by fluoroscope, Fasting blood sugar, Exang(exercise included angina), Fasting Blood Pressure, ST depression (finding on an electrocardiogram, trace in the ST segment is abnormally low below the baseline),Threshold Pressure (high blood pressure), Age, Sex, Chest Pain type, Serum Cholesterol(determine the risk for developing heart disease), Thalach(maximum heart rate achieved) ,painless (chest pain location substernal=1, otherwise=0)),, smoke, Hypertension, Food habits, weight, height and obesity.

II. LITERATURE SURVEY

A lot of research works have been done on prediction of cardiovascular disease in the past. They used different data mining techniques for identification & attained different results for different data mining techniques.

G Purusothaman et al. [1], have studied and compared different classification techniques for cardiovascular disease prediction. In case of applying models such as Decision tree, artificial neural network and Naïve Bayes, the authors focus on the working of hybrid models i.e. models which uses more than one classification technique. They studied

DOI : 10.35291/2454-9150.2020.0076

the works of researchers who studied about the effectiveness of hybrid models. The performances of single models such as Decision tree, ANN and Naïve Bayes are 76%, 85% and 69% respectively. Though, hybrid approaches show an accuracy of 96%. Therefore, hybrid models lead to reliable and promising classifiers for predicting heart diseases with a better accuracy.

Ashish Chhabbi et al. [2], have studied different data mining techniques for discovering hidden patterns from a dataset that can answer queries in prediction of heart disease. The dataset has been collected from University of California, Irvine repository. They have used Naive Bayes and modified k-means algorithm. Experimental results show that modified k-means give better accuracy than simple k-means.

Boshra Baharami et al. [3], have evaluated different classification techniques such as J48 Decision tree, k-Nearest Neighbors(k-NN), Naive Bayes(NB) and SMO(SMO is widely used for training SVM). On the dataset feature selection technique is used to extract the important features. WEKA software has been used for implementing those classification algorithms. 10-fold cross-validation technique is used for testing the data mining techniques. J48 shows the highest accuracy of 83.732%.

Jayami Patel et al, [5], urged cardiopathy prediction using data mining and machine learning algorithmic program. The goal of this study is to extract hidden patterns by applying processing techniques. The best algorithmic program J48 supported UCI information has the most effective accuracy rate compared to LMT.

Ashwini shetty et al [6], suggested to develop the prediction system which is able to diagnose the heart disease from patient’s medical dataset. 13 risk factors of input attributes have taken into thought to form the system. After the analysis of knowledge from the dataset, information cleansing and information integration was performed.

Sharan Monica.L [7], planned an analysis of cardiovascular disease. This paper projected processing techniques to predict the diseases. It will provide the survey of current techniques to extract data from dataset and it will be useful for attention practitioners. The performance is also obtained on the basis of time taken to develop the decision tree for the system.

III. Data Set

The dataset we used in this paper is Cleveland dataset containing 303 instances. The dataset that is recently updated in the data world library is taken. This dataset contains total 76 attributes out of which only 14 attributes are used in this paper for prediction of heart disease.

Attribute Name	Attribute Type	Attribute Description
Age	Continuous	Numerical
Sex	Sex (1-male;0=female)	Numerical
Cp	Chest pain type (1:typical angia, 2:atypical angina, 3:non-anginal pain, 4:asymptomatic)	Numerical
trestbps	Resting blood pressure	Numerical
Chol	Serum cholesterol measured in mg/dl	Numerical
Fbs	Fasting blood sugar (>120mg/dl 1:true,0:false)	Numerical
restecg	Resting electrographic results [0-2]	Numerical
thalach	Heart rate maximum achieved	Numerical
exang	Induced angina due to exercise(1: yes,0:no)	Numerical
oldpeak	ST depression induced due to exercise relative to rest Numerical Slope Slope	Numerical
Slope	Slope of peak ST exercise segment(1:upsloping,2:flat,doensloping)	Numerical
Ca	No. of blood vessels colored by fluoroscopy(0-3)	Numerical
Thal	3:normal,6:fixed defect,7:irreversible defect	Numerical
Num	Diagnosis of heart disease (prediction attribute)	Numerical

Table 1: Dataset Attributes

IV. Data mining techniques used for prediction

Three different data mining classification techniques namely KNN, Naive Bayes and Random Forest are used to analyse the dataset.

4.1 Random Forest Algorithm

Random forest, like its name consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction. The output of the classes is represented by individual trees. This algorithm combines with random selection of features to construct a decision trees with controlled variations. The tree is constructed using algorithm given below.

Let M be the number of training classes and N be the number of variables in classifier.

