Prediction of Hepatitis Disease Using Machine Learning Algorithm

[1]Prof. Satish Manje, [2]Mr. Ansari Md Talha, [3]Mr. Gaikwad Milind , [4]Mr. Pandey Aniruddh.

[1]Asst.Prof.,[2,3,4]UG Student,[1,2,3,4]Computer Engg. Dept. Shivajirao S. Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India.

[1]satishmanje93@gmail.com, [2]mdtalha543210@gmail.com, [3]milindgaikwadm23@gmail.com, [4]annabhai2001 @gmail.com

Abstract - Hepatitis is currently one of the worst diseases that kill people all around the world. The human liver's inflammation is brought on by it. If people are successful in identifying this dangerous condition early on, can prevent many individuals from dying from it. This project review a variety of data mining methods is used to predict the hepatitis disease[1]. In addition, project offered a respectable method for enhancing the effectiveness of prediction models. For the hepatitis disease sample dataset, different classification algorithms are used to calculate prediction accuracy. F1-score, precision, recall, accuracy, and ROC are calculated to compare the performance of categorization models. The algorithm which gives the highest accuracy will be the best classification algorithm.

Keywords- Data Mining, KNN, Naive Bayes, SVM, MLP, Random Forest.

I. INTRODUCTION

The liver is a crucial organ in humans, which participates in a variety of biological functions. The cause of liver damage is hepatitis. Due to this reason, a patient can die. In medical science, the detection of hepatitis disease within a patient's body at an early stage is a challenging task. At present, the medical industry can see that day-by-day amount of data related with health is increasing. Data mining is a field related to machine learning has the ability to manage huge data as well as solve the complex problem very efficiently so that researchers can take correct decision from a huge database. It is applied to identify unknown patterns and find out valuable information from a huge dataset [2]. However, health care industries gather significant information from different clinical reports and patient's diagnostic test results. It is applied to know the class name from the dataset by noticing the unseen pattern along with correlated features present in dataset. The correlated features helps to distinguish either the patient is affected by hepatitis disease or not. Its working approach has the similarity with an expert system. Besides this, it will also save cost and diagnosis time. However, there are many machine learning algorithms that are used for prediction purposes. It is a difficult work for us to find out the best technique. This project have applied Naive Bayes, Support Vector Machine, KNN, Random Forest and Multi-Layer Perceptron in order to predict hepatitis disease is present or not in the patient's body. This project have collected real world diagnostic datasets of hepatitis disease having different types of features of 155 patients from the UCI repository of machine learning [1]. Secondly, have found out the unnecessary features and only highly related features are selected to increase the classification model performance [1]. And at last made a performance comparison of our five techniques as well as compare the performance result with the previous research result and also evaluate the prediction outcome based on different risk factors.

II. AIMS AND OBJECTIVE

a) Aim

The main aim of the project is to detect and classify the accuracy to predict hepatitis disease [1]. There are many methods to do so. Testing and detecting various factors affecting classification of dataset, changes can be done in the future to help better understand this social problem. This project is going to use different algorithms for classification and prediction. The project also focuses to compare the various classification algorithms depending upon the performance factors.

b) Objective

The main motive of this project is to detect and classify the accuracy of different algorithm use to predict the hepatitis disease on the base of the data using different algorithm. The classification of data set is and various algorithm is applied and the algorithm with better performance factor will be the best algorithm to predict hepatitis[1]. This will help in decreasing death rate and making the diagnostic process faster. All this classification of data and prediction will be accurate and cost friendly and will help in taking better decision.

 DOI : 10.5281/zenodo.7857144

III. LITERATURE SURVEY

Paper 1: An empirical analysis of decision tree algorithm: Modeling Hepatitis data:

This paper describes the classification of dataset to detect disease. Compared with most commercial methods, data mining which is the procedure of gaining knowledge from huge databases by detecting patterns. The performance of seven decision tree classification methods is highlighted in this research. J48, Hoeffding Tree, Logistic Model Tree[LMT], REP (Reduced Error Pruning) Tree, Random Forest, Decision Stump and Random Tree on the Hepatitis prognostic dataset that enables the classifier to accurately carry out categorization of medical data[6]. The classification accuracies are evaluated using cross validation method using 10 folds.

