International Journal of Innovative Research in Science and Engineering
Volume 7, Issue 12, 2021
www.ijirse.com
ISSN: 2454-9665

HEPATITIS DISEASE ANALYSIS USING MACHINE LEARNING TECHNIQUES

PARTHASARATHI P[1] ,AKALYA S[2] , DHARINI P[3], GAYATHRI S[4]

[1,2,3,4] *CSE, Bannari Amman Institute Of Technology, (India)*

ABSTRACT:

Viral hepatitis is one of the dreadful diseases in the world.Due to hepatitis B virus more than 350 million people get affected .Hepatocytes in the liver would get damaged due to hepatitis B virus . So far, many studies have been performed in the diagnosis of hepatitis disease.Expert doctors can only do the visual task as it is very difficult to diagnose this virus .We can analyze this virus through machine learning algorithms. The main objective is to analyze the disease using three machine learning algorithms. The dataset has already been analyzed using Random forest , MLP/neural network and K-nearest neighbors algorithms. It is found that K-nearest neighbors performs better than the other two algorithms. In this project Random Forest algorithm is chosen for comparison with K-nearest neighbors algorithm to choose an efficient one for diagnosis.

Keywords**: **Hepatitis B virus, Random Forest, K-nearest neighbor, neural network, Data Mining, Machine learning.

I.INTRODUCTION

Hepatitis B (HB) is a disease caused by the hepatitis B virus (HBV). It affects the liver causing acute and chronic infections as well. People do not find the symptoms in the early stage. In case of acute infection people tend to be sick associated with pain in the abdomen, tiredness, and yellowish skin. The early infection may sometimes result in death and sometimes it may last for weeks. This virus mainly spreads through body fluids of the affected person to the healthy person. Hepatitis B virus symptoms won't be visible during the initial stage it may cause jaundice, stomach pain, etc.The main objective is to analyze the hepatitis dataset using machine learning algorithms and finding the efficient algorithm. This can be done by calculating the accuracy score of each algorithm and compared to the others.

II.PROJECT DESCRIPTION

Nowadays, effective business decisions are taken by extracting valuable information from the raw data The need of processing and exploring the useful information obtained from raw data has arisen in many fields of life; business, medicines, science, and engineering. Today's intelligence technologies analyze the data, explore the information and then convert the information into knowledge. At that point, Data Mining (DM), Machine Learning (ML) mainly used as it accurately extract the information from the huge amount of data.Data mining is used to find out the missing values or hidden patterns from the huge amount of data . DM is effectively analyzing a large amount of data, complex data that contain multiple variable and non linear relations. The

ISSN: 2454-9665

performance of ML algorithms and DM approaches are analyzed on the hepatitis dataset. We then compare the performance of ML classifiers in terms of accuracy and elaborate which techniques effectively analyze the data.

III.OBJECTIVE

To analyze and predict the Hepatitis B Virus data to get high accuracy.

3.1 SYSTEM REQUIREMENTS

3.1.2 HARDWARE REQUIREMENTS:

Processor: 1.1 GHz Dual-Core Intel Core i3

Memory: 1TB

RAM: 8 GB

3.1.3 SOFTWARE REQUIREMENTS:

OS : Windows 10

Browsers : Google, Mozila Firefox

Online Compiler : Google colab

Python packages : Pandas, Numpy, Matplotlib, Seaborn

IV. DATA WRANGLING

We have used the Weka tool for verifying the data as it should not contain any missing values and also it should not have any null values.

4.1.2 PREREQUISITES

- Pandas
- Scikit-learn (Sklearn)
- Numpy
- Matplotlib
- Random forest algorithm
- K-Nearest Neighbors algorithm
- Neural network algorithm

Here we have used google colab to run our source code.

4.1.3 DATASET

Here we have used hepatitis 2.csv dataset which consists of 142 rows and 20 columns(class,age,sex,steroid,antivirals,fatigue,malaise,anorexia,liver_big,liver_firm,spleen_palable,spiders,ascites,varices,bilirubin,alk_phosphate,sgot,albumin,protime,histology). The data in the dataset is from 01/06/2020 to 05/10/2021.

Let us discuss the Python libraries used in Hepatitis B virus to analyze the data.

4.1.4 PANDAS

Pandas module is python library It works with tabular data. It is used to evaluate data.

4.1.5 SKLEARN

Sklearn is a most used library in machine learning for statistical output in which it includes classification and

ISSN: 2454-9665

regression .It is used for grouping the unlabelled data

4.1.6 NUMPY

NumPy is a python package to make mathematical operations, that is it performs scientific estimations.

