A Prediction Model of Detecting Liver Diseases in Patients using Logistic Regression of Machine Learning

PSM Keerthana, Nimish Phalinkar, Riya Mehere, Koppula Bhanu Prakash Reddy, Nidhi Lal

IIIT,Nagpur,Maharashtra.

Abstract: **Liver Diseases are prevalent in India accounting for 2.4% of Indian deaths per year [1]. According to the WHO, liver disease is one of the most common causes of death in India. Liver diseases have become a menacing threat in India with around 10 lakh patients being newly diagnosed with it every year. Liver disease owing to its subtle symptoms remains obscure and hence leading to an onerous diagnosis, often the symptoms become apparent when it is too late [2]. Therefore, an endeavour is made for the forecast of liver sickness in patients utilizing machine learning techniques. In this paper, we thus used the Machine Learning method of Logistic Regression to predict liver disease in patients**.

Keywords - Liver Diseases, Logistic Regression, Machine Learning, Confusion Matrix, Cross-Validation.

1. Introduction

The diagnosis of liver diseases in the early stages is a perplexing task as the symptoms are unnoticeable. Complications of liver diseases are not identified quickly enough as the liver functions normally even when it is partially damaged [3]. Although not determinable even to an experienced medical practitioner, the premature symptoms can be detected. Early diagnosis forms the crux in increasing the patient's life span substantially. In this day and age Machine Learning techniques are quintessential in preventive medicine to predict disease from the health care database [3]. Although not determinable even to an experienced medical practitioner, the premature symptoms can be detected. Early diagnosis forms the crux in increasing the patient's life span substantially. In this day and age Machine Learning techniques are quintessential in preventive medicine to predict disease from the health care database. It plays a salient part in medical decision making and also specializes in the integration of multiple risk factors into predictive tool [6,7]. As the healthcare data is increased gradually, machine learning give access to analysing massive amounts of data to rapidly [12]. Many industries are leveraging machine learning to improve medical diagnostics. This paper aims to predict liver disease in patients using Machine Learning method of Logistic Regression. Inclusive of Logistic Regression, pre-processing techniques such as Removing Duplicate Values, Null values, dealing with categorical data using Encoding method and Scaling features have been used. This model can be used as a valuable tool for clinical decision making [4].

2. Related Work

A considerably eminent amount of research is being carried out on Liver disease prediction which is of paramount importance in today's scenario.

Amidst the most influential work in Machine learning with reference to this topic can be attributed to Thirunavukkarasu K. [5]. Their work ascribes to different classification algorithms namely Logistic Regression, Support Vector Machine and K-Nearest Neighbour have been used for liver disease prediction. The comparison of all these algorithms been done based on classification accuracy which is found through the confusion matrix as [4]:

- K-nearest neighbour model, calculated accuracy is 73.97%. Sensitivity value is 0.904 and specificity values is 0.317.
- Logistic Regression model, accuracy has been calculated and is 73.97%. Sensitivity value is 0.952 and specificity values is 0.195.
- Support Vector Machine model, accuracy has been calculated and is 71.97%. Sensitivity value is 0.952 and specificity values is 0.195.

From the experiment, it is concluded that Logistic Regression and K-Nearest Neighbour has equal accuracy. Since in medical term, test sensitivity is the ability of the test to correctly identify those with the disease thus logistic regression is the best model for predicting liver disease.

Another appreciable work in this field can be ascribed to Md. Mohaimenul Islam [9]. Their work incorporates various machine learning algorithms to improve prediction of liver diseases that provided significant insights along with traditional statistical models. In addition, the model could provide an uncomplicated, rapid, budget friendly, and non-invasive method to accurately diagnose liver diseases [10].

Their model deals with various machine learning techniques to predict liver disease, and logistic regression model showed better performance of 0.763. The confusion matrix has been used to determine the equation between the actual values and predicted values [9]. The whole dataset was predicted using 10-fold cross-validation. The AUC of random forest (RF), support vector machine (SVM), artificial neural network (ANN), and logistic regression (LR) were 0.708, 0.657, 0.7333, and 0.763 respectively.

In their study, logistic regression model displayed better performance in comparison with other classification techniques. Their prediction outcome has the potential to help clinicians make more precise and meaningful decisions about the liver disease diagnosis and treatment.