- The input variable m is used to determine the node of the tree. Note that m<N.
- Choosing n times of training sets with the replacement of all available training cases M by predicting the classes, estimate the error of the tree.
- Choose m variable randomly for each node of the tree and calculate the best split.
- At last the tree is fully grown and it is not pruned. The tree is pushed down for predicting a new sample. When the terminal node ends up, the label is assigned the training sample. This procedure is iterated over all trees and it is reported as random forest prediction.

4.2 KNN

K nearest neighbors or KNN Algorithm is a simple and easy to understand algorithm which uses the entire dataset in its training phase. Whenever a prediction is required for an unseen data instance, it searches through the entire training dataset for k-most similar instances and the data with the most similar instance is finally returned as the prediction

- Determine the parameter K= number of nearest neighbors.
- Compute the distance between the query-instance and all the training samples.
- Sort the distance and determine nearest neighbours based on the k-th minimum distance.
- Gather the category r of the nearest neighbours.
- Use simple majority of the category of the nearest neighbours as the prediction value of the query instance.

4.3 Naïve Bayes

Naive Bayes classifier is a powerful algorithm for the classification. Even if we are working on a data set with thousands of records with some attributes, it is suggested to try Naive Bayes approach. A Naive Bayesian model is easy to build, and understand with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does extensively well and is widely used because it often outperforms more complicated classification methods.

Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

V. Results and Discussion

Python programming is used for the implementation of classification techniques. The UCI dataset consists of 302 records in the Heart diseases database. The data mining classification algorithms namely KNN, Naïve Bayes and Random Forest are implemented and compared for their accuracy. For the experimental grounds the dataset is been divided into training dataset and testing dataset in the ratio of 70:30 respectively.

In this study, the accuracy of three data mining techniques is compared. The goal is to have high accuracy, besides high precision and recall metrics. Although these metrics are used more often in the field of information retrieval, here we have considered them as they are related to the other existing metrics such as specificity and sensitivity. In machine learning, a confusion matrix, also called an error matrix, is a specific table design that permits insights of the execution of a calculation. Each line of the matrix speaks to the examples in an anticipated class while every section speaks to the cases in a real class. Class 0 represents heart diseases and Class 1 represents no heart disease.

Receiver Operating Characteristics (ROC) curve is a curve featuring true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the “ideal” point - a false positive rate of zero, and a true positive rate of one. This is not very realistic, but it does mean that a larger area under the curve (AUC) is usually better. AUC-ROC curve is one of the most commonly used metrics to evaluate the performance of machine learning algorithms and for measuring the accuracy, Accuracy classification score(Accuracy_score) from sklearn metrics is used.

- For KNN algorithm

	Class 0	Class 1
Class 0	27	17
Class 1	13	34

Table 2: Confusion matrix of KNN Algorithm

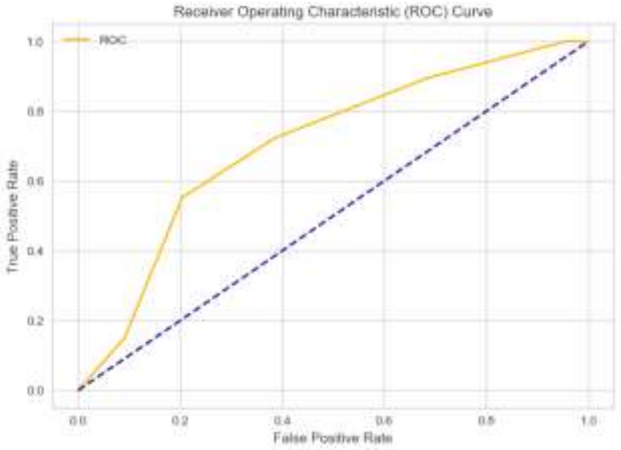

Fig 1: ROC curve for KNN Algorithm

Area under curve for KNN: 0.71
In other words, our model is 71% accurate for instances and their classification.

- For Random Forest Algorithm

	Class 0	Class 1
Class 0	32	12
Class 1	6	41

Table 3: Confusion matrix of Random Forest Algorithm

DOI : 10.35291/2454-9150.2020.0076

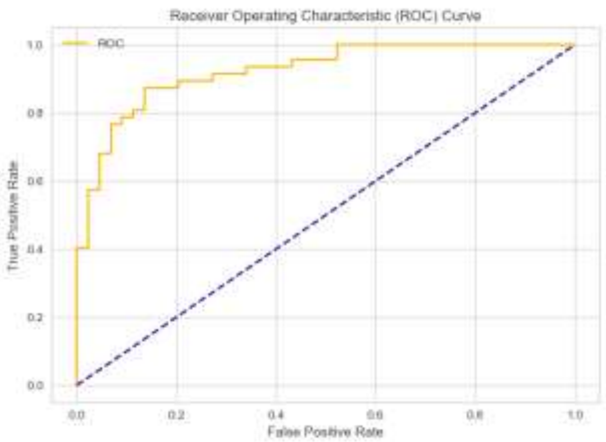

Fig 2: ROC Curve for Random Forest Algorithm

Area under curve for Random Forest: 0.93

In other words, our model is 93% accurate for instances and their classification.