Paper 2: Prediction of hepatitis prognosis using Support Vector Machines, Wrapper Method:

Patients with hepatitis require ongoing specific medical therapy to lower their fatality rate. The motive of this project was to explore the link between hepatitis prediction and SVM. Using clinical test findings data and machine literacy technology similar as SVM, the bracket and vaticination of their life prognostic can be done. Still, cannot pledge that all the features values in the data are identified to each other. Therefore, embodgy Wrapper Methods for removing noise quality before classification. This study demonstrates how integrating the feature selection strategy with the classification process improves prediction between data. Thus it has given improved results using SVM method[8].

Paper 3: Application of CART algorithm in Hepatitis disease diagnosis:

The healthcare sector gathers a huge quantity of data which is not perfectly mined and not put to the ideal use. Locating these hidden ornaments and co-relation often goes unutilized. This paper focuses on this feature of Medical prognosis by learning ornaments through the gathered facts of hepatitis and for developing clever medical decision assist structure to help the doctor. In this paper, the author proposes the use of C4.5, ID3 and CART classifier [7] to analyze these diseases and differentiate the effectiveness and compare them.

IV. EXISTING SYSTEM

Health care industries gather significant information from different clinical reports and patient's diagnostic test results. It is applied to know the class name from the dataset by noticing the unseen pattern along with features available in the dataset. Its working approach has the similarity with an expert system. There were many different features used which may not have that importance in determining the hepatitis [1]. There are many machine learning algorithms that are used for prediction and classification purposes but has less classification accuracy.

V. COMPARATIVE STUDY

Table.1: Comparative Analysis

Sr. No	Paper Name	Author/ Publication	Technology	Advantage	Disadvantage
1.	An empirical analysis of decision tree algorithms: Modeling hepatitis data.	Manickam Ramasamy; Shanthi Selvaraj; M. Mayilvaganan. (IEEE,2015)	Decision Stump, Hoeffding Tree, J48, LMT, Random Forest.	Examine how various decision tree effect the classification accuracies.	Super high maintance cost.
2	Prediction of hepatitis prognosis using Support Vector Machines and Wrapper Method.	A.H. Roslina; A. Noraziah. (IEEE,2010)	SVM; Wrapper Method	To predict the life expectancy for patients with hepatitis.	Applications are limited only for study purpose.
3.	Application of CART algorithm in hepatitis disease diagnosis.	G.Sathyadei. (IEEE,2011)	CART algorithm, C4.5, ID3	This paper purposes to classify dataset and develop decision support.	Difficult to adapt/implement.

VI. PROBLEM STATEMENT

There are several applications and websites available to detect and predict hepatitis. However, most of them are not free and have terrible user interface and the subscription cost is high, because of high cost the service became less accessible to user. Software like this should be and must be available to everyone whoever needs it. To develop a project that use machine learning approaches and algorithms for the diagnosis of hepatitis disorders using several algorithms like SVM, Naive Bayes and K-Means algorithm. Early identification of hepatitis illness is critical for effective therapy. Because of the basic precise symptoms, it is difficult for medical experts to foresee the disease in the early stages. This warning is frequently missed until it is too late. To address this issue, machine learning technologies must be used to enhance hepatitis illness prediction.

VII. PROPOSED SYSTEM

In this project, Support Vector Machine, KNN, Naive Bayes, Multi-Layer Perceptron, and Random Forest algorithms are used to predict whether or not hepatitis disease exists in the patient's body. For the hepatitis illness dataset, classification algorithms such as K-Nearest Neighbor , Naive Bayes,

 DOI : 10.5281/zenodo.7857144

Support Vector Machine (SVM), Multi Layer Perceptron (MLP), and Random Forest are used to calculate F1-score, precision, recall, ROC and accuracy to compare the performance of categorization models. There are different attributes in the dataset, they are: age, sex, protime, antivirals, fatigue, sgot, anorexia, liver big, varices, spleen palpable, histology, ascites, liver firm, bilirubin, alk phosphate, malaise, albumin, steroid, spiders. The dataset is passed to all the algorithms and the performance measure is calculated by the result given by the algorithms. [2] The algorithm with highest accuracy will be decided on the basis of performance measure accuracy.