3. Proposed work

The given data set comprises of 583 Indian Patient details, it consists 10 variables of which 1 is a dependent variable, and the remaining 9 are independent variables used for predicting whether the person is affected by liver disease or not shown by Fig 1. The dependent variable is "is patient" which is indicated as '1', indicating that the person has liver disease and '2', indicating that the person is not a liver patient which are converted to '1' for yes and '0' for no respectively. The given data set has 26 duplicate patient's details which are deleted and then correspondingly used in the model. After removing the duplicates, the data set row size is reduced from 583 to 570. The data set contains 4 null values for alkphos the patient's data with these Null values are deleted completely, for improving the accuracy of the prediction. The data set is partitioned 80% to train and 20% to test the model with Logistic Regression Algorithm.

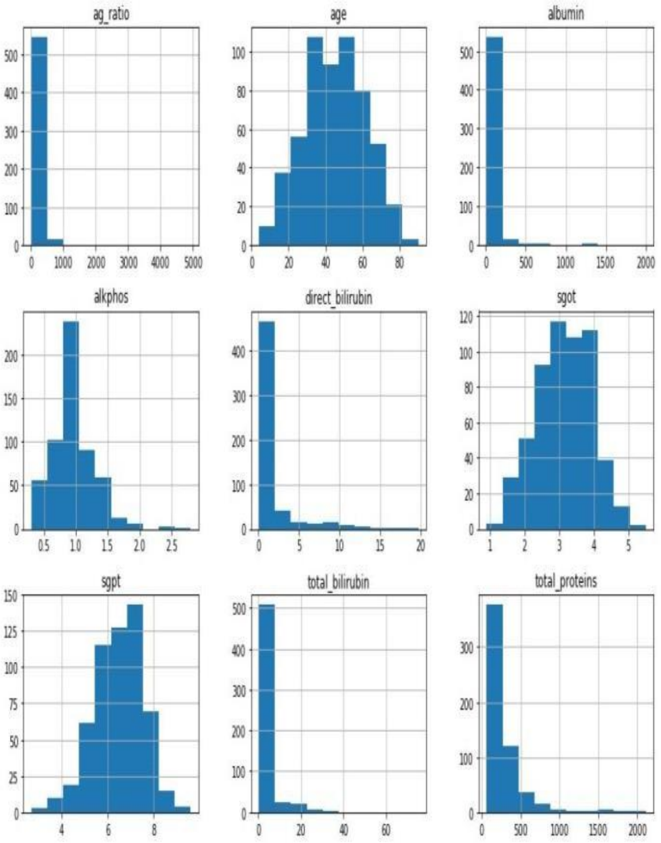


Fig.1- Data visualization of the given set

Logistic Regression comes under the supervised machine learning algorithms. This is used in binary classification. Fig 2 shows the Math behind simple Logistic Regression classifier. Logistic Regression algorithm takes a data set line and then calculates the probability for classifying the sample among the two classes with respect to the threshold value.

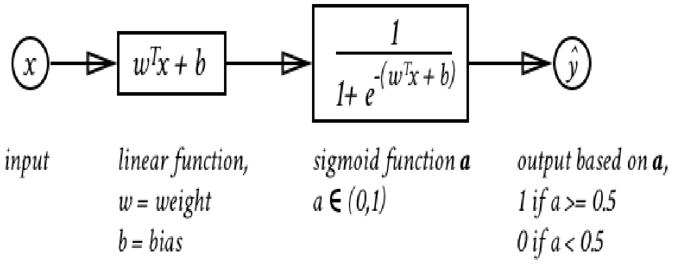


Fig.2- Logistic Regression weights bias

4. *Results and Discussion*

In this paper, Logistic Regression model is implemented for the prediction of liver disease. The dependent variable 'is_patient' as '1'for yes and '0' for no respectively are divided based on gender and plotted against the count of the patients as given by Fig 3.

As per the given data-set, the count of the total number of patients predicted to have liver disease and not are represented as a graph in Fig 4.

From Fig 5, it is observed that the accuracy score of the model had increased until the random state reached the value 3444 and decreased when increased further.

Confusion matrix provides the details of the performance of a classification model on a test data for which the values are known. The True Positives = 97, True Negatives = 1 for the model which is explained using the heat map in Fig 6 where the true labels are represented by Y-axis and predicted labels are represented by X-axis.

