HEPATITIS DISEASE ANALYSIS USING MACHINE LEARNING TECHNIQUES

[1]Nagamani T, [2]Rithuna K

Department of Computer Science and Engineering

Bannari Amman Institute of Technology, Tamilnadu, India-638401

ABSTRACT

Viral hepatitis is one of the most important infectious diseases in the world. It causes an estimation of 1.5 million deaths worldwide each year. Viral hepatitis is an inflammation which tends to damage the hepatocytes in the liver caused by at least six different viruses. So far, many studies have been performed in the diagnosis of hepatitis disease. 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. The main objective is to analyse the disease using two machine learning algorithms. The dataset has already been analysed using Logistic Regression, Support Vector Machine and Naïve Bayes algorithms. It is found that Naïve Bayes performs better than the other two algorithms. In this project Random Forest algorithm is chosen for comparison with Naïve Bayes algorithm to choose an efficient one for diagnosis. Both the algorithms are used for predictive analytics of the hepatitis disease and finding the efficient out of them for better diagnosis.

Keys: compiler - jupyter notepad,

browser – google, python

packages **1. INTRODUCTION**

Hepatitis B (HB) is an infectious disease caused by the hepatitis B virus (HBV).It affects the liver causing acute and chronic infections as well. People do not find the symptom 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. It may take 30 to 180 days for symptoms to begin. In those who get infected around the time of birth 90% develop chronic hepatitis B while less than 10% of those infected after the age of five do.

Most of those with chronic disease have no symptoms; however, cirrhosis and liver cancer may eventually develop. Cirrhosis or liver cancer occur in about 25% of those with chronic disease. It is not possible, on clinical grounds, to differentiate hepatitis B from hepatitis caused by other viral agents, hence, laboratory confirmation of the diagnosis is essential.

A number of blood tests are available to diagnose and monitor people with hepatitis B. They can be used to distinguish acute and chronic infections. Laboratory diagnosis of hepatitis B infection focuses on the detection of the hepatitis B surface antigen HBsAg. WHO recommends that all blood donations be tested for hepatitis B to ensure blood safety and avoid accidental transmission to people who receive blood products. Acute HBV infection is characterized by the presence of HBsAg and immunoglobulin M(IgM) antibody to the core antigen, HBcAg. During the initial phase of infection, patients are also seropositive for hepatitis B e antigen (HBeAg). HBeAg is usually a marker of high levels of replication of the virus.

2. OBJECTIVE

The main objective is to analyse 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. Firstly, the dataset is analysed using Logistic Regression, Support Vector Machine and Naïve Bayes algorithms and comparison is made to find the better performing algorithm. It is found that Naïve Bayes performs better than the other two algorithms. Finally, Random Forest algorithm is chosen for comparison with Naïve Bayes algorithm to choose an efficient one for diagnosis.

3. LITERATURE REVIEW

Li Sijia and Tan Lan proposed the Comparison of the prediction effect between the Logistic Regressive model and SVM model. The paper explains about the financial crisis forewarning. It states that the Financial crises forewarning has important practical significance both for the investors and for the lenders. This paper uses the financial forewarning models, including the Logistic Regressive model and SVM model, to verify the feasibility of the short-term forecast for the financial situation of enterprises. And the paper also gives comparisons between these two models. The results of the study suggest that these two models are both feasible, and the SVM model can achieve better forecasting effects than the Logistic Regressive model. The paper uses two econometric models, the Logistic Regressive model as therepresentation of the traditional financial early warning models, and SVM model (support vector machine model) as the representation of the emerging model of the financial early warning models, to forecast financial situation for some sample firms, summarizes the features of the two models, and then gives a comparison of their prediction effect. Finally, the paper gives some advices on how to approve the financial early warning models so as to give a better prediction.

Fitriana Harahap made an analysis on Implementation of Naïve Bayes Classification Method for Predicting Purchase. It is explained that to choose the right vehicle according to the needs and funds owned by consumers, requires a careful analysis that takes into account many criteria and factors. The criteria used as a benchmark in choosing a vehicle, among others, price, spare parts, cylinder volume, the power of the vehicle. To process all these criteria required a system that can select and classify criteria chosen by consumer, so that can assist consumer in choosing the most appropriate vehicle, therefore needed a system for decision making in making car purchase. The Naive Bayes algorithm is a simple probabilistic classifier that computes a set of probabilities by summing the frequency and value combinations of the given dataset.

VijiyaKumar.K, Lavanya.B, et al. performed the prediction of diabetes disease using Random Forest Algorithm. They said that diabetes is taken into account together of the deadliest and chronic disease that causes a rise in glucose. Diabetes mellitus or just sickness may be a disease caused due to the rise of blood glucose level. Many difficulties might occur if the diabetes remains untreated and unidentified by the doctor. The tedious identifying methodology ends up in visiting of a patient toadiagnostic center and consulting the doctor for more treatment. Rise in machine learning approaches solves this essential draw back. Their main objective is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. Random Forest algorithms are often used for each classification and regression tasks and also it is a type of ensemble learning method. The accuracy level is greater when compared to other algorithms. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly.

