**Predicting Liver Disease**

By: Eraldi Skendaj

**Background:**

Liver disease is a term that includes many types of diseases that limit the function of the liver. It constitutes a major health problem worldwide. The liver plays an important role in the body. It destroys toxins, metabolizes alcohol and regulates the levels of most chemicals in the blood. In that sense it acts as a filter in the circulation of the blood through the body. The liver’s function can be limited when infected by viruses, or inflamed as a cause of other diseases that end up affecting the liver. Excessive consumption of alcohol, drug abuse, contaminated food sources, pollution and as a result the inhalation of harmful gases have resulted in an increase in the number of liver patients. Chronic liver disease is a common condition throughout the world. Liver diseases are not always easily discovered as the liver often maintains its function even when it may be partially damaged. So early detection is of outmost importance in the treatment of liver diseases. In India 2.44% of all deaths in 2014 were caused by liver disease, according to the World Health Organization. That puts liver disease as the 10th most common cause of death in India. Early detection can improve the survival rate of patients. Eliminating the need for liver biopsy and accelerating the diagnosis of the disease is key in helping liver disease patients.

Enzyme levels in the blood, among other attributes such as age, lifestyle, gender, etc. are often used to detect liver disease. In the days of computers, mobile devices and sensors; machine learning and different classification methods can be used to automate the prediction of disease and thus be crucial in saving lives, especially in remote areas of the world.

**About this project:**

This liver disease prediction model will be done using the ILPD (Indian Liver Patient Dataset) from the UCI Machine Learning Repository. The goal is to develop a machine-learning algorithm that would classify a person as suffering from the disease or not based on the features provided in the dataset. The data contains 416 liver patients records and 167 no liver patients records. The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into groups(liver patient or not). This data set contains 441 male patient records and 142 female patient records.

**Attribute Information:**

1. Age - Age of the patient   
2. Gender - Gender of the patient   
3. TB - Total Bilirubin   
4. DB - Direct Bilirubin   
5. Alkphos - Alkaline Phosphotase   
6. Sgpt - Alamine Aminotransferase   
7. Sgot - Aspartate Aminotransferase   
8. TP - Total Protiens   
9. ALB - Albumin   
10. A/G - Ratio Albumin and Globulin Ratio   
11. Selector - Field used to split the data into two sets (labeled by the experts)

**SGOT**- serum glutamic oxaloacetictransaminase (SGOT) is part of an initial screening for liver disease. This test is done to check for liver damage and Help identify liver disease, especially hepatitis and cirrhosis.

**SGPT** - serum glutamic pyruvic transaminase is measured to see if the liver is damaged or diseased. Alamine Aminotransferase is mainly found in the liver.

**Alkphos -**An alkaline phosphatase (ALP) test. ALP is made mostly in the liver.

**TB and DB** - A bilirubin test is used to detect an increased level in the blood. It may be used to help determine the cause of [jaundice](https://labtestsonline.org/understanding/conditions/jaundice/) and/or help diagnose conditions such as [liver disease](https://labtestsonline.org/understanding/conditions/liver-disease/), [hemolytic anemia](https://labtestsonline.org/understanding/conditions/anemia/?start=6), and blockage of the bile ducts.

**TP** - The total protein test measures the amount of the proteins albumin and globulin in your body.

**ALB** - Albumin is a [protein](https://labtestsonline.org/glossary/protein/) made by the liver that keeps fluid from leaking out of blood vessels, nourishes tissues, and transports [hormones](https://labtestsonline.org/glossary/hormone/), vitamins, drugs, and substances like calcium throughout the body.

**A/G** - The albumin to globulin (A/G) ratio has been used as an index of disease state, however, it is not a specific marker for disease because it does not indicate which specific proteins are altered.

**Methodology:**

In order to complete this project the Python programing language was used; along with its various visualization, statistical, and machine learning libraries.

Pre-Processing techniques were used to impute missing data. A pre-analysis of the data was completed using various visualization techniques. This gave insights on possible correlations between the different features of the dataset. The categorical data that was of character type was encoded into an integer type thus preparing the data for machine learning algorithms.

