Liver Patient Analysis Using Machine Learning Techniques

1. Anusha J, 2.Karthik Raju S R, 3.Megha, 4.Neethushree V, 5.T N Jayanth

Co author. MAHENDRA KUMAR B,

1,2,3,4,5 PG SCHOLORS, DEPT. OF MCA, DSCE
CA -ASST. PROF., DEPT. OF MCA,DSCE

EMAIL ID

1.anushajk06@gmail.com, 2.kcoll984@gmail.com, 3.megha.s3377@gmail.com, 4.neethuvgowda8@gmail.com, 5.jayanth.tn1996@gmail.com.

CA - mahendra-mcavtu@dayanandasagar.edu

Abstract –

Analysis of liver disease at a preliminary level is essential for higher remedy. It's miles a very tough task for medical researchers to expect the disorder within the early degrees as a result of subtle symptoms. Frequently the signs and symptoms end up obvious while it's miles too overdue. To overcome this difficulty, this mission aims to improve liver sickness diagnosis using machine learning methods. The primary objective of this research is to useclassification algorithms to pick out the liver patients from wholesome people. This project additionally pursuits to examine the classification algorithms primarily based on their overall performance elements.

Key Words: Logistic regression, Random Forest, Machine Learning, Liver Patient Analysis

1. INTRODUCTION

Troubles with liver patients are not effortlessly observed in an early degree because it will be functioning generally even if it's far partly damaged. An early recognition of liver issues will boom patient's survival charge. Most of the Indians are victim of liver disease. It's far predicted that via 2025 India may turn out to be the arena capital for liver diseases. The big incidence of liver infection in India is contributed due to deskbound way of life, accelerated alcohol intake and smoking. There are about one hundred kinds of liver infections. Consequently, developing a machine to predict whether the patient has liver disease or not might be helpful for the medical area.

Machine learning techniques are much prominent in medicinal conclusion and anticipating diseases.Paul R Harper announced that there isn't essential a solitary best order instrument yet rather the best performing calculation will rely upon the highlights of the dataset to be broke down.

The main goal of this studies is to apply classification techniques to identify the liver sufferers from healthful people. ThePatients with liver disease are denoted by 1 and healthy people are denoted by 2.In this have a look at, two classification algorithms Logistic Regression and Random Forest had been considered for comparing their overall performance based totally at the liver affected person facts.The dataset used is the IndianLiver Patient Dataset (ILPD) which changed into selected from UCI system studying repository for this look at. It is a pattern of the entire Indian populace amassed from AndhraPradesh location and contains of 584 patient records.

2. RELATED WORKS

In recent research works, several neural network models have been developed to aid in diagnosis of liver diseases in the medical field by the physicians such as diagnosis support system , expert system, intelligent diagnosis system, and hybrid intelligent system. In addition, Christopher N. proposed a system to diagnose medical diseases considering 6 benchmarks which are liver disorder, heart diseases, diabetes, breast cancer, hepatitis and lymph. The authors developed two systems based on Pression and Support, an accuracy of 72.00%%with 19 rules of liver disorder dataset and 74.35% with 43 rules which was obtained from the Pression and Support respectively. Ramana also made a critical study on liver diseases diagnosis by evaluating some selected classification algorithms such as Logistic Regression classifier, Support, back propagation neural network, support vector. The authors obtained 74.35% accuracy on Logistic Regression classifier, 74.35% on Random Forest algorithm.

The poor performance in the training and testing of the liver disorder dataset as resulted from an insufficient in the dataset. Therefore, Sug, suggested a method based on oversampling in minor classes in order to compensate for the insufficiency of data effectively. The author considered two algorithms of decision tree for the research work.

3. IMPLEMENTATION

1. DATASET

The Indian Liver Patient Dataset contains of 11 different attributes of 584 patients. The patients with liver disease are denoted by 1 and the non-liver patient are denoted by 2. The detailed description of the dataset is shown in Table. The table provide details about the attribute and attribute type. As clearly visible from the table, all the features except gender are real valued integers. The feature Gender is converted to numeric value (0 and 1) in the data pre-processingstep.

Table-1 Dataset Description

SLNO.	**ATTRIBUTES**	**ATTRIBUTE TYPE**
1.	Age	Numeric
2.	Gender	Nominal
3.	Total_Bilirubin	Numeric
4.	Direct_Bilirubin	Numeric
5.	Alkaline_Phosphotase	Numeric
6.	Alamine_Aminotransferase	Numeric
7.	Aspartate_Aminotransferase	Numeric
8.	Total_Protiens	Numeric
9.	Albumin	Numeric
10.	Albumin_and_Globulin_Ratio	Numeric
11.	Liver_Patient	Numeric (1,2)

2. DATA-PREPROCESSING

Data pre-processing is an important step of solving every machine learning problem. Most of the datasets used with Machine Learning problems need to be processed / cleaned / transformed so that a Machine Learning algorithm can be trained on it. Most commonly used pre-processing techniques are very few like missing value imputation, encoding categorical variables, scaling, etc. These techniques are easy to understand. But when we actually deal with the data, things often get clunky. Every dataset is different and poses unique challenges. All features, except Gender are real valued integers. The last column, Disease, is the label (with ’1’ representing presence of disease and ’2’ representing absence of disease). Total number of data points is 584, with 416 liver

patient records and 167 non-liver patient records. In the description 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 mean values of the column.

