Final Project Topic: Liver Disease Prediction (YiHui Ho)

1. Introduction

With a growing trend of sedentary and lack of physical activities, diseases related to liver have become a common encounter nowadays. In rural areas the intensity is still manageable, but in urban areas, and especially metropolitan areas the liver disease is a very common sighting nowadays. Liver diseases cause millions of deaths every year. Viral hepatitis alone causes 1.34 million deaths every year. Sometimes, liver disease is challenging to diagnose in its early stages. We cannot discover the disease until the liver function is partially damaged. Early diagnosis can be life-saving. Currently, the examination of predicting liver illness has been broadly contemplated. Several studies on artificial immune and genetic algorithms have been reported for liver disease diagnosis. Different classifiers with different data sets indicate differences in diagnoses.

Globally, liver disease has become an alarming and life-threatening issue. Machine learning algorithms can early help in early diagnosis to reduce risk. Analyzing the previous studies showed low performance. Hence this research aims to achieve more satisfactory performance.

This research aims to determine the accuracy of several popular machine algorithms—XGBoost, logistic regression, random forest, Decision Tree, K-NN and MLPclassifier—to predict liver diseases by analyzing different data sets and comparing their performances. The main contributions are to find out (a) the correlation matrix with the outcome, (b) the model performance of the lowest split to the higher split of the training set, and (c) the best split of the training set.

Objectives of Research

In India, delayed diagnosis of diseases is a fundamental problem due to a shortage of medical professionals. A typical scenario, prevalent mostly in rural and somewhat in urban areas is:

1. A patient goes to see a doctor with certain symptoms.
2. The doctor recommending certain tests like blood test, urine test etc. depending on the symptoms.
3. The patient taking the tests in an analysis lab.
4. The patient taking the reports back to the reports back to the hospital, where they are examined, and the disease is identified.

The aim of this project is to reduce the time delay caused due to the unnecessary back and forth shuttling between the hospital and the pathology lab. Various machine learning algorithms will be trained to predict a liver disease in patients.

Problem Statement

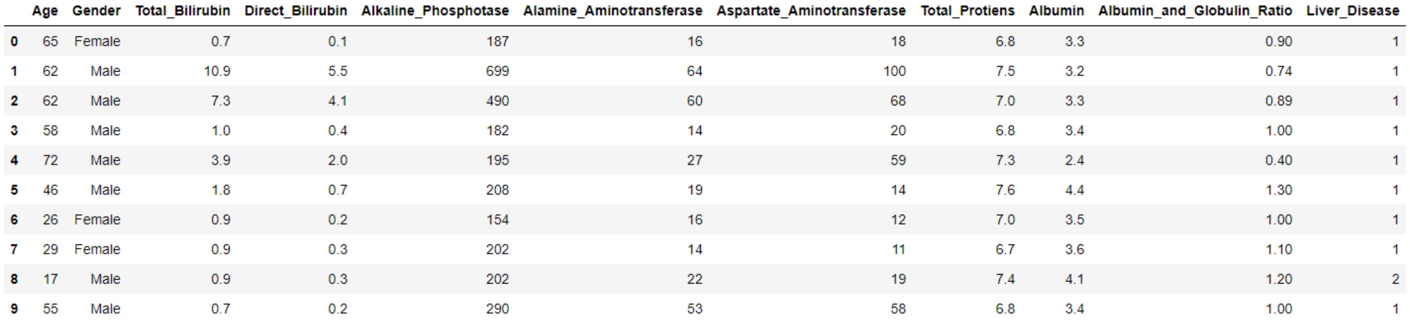
The problem statement is formally defined as:

Given a dataset containing various attributes of 583 Indian patients, use the features available in the dataset and define a classification algorithm which can identify whether a person is suffering from liver disease or not.

2. Materials

2.1 Data Description

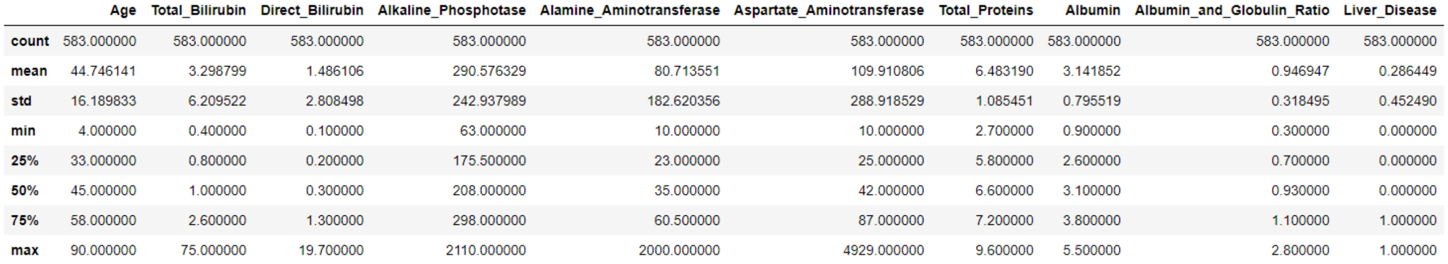
The data are collected from UCI machine Learning Repository and it predicts liver disease based on the given attributes. The data set has eleven attributes which predict the liver disease. The data set is built on both numerical and categorical data types. In my study the dataset contains the attribute such as total bilirubin, direct bilirubin, age, gender, total proteins, albumin, albumin and globulin ratio which is the symptoms of liver disease. The collected data set included 583 liver and non-liver instances with 10 attributes and one outcome. Liver_Disease column value "1" means having liver disease and "2" means without liver disease. A portion of liver dataset is shown below:



	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Liver_Disease
0	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.90	1
1	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
2	62	Male	7.3	4.1	490	60	68	7.0	3.3	0.89	1
3	58	Male	1.0	0.4	182	14	20	6.8	3.4	1.00	1
4	72	Male	3.9	2.0	195	27	59	7.3	2.4	0.40	1
5	46	Male	1.8	0.7	208	19	14	7.6	4.4	1.30	1
6	26	Female	0.9	0.2	154	16	12	7.0	3.5	1.00	1
7	29	Female	0.9	0.3	202	14	11	6.7	3.6	1.10	1
8	17	Male	0.9	0.3	202	22	19	7.4	4.1	1.20	2
9	55	Male	0.7	0.2	290	53	58	6.8	3.4	1.00	1

Tab. 1 presents the features or attributes, with mean and standard deviations.

