International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 394-396, 2019, ISSN:-2319–8656

Applications of Machine Learning for Prediction of Liver Disease

Khan Idris
Student (MSc Big Data Analytics)
MIT-WPU, Pune
Maharashtra,India

Sachin Bhoite
Assistant Professor
MIT-WPU, Pune
Maharashtra,India

Abstract: Patients in India for liver disease are continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. It is expected that by 2025 India may become the World Capital for Liver Diseases. The widespread occurrence of liver infection in India is contributed due to deskbound lifestyle, increased alcohol consumption and smoking. There are about 100 types of liver infections. Therefore, building a model that will help doctors to predict whether a patient is likely to have liver diseases, at an early stage will be a great advantage. Diagnosis of liver disease at a preliminary stage is important for better treatment. We also compare different algorithms for the better accuracy.

Keywords: Indian Liver Patients, Machine Learning, Logistic regression, Support Vector Machine, Random Forest, AdaBoost, Bagging.

1. INTRODUCTION:

As there is growth in Liver Patients in India and it is estimated that till the year 2025 India may be the World Capital for Liver Diseases. We should have solution for this kind of problems and for this it is very important for doctors to identify the liver disease at an early stage. To identify liver disease, at an early stage we are building a machine learning model which will predict whether patient should be diagnosed or not at an early stage. We will be using different algorithms as well as ensemble methods. As, Liver disease can be diagnosed by analyzing the levels of enzymes in the blood. The objective of this model is to increase the survival rate of the Liver Patients by using the previous data about the levels of enzymes in their body. We have record of 583 patients from which 416 were the records of liver patient and 167 records of non liver patient.

2. RELATED WORK:

Ramana made a critical study on liver diseases diagnosis by evaluating some selected classification algorithms such as naïve Bayes classifier, C4.5, back propagation neural network, K-NN and support vector. The authors obtained 51.59% accuracy on Naïve Bayes classifier, 55.94% on C4.5 algorithm, 66.66% on back propagation neural network, 62.6% on KNN and 62.6% accuracy on support vector machine. The poor performance in the training and testing of the liver disorder dataset as resulted from an insufficient in the dataset [1]. We, have also gone through a research paper Diagnosis of Liver Disease Using Machine Learning Techniques by Joel Jacob1, Joseph Chakkalakal Mathew2, Johns Mathew3, Elizabeth Issac4. They have used FOUR classification algorithms Logistic Regression, Support Vector Machines (SVM), K Nearest Neighbor (KNN) and artificial neural networks (ANN) have been considered for comparing their performance based on the liver patient data. Authors obtained 73.23% accuracy on Logistic Regression, 72.05% on k-NN, 75.04 accuracy on Support vector machine [2]. We have also gone through a paper Liver Patient Classification using Intelligence Techniques by Jankisharan Pahareeya, Rajan Vohra, Jagdish Makhijani, Sanjay Patsariya. In this paper Authors have used six intelligence techniques on the ILPD (Indian Liver Patient) Data Set. Throughout the study ten-fold cross validation is performed [3]. "Machine Learning Techniques on Liver Disease", in this paper authors have shown different types of techniques for disease prediction. Here algorithms Logistic Regression, SVM, Decision tree, Random Forest and ensemble techniques are used [4]. "Liver Classification Using Modified Rotation Forest", in this paper authors have gone through various classification algorithms to increase the accuracy and have done feature selection. Accuracy in this paper was 73.33% [5].

3. PROCESS IMPLEMENTATION:

The work flow process is firstly, we have to preprocess the data, then some data visualization part then we trained the model with different algorithms and selecting the algorithm with best output

.

3.1 Dataset

The Indian Liver Patient Dataset consists of 10 different attributes such as Age, Sex, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alamine Phosphatase, Total Proteins Albumin, Albumin and Globulin Ratio, Dataset (result) of 583 patients. (416 records are of liver patients and 167 non liver patients). The patients were described as either 1 or 2 on the basis of liver disease. The feature Sex is converted to numeric value (0 and 1) in the data pre-processing step.

3.2 Data preprocessing

Data pre-processing is an essential step of solving every machine learning problem. It is said that 80% of the time of a Data Scientist is spend in data preprocessing. Most commonly used preprocessing techniques are very few like missing value imputation, encoding, scaling, etc. Every dataset is different and poses unique challenges. All features, except Gender are real valued integers. Therefore, in this column males are labeled as ‘1’ and females are labeled as ‘0’.

The last column, Disease, is the label with ’1’ representing presence of disease and ’2’ representing absence of disease. This column is then relabeled as ‘1’ for liver patients and ‘0’ for the non liver patients. Total number of records is 583, with 416 liver patient records and 167 non-liver patient records. In the data visiualization of this dataset, it is observed that some values are Null for the Albumin and Globulin Ratio column. The columns which contain null values are replaced with median values of the column.

3.3 Classification Techniques

a. Logistic Regression: Logistic regression is a type of a supervised machine learning algorithm. It makes a prediction that has binary outcome from the past data. Logistic regression usually returns result in very short time, hence it is preferred being used as a benchmarking model [4].

Logistic regression is one of the simpler classification models. Because of its parametric nature it can to some extent be interpreted by looking at the parameters making it useful when experimenters want to look at relationships between variables. A parametric model can be described entirely by a vector of parameters = (0, 1... p). An example of a parametric model would be a straight-line y = mx + c where the parameters are c and m. With known parameters the entire model can be recreated. Logistic regression is a parametric model where the parameters are coefficients to the predictor variables written as 0 +1 +X1 + ...PXp Where 0 is called the intercept [2]. As, it is seen that accuracy achieved by Logistic Regression was 73.23%, Now we are applying Adaboost to the Logistic Regression.

b. Support Vector machines: Support vector machines so called as SVM is a supervised learning algorithm which can be used for classification and regression problems as support vector classification (SVC) and support vector regression (SVR). It is used for smaller dataset as it takes too long to process.