For classifier Random Forest Classifier, ROC score is 0.659091 obtained from the reference work as in Fig 7.2 and the ROC curve for True Positive vs False Positive is given by Fig 7.1

The final accuracy score obtained in this model is 0.859649

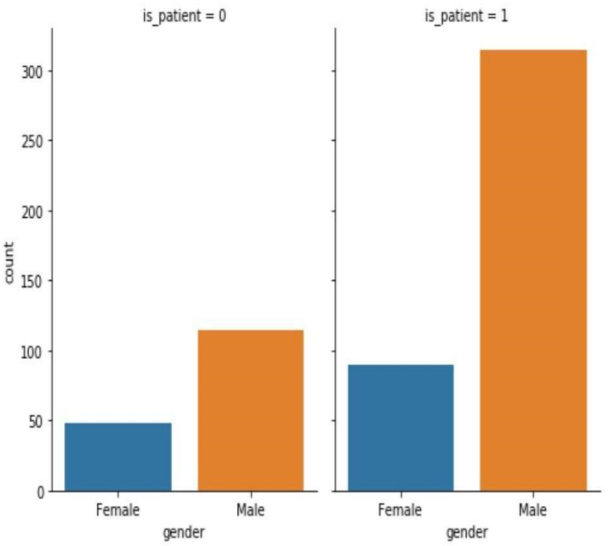


Fig. 3- Frequency vs Independent Factors

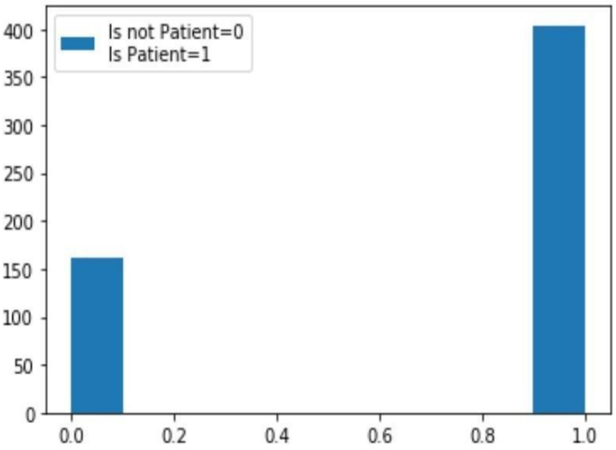


Fig.4- Count of number of patients and not patients

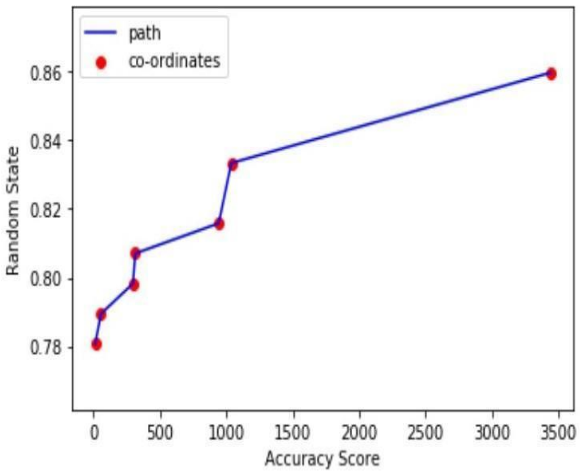


Fig.5- Random state versus the accuracy score

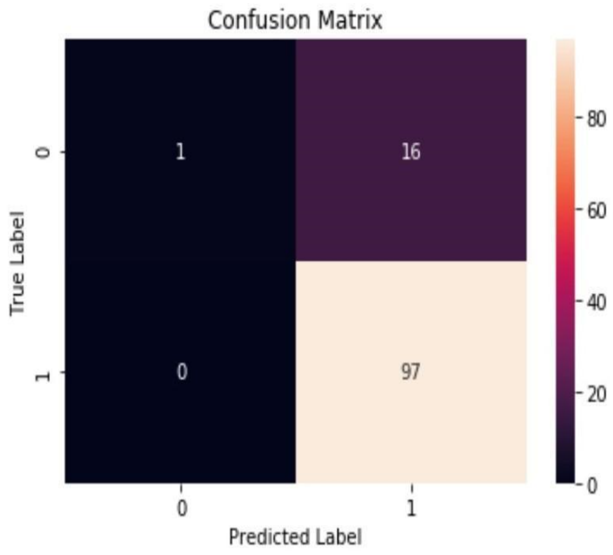
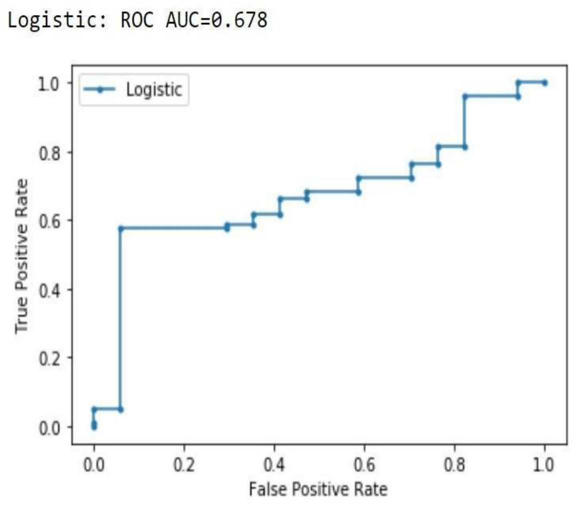