Tab.1



	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Proteins	Albumin	Albumin_and_Globulin_Ratio	Liver_Disease
count	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000
mean	44.746141	3.298799	1.486106	290.576329	80.713551	109.910806	6.483190	3.141852	0.946947	0.286449
std	16.189833	6.209522	2.808498	242.937989	182.620356	288.918529	1.085451	0.795519	0.318495	0.452490
min	4.000000	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000	0.900000	0.300000	0.000000
25%	33.000000	0.800000	0.200000	175.500000	23.000000	25.000000	5.800000	2.600000	0.700000	0.000000
50%	45.000000	1.000000	0.300000	208.000000	35.000000	42.000000	6.600000	3.100000	0.930000	0.000000
75%	58.000000	2.600000	1.300000	298.000000	60.500000	87.000000	7.200000	3.800000	1.100000	1.000000
max	90.000000	75.000000	19.700000	2110.000000	2000.000000	4929.000000	9.600000	5.500000	2.800000	1.000000

From **Fig. 1** and **Fig .2**, we have information the dataset used in the study consist of 167 negative tested for liver disease and 416 are positively tested. Approximately 71.36% of the data set is affected by liver disease. The data set contain 441 male and 142 female patients.

Fig .1

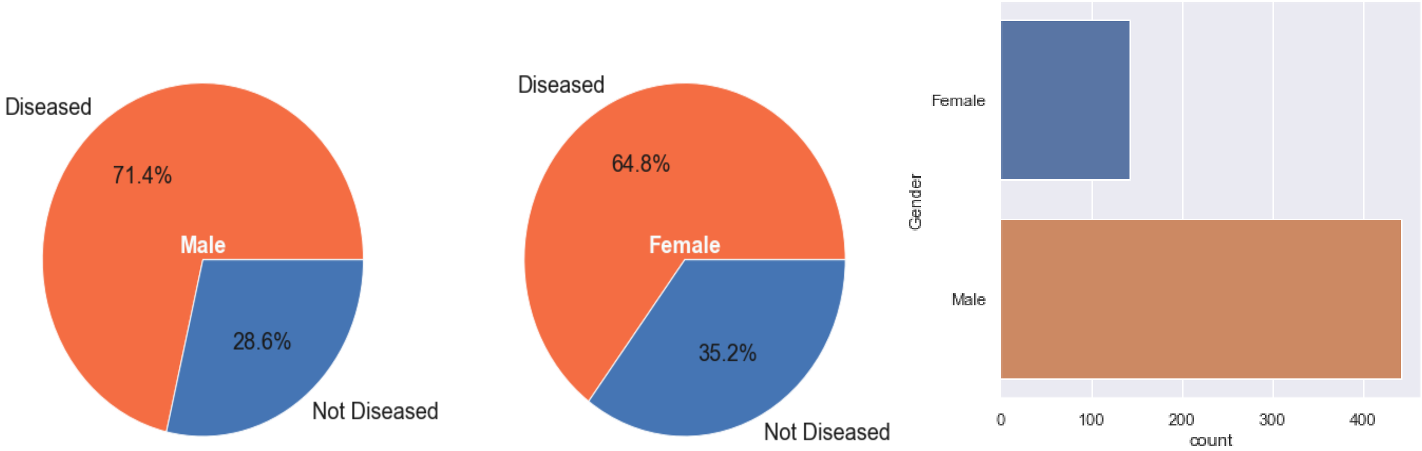
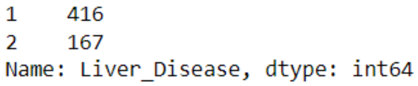
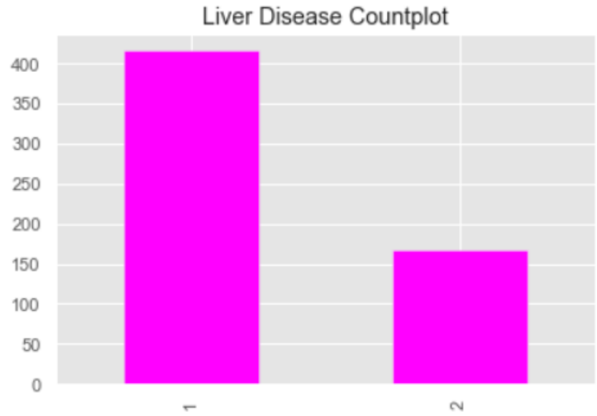