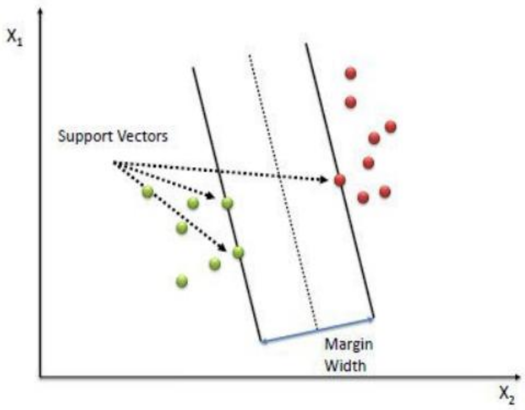


Fig.1 Support Vector Machine

c. Random Forest: Random Forest is a meta estimator that uses the number of decision tree to fit the various sub samples drawn from the original dataset, drawn data with replacement as per the requirements. Decision tree is non-parametric supervised learning. It is used for both classification and regression problem.

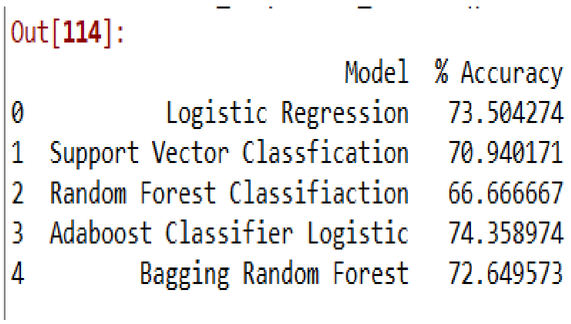
It is flowchart-like structure in which each internal node represents a “test” on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The path between root and leaf represent classification rules. It creates a comprehensive analysis along with each branch and identifies decision nodes that need further analysis.

d.AdaBoost: AdaBoost is one of the first boosting algorithms to be adapted in solving practices. Adaboost helps you combine multiple “weak classifiers” into a single “strong classifier. The weak learners in AdaBoost are decision trees with a single split, called decision stumps. AdaBoost works by putting more weight on difficult to classify instances and less on those already handled well. AdaBoost algorithms can be used for both classification and regression problem[6].

e.Bagging: Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting.

4. RESULTS AND EVALUATION:

The main objective was to predict whether a patient should be diagnosed or not at an early stage with algorithms such as SVM, Logistic Regression and random Forest. These algorithms were also used in previous studies. Now we have improves accuracy of these algorithms by using Bagging and AdaBoost.



```
Out[114]:
                          Model  % Accuracy
0           Logistic Regression   73.504274
1  Support Vector Classfication   70.940171
2  Random Forest Classifiaction   66.666667
3   Adaboost Classifier Logistic   74.358974
4           Bagging Random Forest   72.649573
```

Fig. 2 Accuracies of Algorithms

As, we can see in the above figure increasing accuracies of the algorithms. We got accuracy 73.5% for Logistic Regression, then by applying Adaboost classifier the accuracy has been increased to 74.35%.

For Support Vector Machine we got 70.94%, and for Random Forest Classification 66.67% here we have got a considerable increase in accuracy by using Bagging that is the accuracy of 72.64.

5. CONCLUSION:

We have applied the machine Learning algorithms on the Indian Liver Patient dataset to predict the patients by the enzymes content in their at an early stage. We have used different machine learning classification algorithm as Logistic Regression, SVC, Random Forest and further we have applied bagging to Random Forest and AdaBoost to Logistic Regression. Logistic Regression is fast in processing and gave accuracy of 73.5%. Thus for increasing its accuracy we have used AdaBoost and got accuracy of 74.36%.

6. REFERENCES:

[1] Bendi Venkata Ramana1, Prof. M.Surendra Prasad Babu2 , Prof. N. B. Venkateswarlu3. “A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis” International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011.

[2] Joel Jacob, Joseph Chakkalakal Mathew, Johns Mathew, Elizabeth Issac “Diagnosis of Liver Disease Using Machine Learning Techniques “by International Research Journal of Engineering and Technology (IRJET) 1,2,3 Dept. of Computer Science and Engineering, MACE, Kerala, India 4Assistant Professor, Dept. of Computer Science and Engineering, MACE, Kerala, India Volume: 05 Issue: 04 | Apr-2018.

[3] Jankisharan Pahareeya1, Rajan Vohra2, Jagdish Makhijani3 Sanjay Patsariya4 “Liver Patient Classification using Intelligence Techniques”, International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 2, February 2014 .

[4] V.V. Ramalingam1, A.Pandian2, R. Ragavendran3 “Machine Learning Techniques on Liver Disease - A Survey” 1,2,3Department of Computer Science and Engineering, SRMIST, Kattankulathur. International Journal of Engineering & Technology, 7 (4.19) (2018) 485-495.

[5] Bendi Venkata Ramana1 , Prof. M.Surendra Prasad Babu2 1 Associate Professor, “Liver Classification Using Modified Rotation Forest “Dept.of IT, AITAM, Tekkali, A.P. India. 2 Dept. of CS&SE, Andhra University, Visakhapatnam-530 003, A.P, India.

[6] https://towardsdatascience.com/understanding-adaboost-2f94f22d5bfe