4. PROJECT DESCRIPTION

Nowadays, extracting valuable information from the raw data is essential to take the effective business decision. 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 analyse the data, explore the information and then convert the information into knowledge. At that point, Data Mining (DM), Machine Learning (ML) play a vital role to accurately extract the information from the huge amount of data. Several DM methods exist for prediction these are; "Classification, Clustering, Association rules, Summarizations and Regression".

Data mining is used to finding out previously unknown, even though potentially useful, hidden patterns from the extensive amount of data. DM is effectively analysed a large amount of data, complex data that contain multiple variable and nonlinear relations. DM is mainly used to predict outcomes or behaviours according to the future perspective also explore the relationship and associations that are currently not understood. ML is more effective to explore knowledge, validate the data and their behaviour. When data is available, split it into training and test datasets and trained to explore where it stands in future.

The performance of ML algorithms and DM approaches are analysed on hepatitis dataset. We then compare the performance of ML classifiers in terms of accuracy and elaborate which techniques effectively analyse the data. This can be done using two classification techniques; Naive Bayes (NB) and Random Forest (RF) to predict about the hepatitis diseases as the proposed methodology.

5. PROBLEM DEFINITION

Analysis of the hepatitis dataset has already been performed using various algorithms namely Logistic Regression (LR), Support Vector Machine (SVM) and Naïve Bayes (NB). In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting).

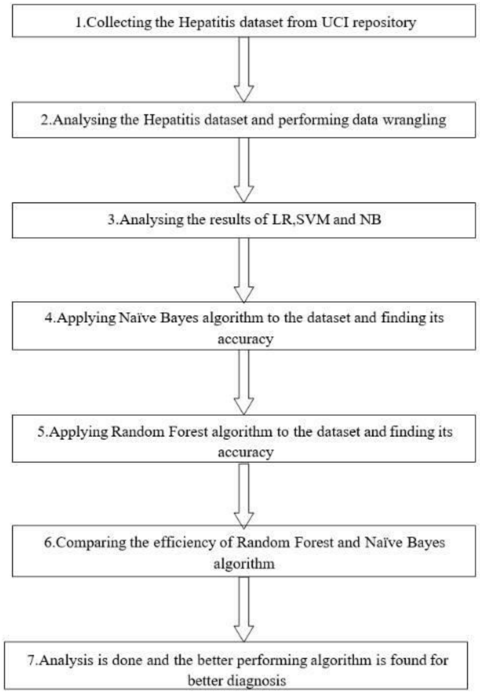
Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. It transforms its output using the logistic sigmoid function to return a probability value. The Naïve Bayes classification technique is based on Bayes' theorem with an assumption of independence between predictors. In simple terms, a Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. The algorithm first creates a frequency table (similar to prior probability) of all classes and then creates a likelihood table.

Then, finally, it calculates the posterior probability. On analysis and comparison of algorithms with hepatitis dataset it is found that Naïve Bayes performs better in analysing the data. This project is intended to find the best algorithm than Naïve Bayes algorithm in analysing the same hepatitis dataset and giving the best accuracy as possible.

6. PROPOSED SYSTEM

The hepatitis dataset is now evaluated to be better performing in Naïve Bayes algorithm. To make the efficiency better Random Forest algorithm is chosen to check if it performs better than Naïve Bayes. 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. The two algorithms namely Naïve Bayes and Random Forest are finally analysed with the same hepatitis dataset and a comparison is made between them based on its accuracy and predicting the better performing algorithm for diagnosis by analysing the disease.

7. FLOW DIAGRAM


1.Collecting the Hepatitis dataset from UCI repository
2.Analysing the Hepatitis dataset and performing data wrangling
3.Analysing the results of LR,SVM and NB
4.Applying Naïve Bayes algorithm to the dataset and finding its accuracy
5.Applying Random Forest algorithm to the dataset and finding its accuracy
6.Comparing the efficiency of Random Forest and Naïve Bayes algorithm
7.Analysis is done and the better performing algorithm is found for better diagnosis

8. 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, Random forest classifier achieved higher accuracy of 87.1% with minimum execution time in classifying and predicting hepatitis infectious disease. Hybrid algorithms with a genetic algorithm can be used in the future for prediction.

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] Mehrbakhsh Nilashi, Othman bin Ibrahim, Hossein Ahmadi, Leila Shahmoradi et al. 2017. 'An analytical method for diseases prediction using machine learning techniques,Computers&Chemical Engineering'.

[6] Amanpreet Singh, Narina Thakur, Aakanksha Sharma 2016 – 'A review of Supervised machine learning algorithms'.

[7]https://medium.com/@synced/how-random-forest-algorithm-works- in machine-learning-3c0fe15b6674.