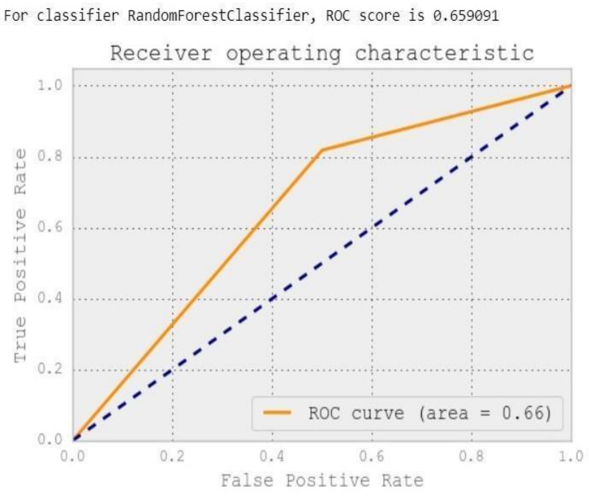
Fig.6- Confusion Matrix

ROC – Curve



```
Logistic: ROC AUC=0.678
```

Fig .7.1- ROC curve for True Positive vs False Positive



```
For classifier RandomForestClassifier, ROC score is 0.659091
```

Fig – 7.2: True Positive vs False Positive ROC Curve

5. *Conclusion*

Prompt and timely detection of liver disease prediction plays a vital role in increasing life span of patient. In this paper, an attempt is made to detect the presence of liver disease using Logistic Regression method of Machine Learning. The accuracy score of the model can be further improved by using decision tree and also by increasing the data set, K-Nearest Neighbor algorithm is also one of the pertinent methods which can be used to predict the disease accurately. It proposes to improve the accuracy further.

6. REFRENCES

[1] www.worldlifeexpectancy.com

[2] Rong-Ho Lin, "An Intelligent Model for Liver Disease Diagnosis, "Artificial Intelligence in Medicine, 2009" Sontakke, S., Lohokare, J., &

Dani, R. (2017). Diagnosis of liver diseases using machinelearning.2017 International Conference on Emerging Trends & Innovation in ICT(ICEI).

[3] Prediction of fatty liver disease using machine learning algorithms. Chieh Chen Wu, Wen Chun Yeh, Wen Ding Hsu, Md Mohaimenul Islam, Phung Anh (Alex) Nguyen, Tahmina Nasrin Poly, Yao Chin Wang, Hsuan Chia Yang, Yu Chuan (Jack) Li, TMU Research Centre of Artificial Intelligence in Medicine, College of Medical Science and Technology, Taipei Municipal Wanfang Hospital TMU Research Centre of Cancer Translational Medicine

[4] Thirunavukkarasu, k., Singh, A. S., Irfan, M., & Chowdhury, A. (2018). Prediction of Liver Disease using Classification Algorithms. 2018 4th International Conference on Computing Communication and Automation (ICCCA).

[5] W. Raghupathi, V. Raghupathi, Big data analytics in healthcare: promise and potential, Health information science and systems (2014), 2-3.

[6] P. Groves, B. Kayyali, D. Knott, S.V. Kuiken, The big data revolution in healthcare: Accelerating value and innovation, 2016.

[7] A. Charleonnan, T. Fufaung, T. Niyomwong, W Chokchueypattanakit, S. Suwannawach, N. Ninchawee “Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques” MITiCON2016.

[8] Md.Mohaimenul Islam a, b, Chieh-Chen Wu a, b, Tahmina Nasrin Poly a, b, HsuanChia Yang b, Yu-Chuan (Jack) Li a, b, c,1 a Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan b International Centre for Health Information Technology (ICHIT), Taipei Medical University, Taipei, Taiwan C Department of Dermatology, Wan Fang Hospital, Taipei, Taiwan

[9] J. Kang, T. Lee, I. Yap, K. Lun, Analysis of cost-effectiveness of different strategies for hepatocellular carcinoma screening in hepatitis B virus carriers. Journal of gastroenterology and hepatology 7(1992),463-468.