- For Naïve Bayes Algorithm

	Class 0	Class 1
Class 0	32	12
Class 1	5	42

Table 4: Confusion matrix of Naïve Bayes Algorithm

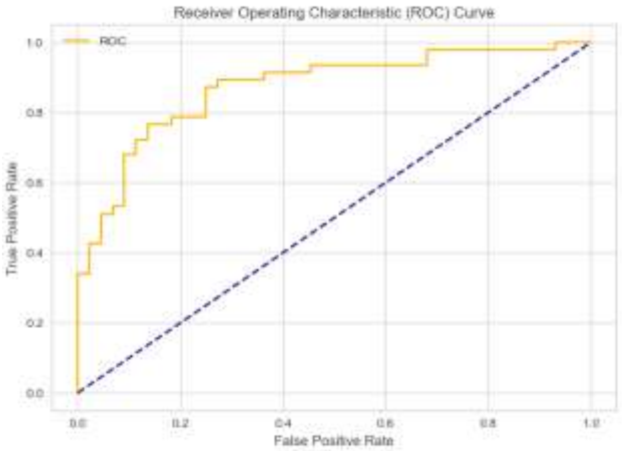

Fig 3: ROC Curve for Naïve Bayes Algorithm

Area under curve for Naïve Bayes: 0.87

In other words, our model is 87% accurate for instances and their classification.

- Accuracy comparison for the three algorithms using accuracy_score(sklearn metrics)

Classification	Accuracy
Naive Bayes	80.2197
Random Forest Algorithm	81.3198
KNN	67.021

Table 5: Accuracy of the classification algorithms

From the experimental results, it was found that Random Forest Classifier has the best accuracy of 81.31% when it comes to prediction of heart diseases and therefore can be used for the prediction purpose based on the 14 attributes as mentioned in the dataset.

VI. Conclusion

In this paper, the heart disease dataset of UCI repository is taken and subjected to various classification and clustering algorithms using python. The main focus is to target all possible combinations of the attributes against various algorithms. Then of all the techniques it is the technique that works the best to predict the heart disease at an early stage is identified. Three supervised machine learning algorithms namely KNN, Naive Bayes and Random forest are compared in terms of accuracy using the UCI heart diseases dataset. From the experimental results, it's found out that Random Forest algorithm predicts the heart disease with the accuracy of 81.319%. In future, the performance of Random Forest algorithm can be compared with various other classification algorithms and hybrid algorithms.

References

[1] G. Purusothaman, and P. Krishnakumari, June 2015,"A Survey of Data Mining Techniques on Risk Prediction: Heart Disease", Indian Journal of Science and Technology, Vol. 8(12), DOI:10.17485/ijst/2015/v8i12/58385, pp. 1-5.

[2] Ashish Chhabbi,Lakhan Ahuja,Sahil Ahir, and Y. K. Sharma,19 March 2016,"Heart Disease Prediction Using Data Mining Techniques", International Journal of Research in Advent Technology,E-ISSN:2321-9637,Special Issue National Conference "NCPC-2016", pp. 104-106.

[3] Boshra Bahrami, and Mirsaeid Hosseini Shirvani,February 2015,"Prediction and Diagnosis of Heart Disease by Data Mining Techniques",Journal of Multidisciplinary Engineering Science and Technology(JMEST), ISSN:3159- 0040, Vol. 2, Issue 2, pp. 164-168.

[4] Siddharth Mundra, Kiran Manjrekar, Nimit Lalwani, Nilesh Rathod. Review on prediction of heart disease using data mining, International Journal of Advance Research, Ideas and Innovations in Technology5.6 (2019), www.IJARIIT.com.

[5] Jayami Patel, Prof. Tejal Upadhay, Dr. Samir Patel, "Heart disease Prediction using Machine Learning and Data mining Technique", March 2017.

[6] Ashwini Shetty A, Chandra Naik, "Different Data Mining Approaches for Predicting Heart Disease", International Journal of Innovative in Science Engineering and Technology, Vol.5, May 2016, pp.277- 281

[7] Sharan Monica.L, Sathees Kumar.B, "Analysis of CardioVasular Disease Prediction using Data Mining Techniques", International Journal of Modern Computer Science, vol.4, 1 February 2016, pp.55-58.

DOI : 10.35291/2454-9150.2020.0076