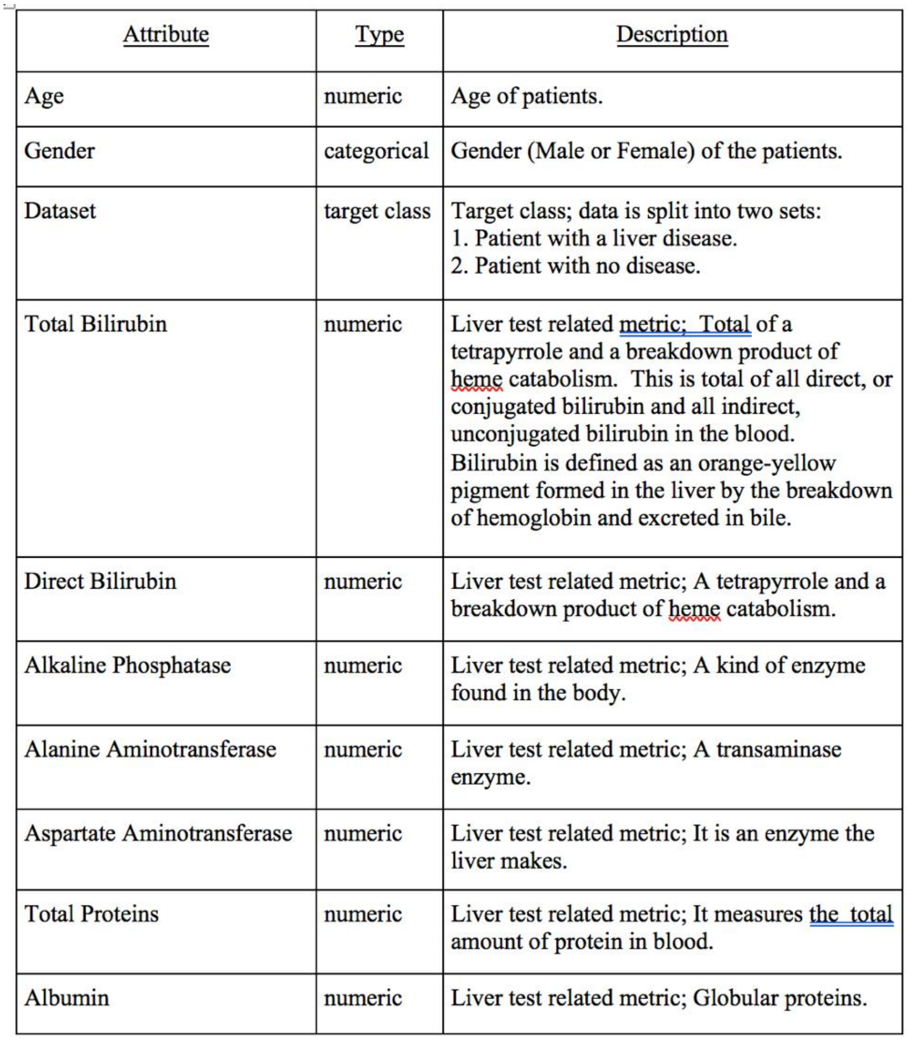
Fig .2



```
1    416
2    167
Name: Liver_Disease, dtype: int64
```



2.2 Definition of Variables



Attribute	Type	Description
Age	numeric	Age of patients.
Gender	categorical	Gender (Male or Female) of the patients.
Dataset	target class	Target class; data is split into two sets: 1. Patient with a liver disease. 2. Patient with no disease.
Total Bilirubin	numeric	Liver test related metric; Total of a tetrapyrrole and a breakdown product of heme catabolism. This is total of all direct, or conjugated bilirubin and all indirect, unconjugated bilirubin in the blood. Bilirubin is defined as an orange-yellow pigment formed in the liver by the breakdown of hemoglobin and excreted in bile.
Direct Bilirubin	numeric	Liver test related metric; A tetrapyrrole and a breakdown product of heme catabolism.
Alkaline Phosphatase	numeric	Liver test related metric; A kind of enzyme found in the body.
Alanine Aminotransferase	numeric	Liver test related metric; A transaminase enzyme.
Aspartate Aminotransferase	numeric	Liver test related metric; It is an enzyme the liver makes.
Total Proteins	numeric	Liver test related metric; It measures the total amount of protein in blood.
Albumin	numeric	Liver test related metric; Globular proteins.

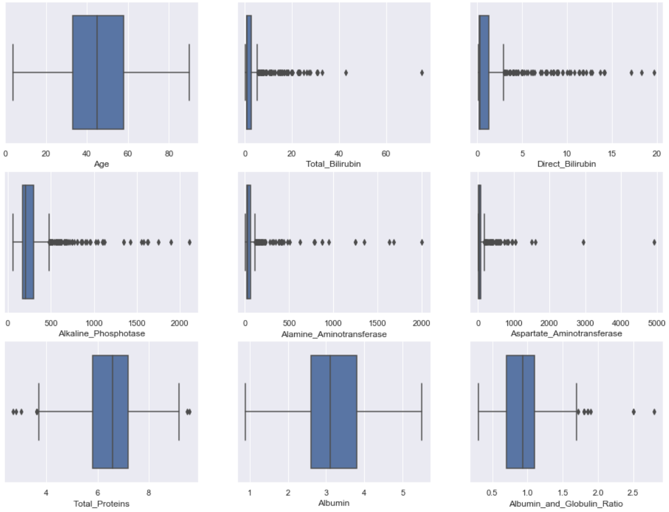
2.3 Feature Engineering

The dataset was manipulated and cleaned using an assortment of libraries that included Pandas and Matplotlib. There was quite some feature engineering to do that included renaming certain columns, dealing null values, handling outliers, and log-transforming and min-max scaling the continuous features. There was also some binning that had to be done regarding the age column to make the results easier to interpret and visualize. One-hot encoding was performed on any discrete features.

Cleaning the data:

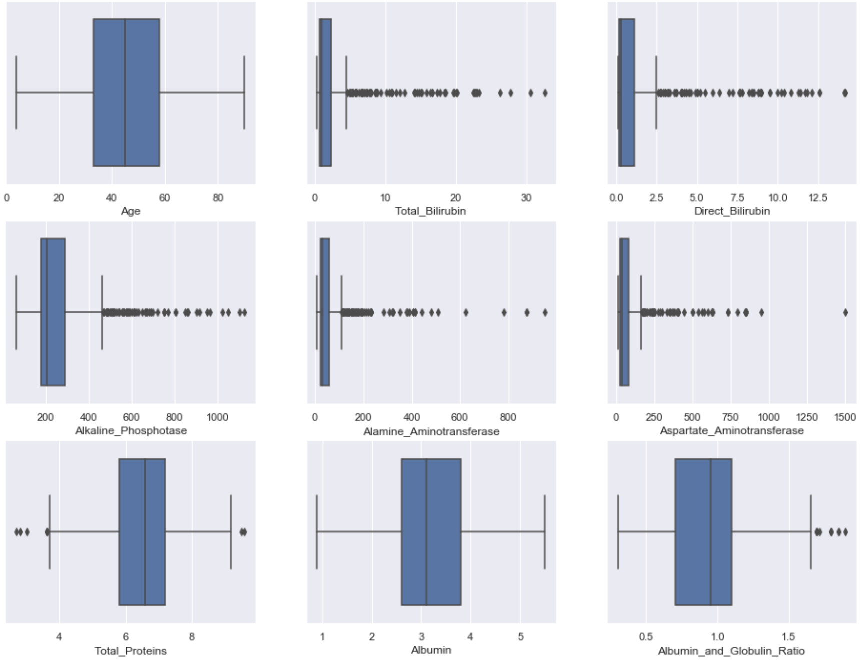
1. Dealing with null values: Only “Albumin_and_Globulin_Ratio” column contains null values, so I replaced na values from albumin and globulin ratio column with median.