VIII. ALGORITHM

The Algorithm for Prediction of hepatitis disease:

Step.1: Start
Step.2: Prediction using Random Forest :

```
df = pd.read_csv('hepatitis.csv')
model = RandomForestClassifier()
model.fit(X_train, y_train)
y_pred = model.predict(X__test)
```

Step.3: Prediction using SVM:

```
model2 = SVC()
model2.fit(X_train, y_train)
y_pred2 = model2.predict(X_test)
```

Step.4: Prediction using Naïve bayes:

```
model3 = GaussianNB()
model3.fit(X_train, y_train)
y_pred3 = model3.predict(X_test)
```

Step.5: Prediction using KNN:

```
mmodel4 = KNeighborsClassifier()
model4.fit(X_train, y_train)
y_pred4 = model4.predict(X_test)
```

Step.6: Prediction using Multi-Layer Perceptron:

```
model5 = MLPClassifier()j
model5.fit(X_train, y_train)
y_pred5 = model5.predict(X_test)
```

IX. MATHEMATICAL MODEL

1) K--Nearest Neighbor [KNN]

This classifier performs classification in three steps. In step-1, it computes K-value. In step-2, for each test sample it computes the distance between all the training data as well as sorts it and finally in step-3, the class name will be provided to the test sample data by applying majority voting approach. The Euclidean distance is computed by:

$$E_d = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

2) Naive Bayes:-

This technique works by using Bayes theorem and it is used for classification purposes. This classifier model is easy to build. We can find out the likelihood of any event that is occurring given the probability of another event that has already occurred by taking the help of Bayes Theorem. The posterior probability is computed by:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

3) Support Vector Machine [SVM]:-

SVM is supervised machine learning approach is popular for both classification as well as regression purposes. This paper, have use this technique for classification purposes. At first, it recognizes patterns from training data. Then it divides the data points properly into their classes by finding a maximal margin hyperplane. This hyperplane helps us in order to predict on test data.

4) Multi-Layer Perceptron (MLP):-

This supervised machine learning approach is popular both in terms of regression and classification purposes. This paper, have use this technique for classification purposes. At first, it trains the dataset using backpropagation method. The gradient descent function which is used to train the model is calculated by using backpropagation. After completing the training phase, the model can predict the class name for the new test sample. A MLP network contains three layers. At first, input layer which accepts input. Next layer is hidden layer. It can be one or more in number. Finally, the output layer, it generates the results after classification.

5) Random Forest:-

This is a supervised machine learning approach which is popular for both classification as well as regression purposes. This project have use it for classification purposes. It works in three steps. In step-1 in the learning phase, a forest of Decision Trees are produced from a number of trees. In step-2 for each test data, the trees which are used to make a forest in previous step predicts a class name. In the last step which is step-3, the correct class name is assigned to test data based on majority of votes .Each of the data present in a dataset is faced the step-3.

X. SYSTEM ARCHITECTURE

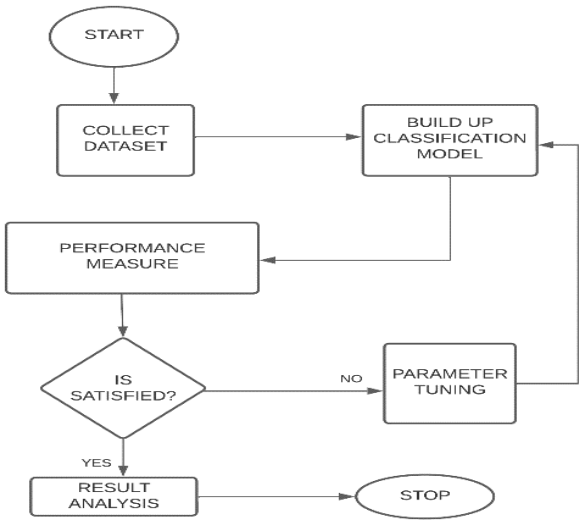

Fig.1: System Architecture

 DOI : 10.5281/zenodo.7857144

Explanation: The dataset taken from the repository will contain all the features and their respective values of the patient, from the dataset the correlated features that plays vital role in detection of disease will be selected. The selected dataset will be passed to all the classification model and the performance measure will be calculated for each models. Accuracy, precision, F1 score, recall, ROC will be calculated and will be used for analysis. The model with highest performance measure will be the best classifier.