3. CLASSIFICATION TECHNIQUES

a) LOGISTICREGRESSION

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 = kx + m where the parameters are k and m. With known parameters the entire model can be recreated. Logistic regression is a parametric model where the parameters arecoefficientstothepredictorvariableswrittenas0+1

+X1 + ...PXp Where 0 is called the intercept. For convenience we instead write the above sum of the parameterized predictor variables in vector form as X. The name logistic regression is a bit unfortunate since a regression model is usually used to find a continuous response variable, whereas in classification the response variable is discrete. The term can be motivated by the fact that we in logistic regression found the probability of the response variable belonging to a certain class, and this probability iscontinuous.

b) RANDOM FOREST

Random Forests are an ensemble of k untrained Decision Trees (trees with only a root node) with M bootstrap samples (k and M do not have to be the same) trained using a variant of the random subspace method or feature bagging method. Note the method of training random forests is not quite as straightforward as applying bagging to a bunch of individual decision trees and then simply aggregating the output. The procedure for training a random forest is as follows:

At the current node, randomly select p features from available features D. The number of features p is usually much smaller than the total number of features D.

Compute the best split point for tree k using the specified splitting metric (Gini Impurity, Information Gain, etc.) and split the current node into daughter nodes and reduce the number of features D from this node on.

Repeat steps 1 to 2 until either a maximum tree depth l has been reached or the splitting metric reaches some extrema.

Repeat steps 1 to 3 for each tree k in the forest.

Vote or aggregate on the output of each tree in the forest.

Compared with single decision trees, random forests split by selecting multiple feature variables instead of single features variables at each split point. Intuitively, the variable selection properties of decision trees can be drastically improved using this feature bagging procedure. Typically, the number of trees k is large, on the order of hundreds to thousands for large datasets with many features.

4. RESULTS AND EVALUATION

Our most important intention going into this task was to predict liver ailment the usage of diverse machine learning strategies. We expected the use of guide vector machine,**Logistic Regression** and **Random Forest.** All of them expected with better effects. With each set of rules, we've located accuracy, precision, F1-Score, Recall & Support which may be defined as follows:

Accuracy: the accuracy of a classifier is the share of the take a look at set tuples which are efficiently categorized by means of the classifier.

Precision: precision is defined as the proportion of the genuine positives towards all of the effective consequences (each real positives and fake positives)

F1-Score: The **F_1 score** (also **F-score** or **F-measure**) is a measure of a test's accuracy. It considers both the precision*p* and the recall*r* of the test to compute the score: *p* is the number of correct positive results divided by the number of all positive results returned by the classifier, and *r* is the number of correct positive results divided by the number of all relevant samples.

F1-Score: The f1 score (additionally f-rating or f-degree) is a measure of a check's accuracy. it considers both the precision ‘p’ and the Recall ‘r’ of the set take a look at to compute , the rating: p is the range of correct fine consequences divided by means of the quantity of all advantageous consequences lower back by using the classifier, and r is the number of accurate advantageous effects divided by the variety of all applicable samples. //

Recall: recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.Recall is therefore based on an understanding and measure of relevance.

Support: This gives the frequency (no. of times the item occurred) of the item in the dataset

Classification Algorithm	Accuracy	Precision	Recall (Sensitivity)	F1-score	Support
Logistic Regression	74.35	72.00	74.00	72.00	117
Random Forest	74.35	77.00	74.00	75.00	117

5. CONCLUSION

In this undertaking, we have proposed strategies for diagnosing liver ailment in patients utilizing machine learning techniques. The twomachine learning strategies that were utilized incorporate Logistic Regressionand Random Forest. The framework was executed utilizing every one of the models and their presentation was assessed. Execution assessment depended on certain presentation measurements. Both Logistic Regression and Random Forest model gave the same accuracy of 74.36%. Contrasting this work and the past research works, it was found that Random Forest demonstrated profoundly productive.

REFERENCES

[1] Michael J Sorich. An intelligent model for liver disease diagnosis. Artificial Intelligence in Medicine2009;47:53—62.

[2] Paul R. Harper, A review and comparison of classification algorithms for medical decision making.

[3] BUPA Liver Disorder Dataset. UCI repository machine learningdatabases.

[4] Prof Christopher N. New Automatic Diagnosis of Liver Status Using BayesianClassification.

[5] Schiff's Diseases of the Liver, 10th Edition Copyright ©2007 Ramana, Eugene R.; Sorrell, Michael Maddrey, WillisC.

[6] P. Sug, On the optimality of the simple Bayesian classifier under zero-one loss, Machine Learning 29 (2–3) (1997)103–130.

[7]16th Edition HARRISON'S PRINCIPLES of Liver Status Using Logistic Regression, Random Forest.