HEPATITIS DISEASE ANALYSIS USING MACHINE LEARNING ALGORITHMS

Akshaya K[1], AbdulMajeed M [2], Divya A [3], Ramya R[4]

[1,2,3] UG Student, Bannari Amman Institute of Technology, Sathyamangalam.

[4]Asst.Professor,Bannari Amman Institute of Technology, Sathyamangalam.

----------***----------

Abstract: Hepatitis B is a liver inflammation disease which can cause both acute and chronic effects through viral infection. The virus is generally transmitted through blood contact or contact with other fluids in the body. It is also transmitted from mother to child during birth. As of 2016, WHO estimates that 27 million people were living with Hepatitis B and were also aware of their infections while 4.5 million of the people diagnosed were on treatment. Hepatitis B can be prevented by vaccines through which the development of complications like chronic disease can be prevented. This immediate treatment requires a lot of medical data analysis. The disease has to be confirmed by various health factors and has to be analyzed to know the presence of disease. Manual analysis requires a lot of time and only medical experts can do it within a stipulated time. Data processing technique is a necessary way of approach to find a solution for this drawback. A number of machine learning algorithms are used for data analysis. Python is employed for analyzing the Hepatitis dataset. Jupyter Notebook is the data processing tool used to predict the presence of disease. Information mining techniques like Logistic Regression, Support Vector Machine, Naive Bayes and Random Forest algorithms are used. Additionally, the better performing algorithm for analysis of the Hepatitis dataset is known through accuracy factors for better diagnosis. It integrates the work of assorted authors in one place and therefore it is helpful for researchers to induce the information of knowledge mining techniques.

1.INTRODUCTION

Viral hepatitis is one of the most important infectious diseases in the world. It is caused by a virus namely hepatitis B which attacks and injures the liver. Hepatitis B causes the most common and serious liver infections. It is an inflammation which tends to damage the hepatocytes in the liver caused by at least six different viruses has some unusual features similar to retrovirus and has a small DNA structure. It belongs to the family of Hepadnaviridae and is a prototype. On comparing its sequence, the virus is classified into eight genotypes from A to H. Each of them has a different geographic distribution.

The virus is transmitted through blood and bodily fluids. It can be spread to others through blood, unprotected sex, illegal drug usage, needles that are unsterilized, and from an infected woman to her newborn during pregnancy or childbirth. Hepatitis B is silent epidemic because most of the people will not have any symptoms when they infected. By this the virus can be spread unknowingly and silently. For those who are severely affected and doesn't incur any symptoms, their liver will still be damaged and can be serious causing liver cirrhosis or liver cancer. It is estimated

that about 1 million people die from hepatitis B each year despite the fact that it is preventable and treatable. The only way to know the conformal of infection is to do a blood test and continue further diagnosis. Many studies are being performed in the diagnosis. Medical diagnostics is quite difficult and visual task is mostly done by expert doctors. The automatic analysis can be approached by using machine learning algorithms. Algorithms like Logistic Regression, Support Vector Machine, Naive Bayes and Random Forest algorithms are used. The effectively performing algorithm is chosen and is used for better diagnosis of disease**.**

2.METHODOLOGY

1.DATA EXTRACTION

The hepatitis dataset is extracted from uci repository. The parameters included within the dataset are class, age, sex, steroid, antivirals, fatigue, malaise, anorexia, size of liver, firmness of liver, bilirubin, alkaline phosphate, SGOT, albumin, spleen palpable, spiders, ascites, varices, protime and histology.

2.DATA WRANGLING

Data wrangling is the process of cleaning, enriching and restructuring the available raw data and formatting into a more usable data. This helps analyzers in quick decision making, and thus get better insights in a short span. Organizing and cleaning data before analysis is extremely useful and helps the firms in quick analysis of larger amount of data. The data has to be analyzed and cleaned. Parameters with no values has to be chosen and filled. It is done manually.

LOGISTIC REGRESSION

Logistic Regression is a Machine Learning algorithm used for the classification problems and regression problems. It is a predictive analysis algorithm and based on the concept of probability. Logistic regression is used to assign observations to a discrete set of classes. Some of the examples of classification problems are to find if the Email is spam or not, to find if the online transactions Fraudulent or not, Tumor being Malignant or Benign. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.

SUPPORT VECTOR MACHINE

In machine learning, support-vector machine are supervised learning models. It is associated with other learning algorithms that analyse the data used for analysing classification and regression problems. Given a set of training examples, each being marked as belonging to any of the two categories, the SVM training algorithm builds a model that assigns new examples to one or the other category, which makes it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

NAIVE BAYES

The Naive Bayes classification technique is based on Bayes theorem assuming that the predictors are independent of each other. A Naive Bayes classifier assumes that a feature in a class is unrelated to any other features. The algorithm first creates a frequency table of all classes. It then creates a likelihood

table. Finally, it calculates the posterior probability. On analysis and comparison of logistic regression algorithms, support vector machine algorithms with Naïve Bayes algorithm on applying it on hepatitis dataset it is found that Naïve Bayes performs better in analyzing the data.