2. Dealing with outliers: from the figure below, we can see each feature contain outlier.

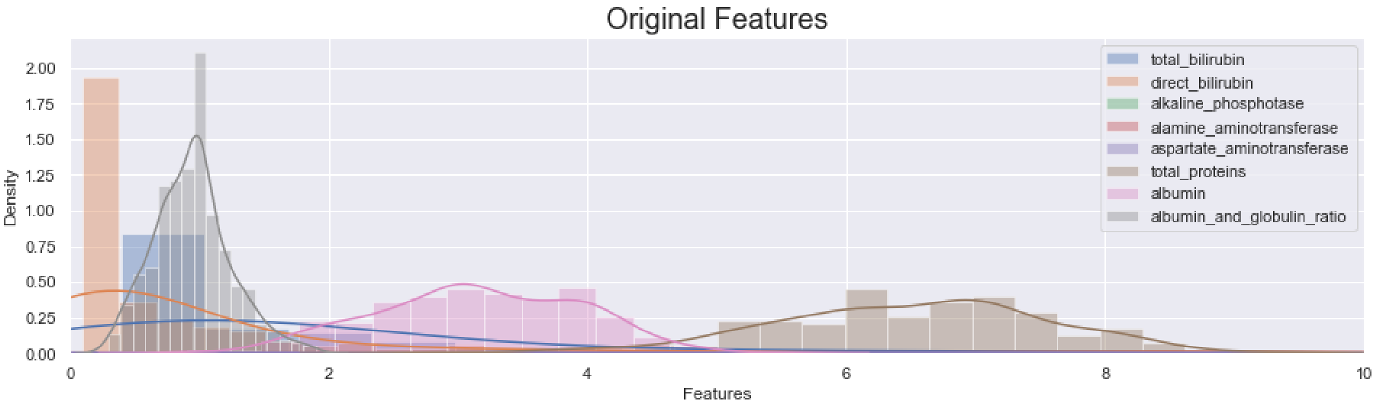


Removed outliers from certain features:
Total_Bilirubin > 40,
Direct_Bilirubin > 15.0,
Alkaline_Phosphotase > 1250,
Alamine_Aminotransferase > 1000,
Aspartate_Aminotransferase > 2000,
Albumin_and_Globulin_Ratio > 2.0.

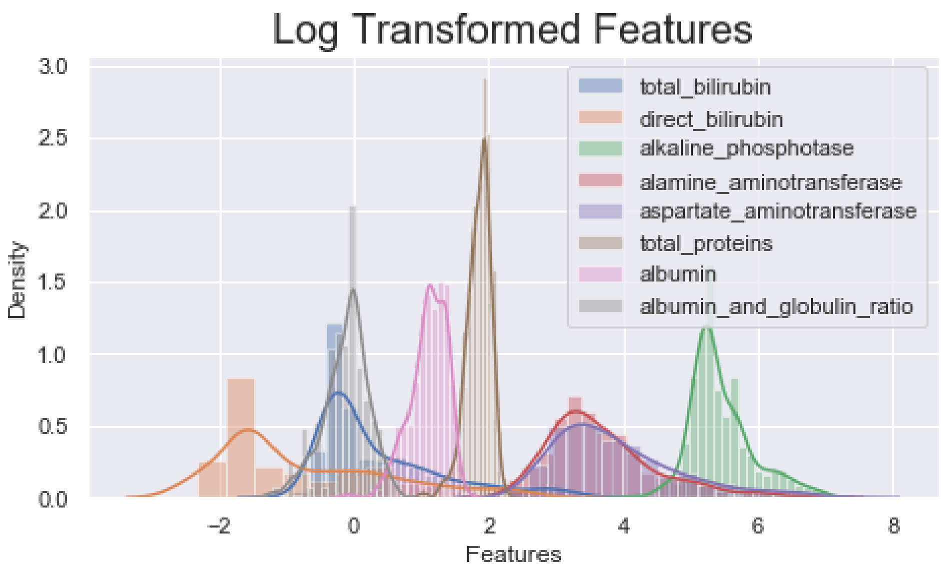
The figure below shows boxplots after removal of outliers.



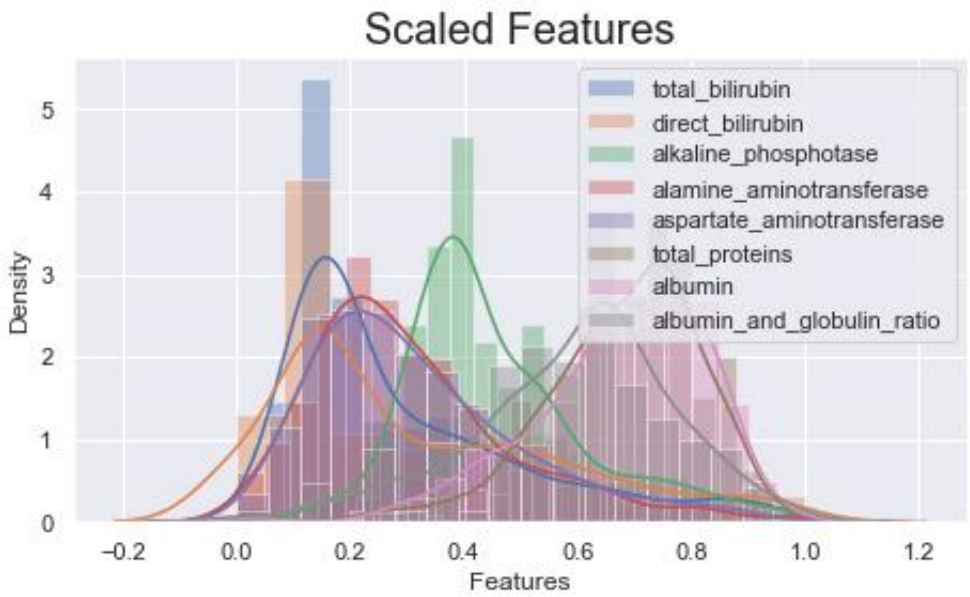
3. Feature Scaling:



Creating log transformation of all features:

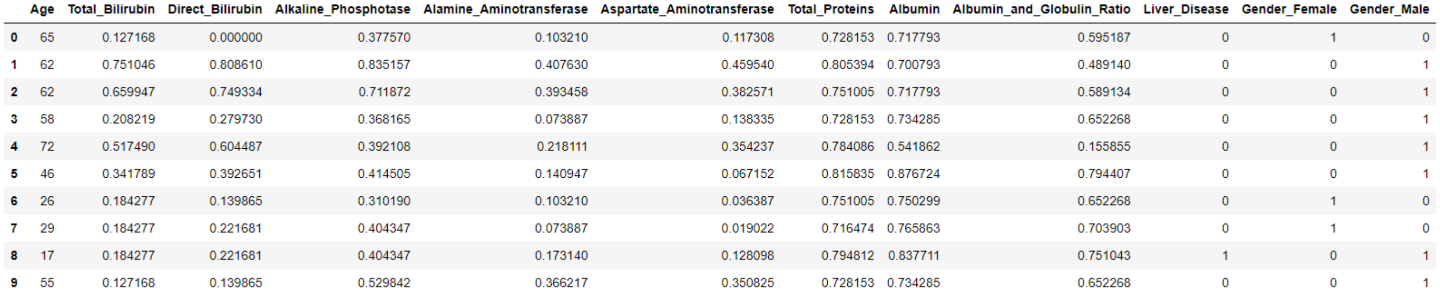


min-max scaling all the features:



4. One-hot encoding: Since “gender” columns now are categorical, created dummies for dataset.

After cleaning the data, feature scaling and one-hot encoding, the final data is shown below:



	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Proteins	Albumin	Albumin_and_Globulin_Ratio	Liver_Disease	Gender_Female	Gender_Male
0	65	0.127168	0.000000	0.377570	0.103210	0.117308	0.728153	0.717793	0.595187	0	1	0
1	62	0.751046	0.808610	0.835157	0.407630	0.459540	0.805394	0.700793	0.489140	0	0	1
2	62	0.659947	0.749334	0.711872	0.393458	0.382571	0.751005	0.717793	0.589134	0	0	1
3	58	0.208219	0.279730	0.368165	0.073887	0.138335	0.728153	0.734285	0.652268	0	0	1
4	72	0.517490	0.604487	0.392108	0.218111	0.354237	0.784086	0.541862	0.155855	0	0	1
5	46	0.341789	0.392651	0.414505	0.140947	0.067152	0.815835	0.876724	0.794407	0	0	1
6	26	0.184277	0.139865	0.310190	0.103210	0.036387	0.751005	0.750299	0.652268	0	1	0
7	29	0.184277	0.221681	0.404347	0.073887	0.019022	0.716474	0.765863	0.703903	0	1	0
8	17	0.184277	0.221681	0.404347	0.173140	0.128098	0.794812	0.837711	0.751043	1	0	1
9	55	0.127168	0.139865	0.529842	0.366217	0.350825	0.728153	0.734285	0.652268	0	0	1

3. Methodology

Main objective of this project is to identify that whether the patient has liver disease or not. This study applied logistic regression, Gaussian Naïve Bayes Classifier, K-NN, SVM, decision tree, random forest, XGBoost, AdaBoost and MLPclassifier algorithms. For each algorithm, I tried out different values of a few hyperparameters to arrive at the best possible classifier. This will be carried out with the help of grid search cross validation techniques. The algorithms are briefly described as follows.

3.1 Logistic Regression (LR)

Logistic regression (LR) is generally a linear model that is used for predicting binary variables. LR technique is used for classifying a new observation of an unknown group. LR is a unique system for gathering data into two random and exhaustive data collections.

3.2 K-Nearest Neighbors (K-NN)

K-NN is an elementary classification algorithm of machine learning. It gives an example of the most preferred class among the neighboring K. K is a constraint for changing the classification algorithms. In this study, I used 44 K neighbors and uniformly used weights. Uniform is a function where all kinds of neighbored points are appropriately weighted. I also used leaf size 30. The leaf size affects the speed of the implementation and the required memory.

3.3 Decision Tree (DT)

A decision tree (DT) is a data excavation model to tackle alliancing and categorizing issues. The DT is generally a prime recursive block to cohere a continuous reckoning implementation. DT includes nodes (parent) and leaves (child). This study used three max-leaf nodes and three criteria for DTs, for which the random state was 0. The random state controls the estimator's randomness, and 0 means that the randomness was false in this case.

3.4 Random Forest (RF)

Random forest (RF) is a versatile, convenient model that exhibits different outputs. RF obstructs the overfitting problem. It is one of the main versions of ensemble learning. Ensemble learning is defined by using the same algorithm multiple times or using numerous algorithms. In this study, I used 5000 estimators, 25 max depth and 10 min samples leaf. Estimators refer to existing tree numbers in a forest. Here, the random state was 42. A random state is used for controlling the randomness of the samples when building trees.

3.5 Extreme Gradient Boosting (XGBoost)

XGBoost is a supervised learning algorithm that implements a method to generate accurate models called boosting. Supervised learning applies from a series of notable training examples to the task of inferencing a predictive model. Here, I used 40 estimators. Estimators are the number of performing boosting stages. The subsample used here was 0.7, and the depth was 12. Subsample refers to the fraction of samples to fit the base learners.