4.1.7 MATPLOTLIB

Matplotlib is a plotting library used in python for data visualization.It is used to create static,animated and interactive visualizations using python.

4.1.7 RANDOM FOREST ALGORITHM

A random forest is used in regression and classification. From a dataset divided it into different modules to form like random forest majority out of those modules will be taken. From that decision tree we are taking the majority one.

4.1.8 K-NEAREST NEIGHBORS ALGORITHM

K-Nearest Neighbors (KNN) algorithm is a classification and regression algorithm. It need some reference and data records. Here also majority will be taken

4.1.9 NEURAL NETWORK

Neural networks are composed of a number of layers of nodes. Each node is called a perceptron which works similar to neurons and is fired based on the activation functions like ReLu or Sigmoid and finally calculates the output. When in the training stage it uses a backpropagation method and adjusts the weights and bias accordingly using the optimizers specified.

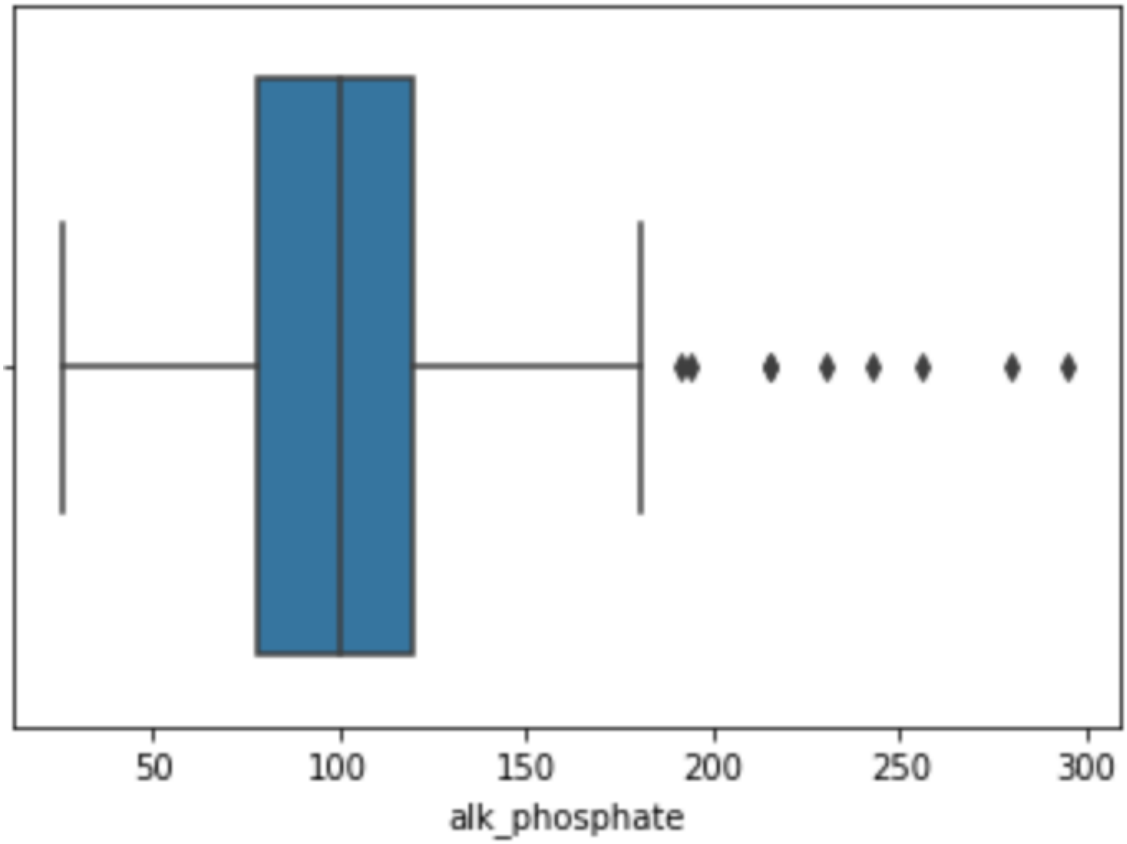
V. METHODOLOGY PROPOSED

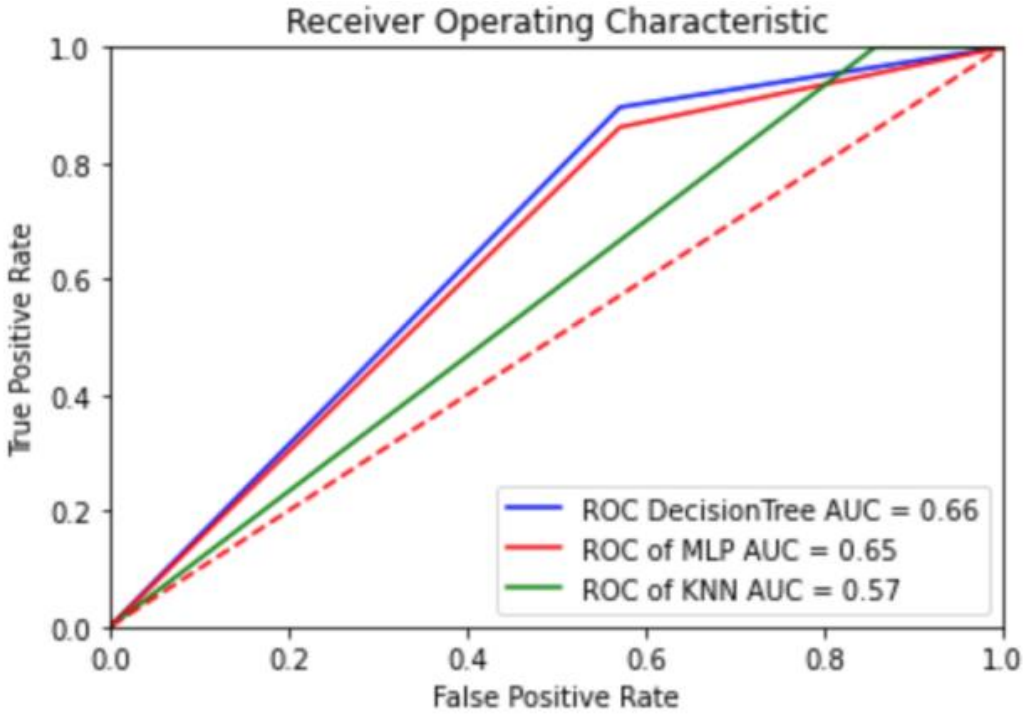
VI. ANALYSIS RESULT

Comparison of random forest k-nearest neighbors and neural networks are made to find the best out of them among those three algorithms K-nearest neighbor has the highest accuracy

	Algorithms	Accuracies	Area Under the Curve	Recall	Precision
2	K Nearest Neighbor	83.333333	57.142857	0.828571	1.000000
0	Random Forest	80.555556	66.256158	0.866667	0.896552
1	MLP/Neural Network	77.777778	64.532020	0.862069	0.862069