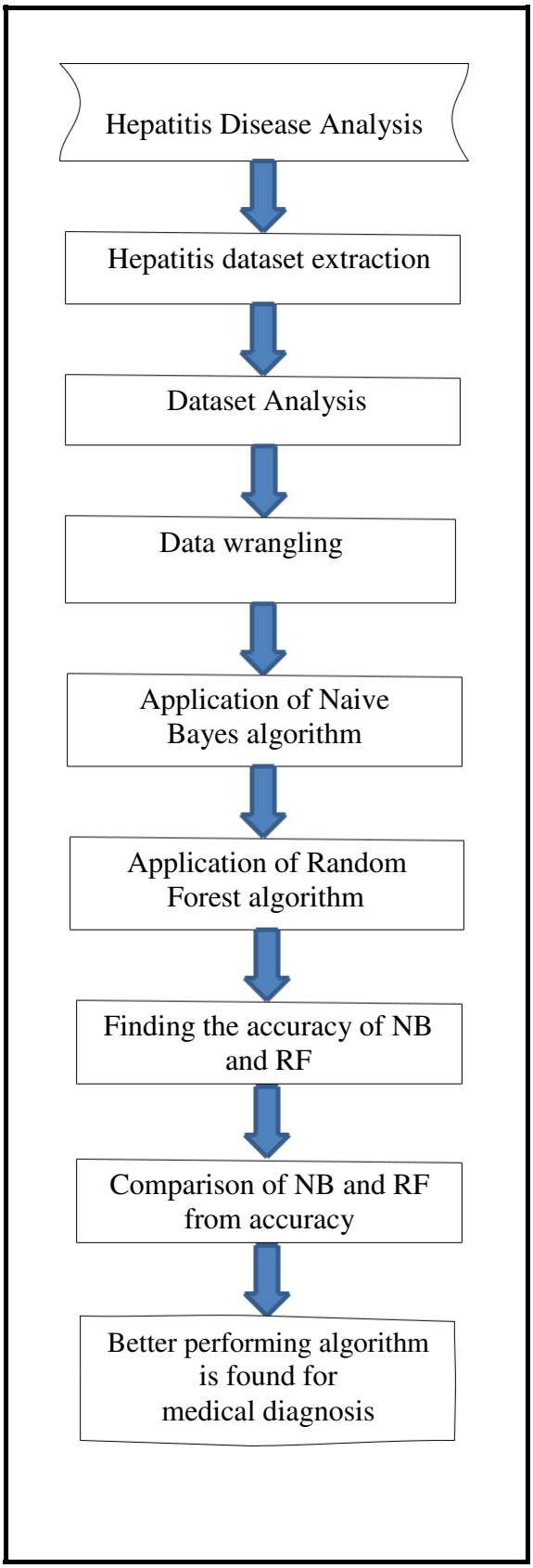
RANDOM FOREST

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

3.ALGORITHM APPLICATION

The Random Forest algorithm is applied to the dataset. It is done using python language and jupyter tool is used as a platform to perform the analysis. Scikit libraries are used to analyze the hepatitis data. Accuracy has to be found for both the algorithms. The accuracy of the algorithm is found. The accuracy of an algorithm is a way of measure to know how often the algorithm classifies the point of data correctly. It is the number of exactly predicted points of data out of all data points. It is found using scikit libraries.

3.FLOW CHART



4.CONCLUSION

Hepatitis disease is a highly contagious liver infection which affects many people in the world. The better the disease is found earlier the better could be the treatment for diagnosis. Machine learning algorithms can predict the disease faster than manual prediction. According to the analysis Random Forest algorithm analyses the data faster comparing to Naïve Bayes algorithm owing to its accuracy. Random Forest algorithm can be used for better diagnosis.

REFERENCES

[1] Li Sijia, Tan Lan, Zhuang Yu, Yu Xiuliang 2010 - Comparison of the prediction effect between the Logistic Regressive model and SVM model

[2] P. R. Visali Lakshmi, G. Shwetha, N. Sri Madhava Raja 2017- 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] VijiyaKumar.K, Lavanya.B, Nirmala.I, Sofia Caroline.S et al. 2019 -Random Forest Algorithm for the Prediction of Diabetes

[5] https://www.who.int/news-room/fact-sheets/detail/hepatitis-b

[6] https://www.niddk.nih.gov/health-information/liver-disease/viral-hepatitis/what-is-viral-hepatitis

[7] https://www.hepb.org/what-is-hepatitis-b/what-is-hepb/

[8]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2809016/