3.6 Multi-layer Perceptron Classifier (MLPClassifier)

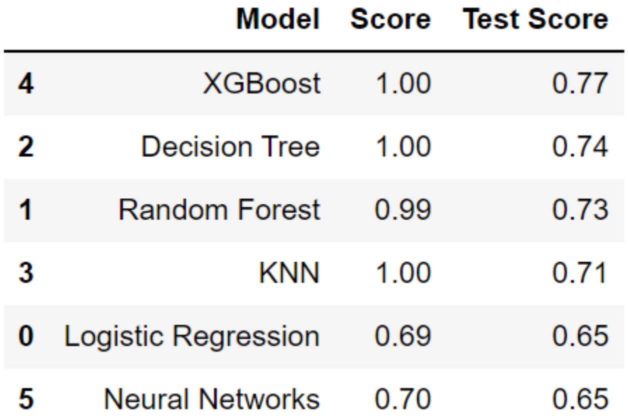
The multilayer perceptron (MLP) is a feedforward artificial neural network model that maps input data sets to a set of appropriate outputs. An MLP consists of multiple layers and each layer is fully connected to the following one. The nodes of the layers are neurons with nonlinear activation functions, except for the nodes of the input layer. Between the input and the output layer there may be one or more nonlinear hidden layers.

4. Results and Discussion

In this section, results are analyzed which are given by several different classification algorithms. In the experiment, the dataset is divided into training set and testing set. The ratio of the training set is 70% and 30% respectively. In this work, 10- fold cross-validation is used to train and test the machine learning model. The experiment is conducted in Python programming language and the library used are pandas and sci-kit learn.

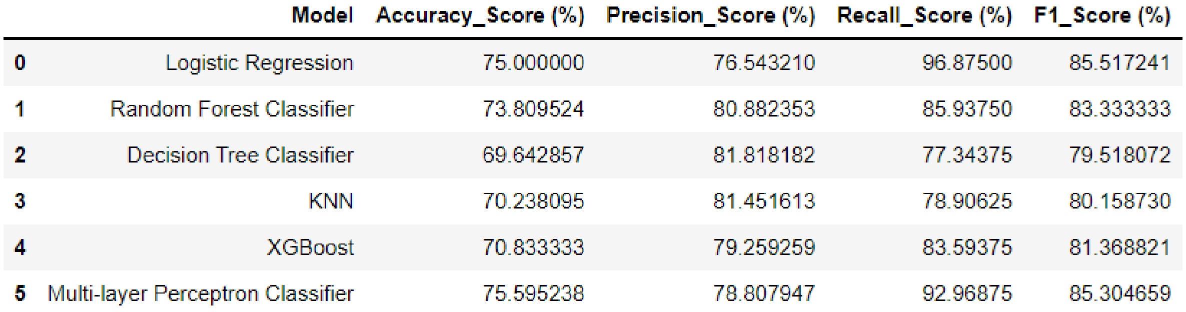
Model Evaluation:

It seems it is overfitting. The training set score got a better score than the testing set score.



	Model	Score	Test Score
4	XGBoost	1.00	0.77
2	Decision Tree	1.00	0.74
1	Random Forest	0.99	0.73
3	KNN	1.00	0.71
0	Logistic Regression	0.69	0.65
5	Neural Networks	0.70	0.65

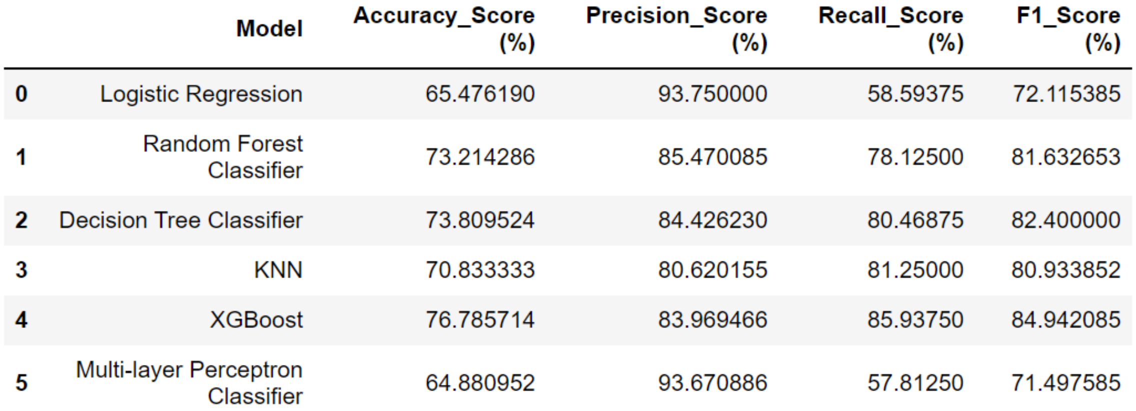
The comparison of various decision tree algorithms performed on liver disease data is shown in below:



	Model	Accuracy_Score (%)	Precision_Score (%)	Recall_Score (%)	F1_Score (%)
0	Logistic Regression	75.000000	76.543210	96.87500	85.517241
1	Random Forest Classifier	73.809524	80.882353	85.93750	83.333333
2	Decision Tree Classifier	69.642857	81.818182	77.34375	79.518072
3	KNN	70.238095	81.451613	78.90625	80.158730
4	XGBoost	70.833333	79.259259	83.59375	81.368821
5	Multi-layer Perceptron Classifier	75.595238	78.807947	92.96875	85.304659

Dealing with unbalance data:

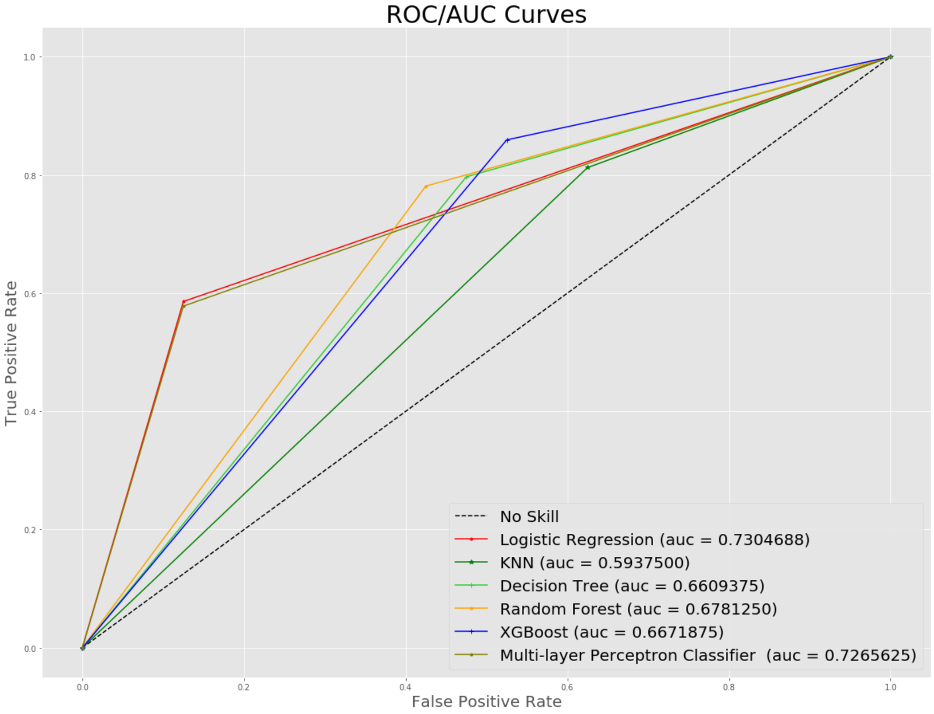
I noticed that the dataset contains information that a scale is not balanced. The labeled data is imbalance since the dataset used in the study consist of 167 is tagged as 0 (no liver disease) and 416 is tagged as 1 (liver disease).
After up-sampled the data, I obtained a better result is shown in below:



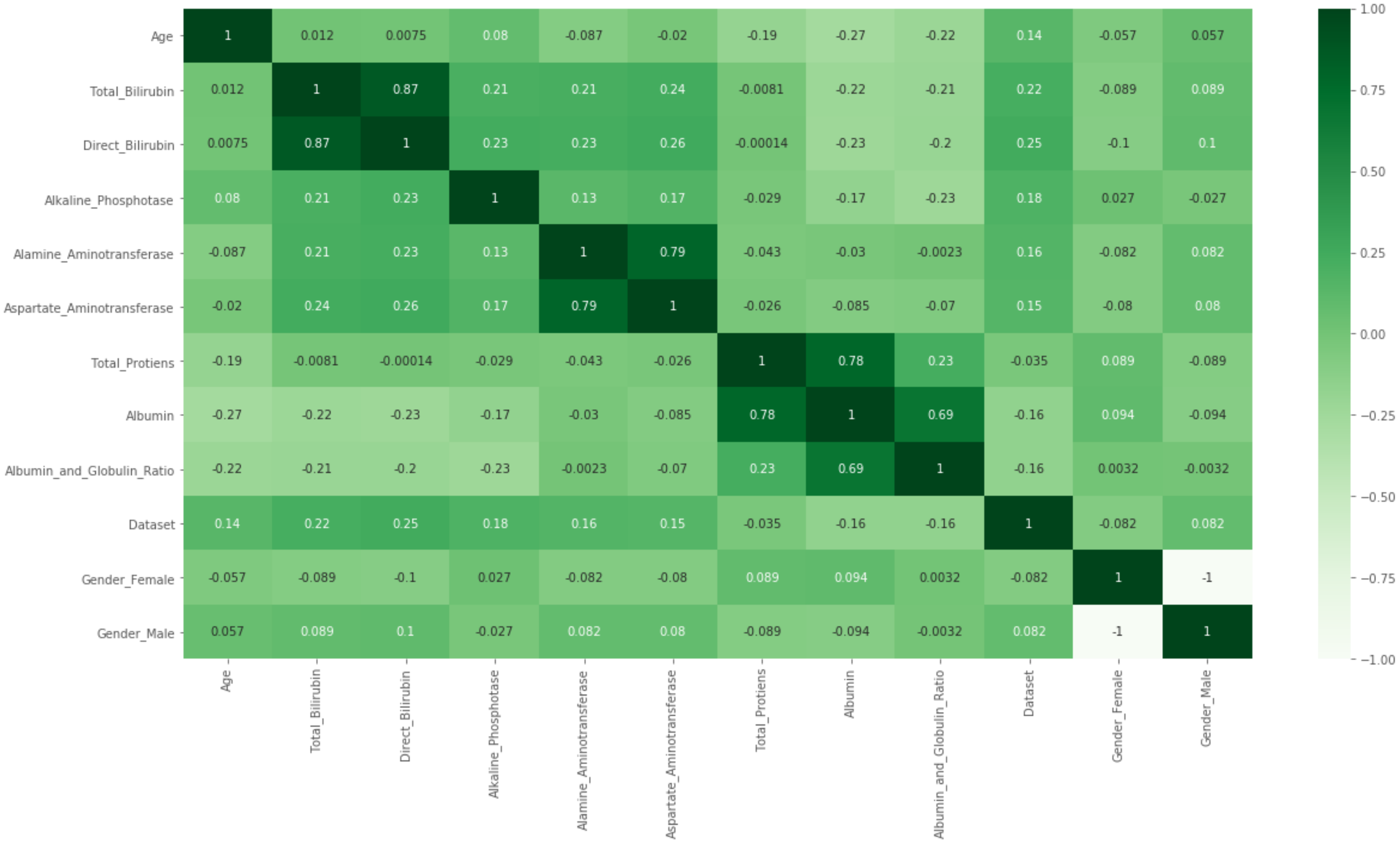
	Model	Accuracy_Score (%)	Precision_Score (%)	Recall_Score (%)	F1_Score (%)
0	Logistic Regression	65.476190	93.750000	58.59375	72.115385
1	Random Forest Classifier	73.214286	85.470085	78.12500	81.632653
2	Decision Tree Classifier	73.809524	84.426230	80.46875	82.400000
3	KNN	70.833333	80.620155	81.25000	80.933852
4	XGBoost	76.785714	83.969466	85.93750	84.942085
5	Multi-layer Perceptron Classifier	64.880952	93.670886	57.81250	71.497585