XI. ADVANTAGES

• This model can predict if a person has hepatitis or not.

• It reduces the time complexity of doctors and also cost effective for patients.

• Delivers faster and accurate results in order to identify profitable opportunities.

• Helps solve complex real world problems with several constraints.

• Provide a path towards accomplishing Artificial Intelligence some day in the future.

• No human supervision is required.

• Provides faster and more reliable findings to discover profitable possibilities.

XII. DESIGN DETAILS

Fig 2: Result

The project aims to find the classification algorithm to predict hepatitis disease with the help of machine learning algorithms. Different machine learning models are used in this project for classification and prediction purpose. The result analysis is done on the basis on the scores obtained by classification models of performance measures (Accuracy, precision, F1 score, recall, ROC). Each models have different score. The model with the overall highest score will be the best classifier and will be used for prediction.

XIII. CONCLUSION

Thus, we have tried to implement the paper “Md. Julker Nayeem; Farjana Alam; Md. Ataur Rahman, Sohel Rana " Prediction of Hepatitis Disease Using K-Nearest Neighbors, Naive Bayes, Support Vector Machine, Multi-Layer Perceptron and Random Forest " 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) Accession Number:20633561DOI:10.1109/ICICT4SD50815.2021.9397013.” and according to the implementation the conclusion is to get the best classifier among our classification models. The less contribution features present in dataset may be the reason for poor classification accuracy so its better to remove them. By removing observations from the dataset as well as only selecting correlated feature from dataset, each of the classification algorithms gives a remarkable performance.Among the five classifiers, Multilayer perceptron has shown better performance with the highest accuracy of 93.16%.

REFERENCE

[1] Md. Julker Nayeem; Sohel Rana; Farjana Alam; Md. Ataur Rahman" Prediction of Hepatitis Disease Using K-Nearest Neighbors, Naive Bayes, Support Vector Machine, Multi-Layer Perceptron and Random Forest " 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) Accession Number:20633561DOI:10.1109/ICICT4SD50815.2021.9397013.

[2]Prof. Vishal R. Shinde “Machine learning algorithm for stroke disease classification” in IJREAM, ISSN: 2454-9150, Volume 08, Issue 01, APR 2022 Special Issue

[3] K. S. Bhargav, T. D. Kumari, D. S. S. B. Thota, and V. B "Application of Machine Learning Classification Algorithms on Hepatitis Dataset." *International Journal of Applied Engineering Research*, vol. 13, no. 16, pp. 12732-12737, 2018

[4] S. Hashem, G. Esmat, W. Elakel, S. Habashy, S. A. Raouf, M. Elhefnawi, M. El-Adawy, and M. Elhefnawi "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients." *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15 no. 3, pp. 861-868, 2018.

[5] S. M. M. Hasan, M. A. Mamun, M. P. Uddin, and M. A. Hossain, "Comparative Analysis of Classification Approaches for Heart Disease Prediction," IEEE International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, pp. 1-4, 8-9 February. 2018.

[6] M. Ramasamy, S. Selvaraj, and Dr. M. Mayilvaganan. "An empirical analysis of decision tree algorithms: Modeling hepatitis data." IEEE International Conference on Engineering and Technology (ICETECH), India, pp. 1-4, 20 March. 2015.

[7] G. Sathyadevi"Application of CART algorithm in hepatitis disease diagnosis." IEEE International Conference on Recent Trends in Information Technology (ICRTIT), India, pp. 1283-1287, 3-5 June. 2011.

[8] A. H. Roslina, and A. Noraziah. "Prediction of hepatitis prognosis using Support Vector Machines and Wrapper Method." IEEE Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), China, Vol. 5, pp. 2209-2211, 10-12 August. 2010.

 DOI : 10.5281/zenodo.7857144