Looking at the dataset it can be seen that it contains 416 liver patients records and 167 no liver patients records. This shows that the data is skewed an unbalanced. Unbalanced data results in models behaving poorly. As a result the minority class was up-sampled to balance the data.

Machine learning algorithms require that the data be split into dependent and independent variables. In this case the dataset was split so the dependent variable is the column that predicts whether a person has the disease; and the independent variables are the rest of the features in the dataset.

A quick look at the data shows that the data is not scaled. As a result, depending on the algorithm, certain features may overtake the others and introduce bias and result in poor modeling. So feature scaling was employed and the data was standardized.

Looking at the correlations between the features it is clear that the dimensions of this dataset can be reduced to result in better predictions. Primary Component Analysis (PCA) was used for dimensional reduction.

The dataset is then split into a train and test set that the various machine learning algorithms can use to train on and then test their predictions. The dataset is at this point ready to be used to model and predict diseased state using various machine-learning techniques.

**Metrics:**

Along with classification reports and confusion matrices, cross validation accuracies and Area Under the Curve of the Receiver Operating Characteristic (ROC) were used to validate the results of the different machine-learning algorithms. A higher area under the curve for the ROC metric indicates a better true positive rate for the same false positive rate. All of these metrics were used to determine the algorithm that gave the best predictions.

**Machine-Learning algorithms:**

The following machine learning algorithms were used on thus dataset.

1. Logistic Regression

2. Decision Tree

3. Random Forest

4. Support Vector Machine (SVM)

5. K-Nearest Neighbors (K-NN)

6. Naive Bayes

7. Neural Networks

**1. Logistic Regression** – The logistic regression algorithm uses the logistic (sigmoid function) to classify binary data. This model was used as a baseline in determining whether more machine-learning techniques needed to be used to improve the accuracy of the predictions. Looking at the metrics results **prediction accuracy is between 69-72%.**

|  |
| --- |
|  |

**2**. **Decision Tree** - is a hierarchical tree structure that is used to classify data based on a series of rules about the attributes of the class. This model improves on the logistic regression and results **prediction accuracy is between 81-83%.**

**3.** **Random Forest** – makes use of many decision trees on a dataset to make predictions on a model. This algorithm results in a **prediction accuracy between 83-89%.**

**4. Support Vector Machine** (SVM) - A hyperplane separates and classifies the data into different categories. This algorithm results in a **prediction accuracy between 69-73%.** An attempt to improve the accuracy of this model was made. Using the grid search technique, the optimal parameters were found. The accuracy scores did not improve.

**5. K-Nearest Neighbors** (K-NN) – An object is classified based on the majority of its nearest neighbors. This algorithm results in a **prediction accuracy between 80 to 83%** using a neighbor of 1. Applying the elbow method to determine the optimal number of neighbors did not result in improved accuracy.

**6. Naive Bayes** – Bayesian classifiers work on the assumption that each attribute contributes to the total outcome independent of other attributes. This algorithm results in a **prediction accuracy between 64 to 67%**

**7. Neural Networks –** This algorithm mimics the neuron-synapse relationship found in biological organisms. It has an input layer, many hidden layers, and an output layer. Resulting in a network of layer; which gives its name. This algorithm results in a **prediction accuracy between 80 to 81%**

**Discussion:**

As it can be seen from the results of the various algorithms, Decision Trees, Random Forest, K-Nearest Neighbors and Neural Networks give better results then the other algorithms. Further insight can be gained if these same algorithms were applied to a dataset where the majority class was under-sampled during the pre-processing stage of this machine-learning project. The accuracies can possibly be improved if more hidden layers were added to the neural network algorithms, or if other parameters, such as epochs were further optimized. The same parameter discussion can had about the other models that performed the best. However care should be taken no to over-fit this data. This would result in poor predictions when new data is introduced into the models. The accuracy ranges obtained (80-89%) through these algorithms are quite good and would give a doctor and their patients a quick diagnosis, resulting in earlier and better treatment and possibly reducing crowds at the specialists office.

**Conclusion:**

As it has been shown through this project, it is quite possible to help diagnose liver disease by quickly predicting disease state through machine learning algorithms. Early diagnosis through these methods is not only possible but also fairly accurate and this accuracy can be further improved by optimizing the machine-learning algorithms.