From the table above, I obtained the accuracy, precision, recall, and F1 score for those proposed algorithms. From the results of those applied algorithms, the XGBoost showed satisfactory results among all applied algorithms.
Fig. 3 represents the AUC scores of the proposed algorithms from the ROC curve. I observed that Logistic regression showed the highest AUC score from the obtained AUC scores, which was 72.7%, and KNN showed the lowest score, which was 59.3%, among those six algorithms. From all the statistical results, this analysis showed that the XGBoost algorithm had the highest accuracy score and a not bad AUC score , respectively.

Fig. 3



1. There is evidence of correlations between certain features. What this tells us is that certain features can be indicative of other features being elevated as well. An example of this would be the high positive correlation between direct bilirubin and total bilirubin. If a patient were to come in with high levels of direct bilirubin, we would be safe to assume that the likelihood that they also have a high incidence of total bilirubin is quite high. The health care practitioner could choose to only administer certain tests and not others, which could potentially save a both the healthcare practitioner and patient vast amounts of time and resources.

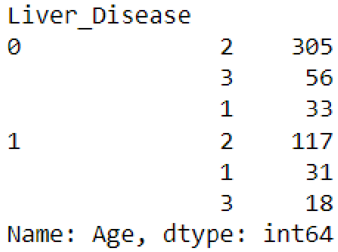


	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Dataset	Gender_Female	Gender_Male
Age	1	0.012	0.0075	0.08	-0.087	-0.02	-0.19	-0.27	-0.22	0.14	-0.057	0.057
Total_Bilirubin	0.012	1	0.87	0.21	0.21	0.24	-0.0081	-0.22	-0.21	0.22	-0.089	0.089
Direct_Bilirubin	0.0075	0.87	1	0.23	0.23	0.26	-0.00014	-0.23	-0.2	0.25	-0.1	0.1
Alkaline_Phosphotase	0.08	0.21	0.23	1	0.13	0.17	-0.029	-0.17	-0.23	0.18	0.027	-0.027
Alamine_Aminotransferase	-0.087	0.21	0.23	0.13	1	0.79	-0.043	-0.03	-0.0023	0.16	-0.082	0.082
Aspartate_Aminotransferase	-0.02	0.24	0.26	0.17	0.79	1	-0.026	-0.085	-0.07	0.15	-0.08	0.08
Total_Protiens	-0.19	-0.0081	-0.00014	-0.029	-0.043	-0.026	1	0.78	0.23	-0.035	0.089	-0.089
Albumin	-0.27	-0.22	-0.23	-0.17	-0.03	-0.085	0.78	1	0.69	-0.16	0.094	-0.094
Albumin_and_Globulin_Ratio	-0.22	-0.21	-0.2	-0.23	-0.0023	-0.07	0.23	0.69	1	-0.16	0.0032	-0.0032
Dataset	0.14	0.22	0.25	0.18	0.16	0.15	-0.035	-0.16	-0.16	1	-0.082	0.082
Gender_Female	-0.057	-0.089	-0.1	0.027	-0.082	-0.08	0.089	0.094	0.0032	-0.082	1	-1
Gender_Male	0.057	0.089	0.1	-0.027	0.082	0.08	-0.089	-0.094	-0.0032	0.082	-1	1

2. Males comprised most of the dataset by a vast amount. There are many more males than there are females affected by the liver disease as well. We are not made aware of how the data was acquired, but the disparity between genders is astonishing. The measures for all features associated with the disease were much greater in males than in females. These types of discoveries can lead to targeted preventive care for male subjects when they come in for rudimentary check-ups or health issues. Healthcare facilities and various other organizations can take a step in addressing the issue-at-hand and make the public aware of the consequences that are associated with liver disease.

3. Adults between the ages of 24-63 years old seem to be the most in danger. This can be attributed to these adults being tested more often for liver disease than are young people and the elderly. We may need to focus more on testing these other demographics. There is a strong possibility that we are missing the greater picture here. If I were to focus on testing the youth and improving their dietary intake, decreasing alcohol consumption, addressing pollution, decontaminating food, and eradicating drug use, I may arrive at a stage where we can prevent our youth from developing liver disease later in their lives.

age group 1 = youth, youth = 0-23
2 = adult, adult = 24-63
3 = elderly, elderly = 64-99.



```
Liver_Disease
0              2    305
               3     56
               1     33
1              2    117
               1     31
               3     18
Name: Age, dtype: int64
```

From all the people that were tested, 75.36% are adults.

Percent of people with liver disease that are in the adult age group of 24-63 years of age is 70.48%

5. Conclusions:

After implementing a plethora of classification machine learning algorithms on the final dataset, the default XGBoost algorithm proved to be the one that performed the best. It had the highest accuracy score with 0.73 and with AUC score 0.61. I then went ahead and focused on tuning the hyperparameters on the XGBoost. I specified the parameters associated with the grid beforehand. These can always be changed according to one's preference. I chose the ones I deemed most important in XGBoost algorithm. After performing the XGBoost with tuning parameters such as learning rate with 0.1, max depth with 9, n_estimator with 509, the accuracy with 0.77 was great than it was for the XGBoost algorithm on default parameters, although not by much. The AUC with 0.67 was also great than the default XGBoost. In general, it made some progress overall. The fine-tunning XGBoost algorithm shows promise when comparing it to all the other classification algorithms I used, but there is still significant room for improvement. I would play around with the hyperparameter limits and possibly even try several hyperparameter tuning tool. After reading a blog on hyperparameter tuning XGBoost algorithms, I realized that there are additional hyperparameters available to tune that can improve model accuracy and computational efficiency. However, it is time-consuming process, so I only show all of work I've done so far.