International Journal of Innovative Research in Science and Engineering
Volume 7, Issue 12, 2021
www.ijirse.com
ISSN: 2454-9665





ISSN: 2454-9665

VII. CONCLUSION

The performance comparisons of the two classifiers are presented. The classifier performance was based on the percentage of the Correctly Classified Instances or Accuracy. In this performance analysis, K-nearest neighbor classifier achieved higher accuracy of 83.3% with minimum execution time in classifying and predicting hepatitis infectious disease.

REFERENCE

[1] Li Sijia, Tan Lan, Zhuang Yu, Yu Xiuliang 2020 – 'Comparison of the prediction effect between the Logistic Regressive model and SVM model'

[2] P. R. Visali Lakshmi, G. Shwetha, N. Sri Madhava Raja 2021 – 'Preliminary big data analytics of hepatitis disease by random forest and SVM using r-tool'

[3] Fitriana Harahap, Ahir Yugo Nugroho Harahap, Evri Ekadiansyah et al. 2018 'Implementation of Naïve Bayes Classification Method for Predicting Purchase'

[4] VijayaKumar.K, Lavanya.B, Nirmala.I, Sofia Caroline.S et al. 2019 – 'Random Forest Algorithm for the Prediction of Diabetes'

[5] Mehrbakhsh Nilashi, Othman bin Ibrahim, Hossein Ahmadi, Leila Shahmoradi et al. 2019. 'An analytical method for diseases prediction using machine learning techniques, Computers & Chemical Engineering'.

[6] Amanpreet Singh, Narina Thakur**,** Aakanksha Sharma 2020 – 'A review of Supervised machine learning algorithms'.

[7]https://medium.com/@synced/how-random-forest-algorithm-works inmachine-learning-3c0fe15b667.