Liver Disease Prediction

Final Project

Group 15

Karthik Macha

Vignesh Sivakumar

857-204-9749 (Tel of Student 1)
617-201-2402 (Tel of Student 2)

macha.k@northeastern.edu

sivakumar.vig@northeastern.edu

Percentage of effort Contributed by Student 1: 50

Percentage of effort Contributed by Student 2: 50

Signature of Student 1: Karthik Macha

Signature of Student 2: Vignesh Sivakumar

Submission Date: 12/09/2022

Project selection and problem definition

Problem Setting

Liver diseases are rapidly becoming one of the most lethal diseases in several countries. Patients with liver disease have been steadily increasing because of excessive alcohol consumption, inhalation of harmful gases, and consumption of contaminated food, pickles, and drugs. This dataset was used to test prediction algorithms to reduce the burden on doctors.

Problem Definition

The liver damage caused by various diseases is generally irreversible. However, if liver disease is detected early and the underlying cause is treated, further damage can be avoided. Treatment is determined by the cause of the disease and the extent of the damage. In this project, we will identify the best models based on a patient's lifestyle and health conditions to predict liver disease.

Problem Sources

The liver disease prediction dataset is taken from the website Kaggle, which is an online community of data scientists and machine learning practitioners and an open-source data-science platform to publish analytics projects.

(https://www.kaggle.com/datasets/abhi8923shriv/liver-disease-patient-dataset).

Problem Description

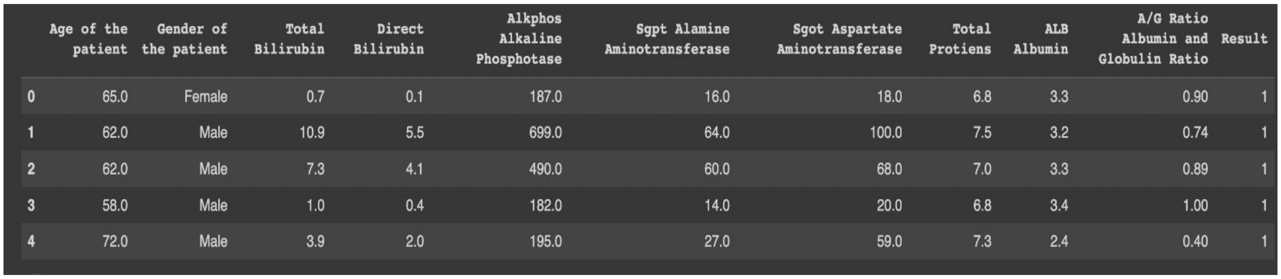
This data set (Liver Patient Dataset (LPD)_train.csv) contains 11 features or columns and 30,691 observations or rows. These characteristics are related to the patient's demographic information, including age, gender, and blood samples like bilirubin and albumin. The predictor variables are all characteristics related to the patient's demographic information, including age, gender, and blood samples. The Result is the Response Variable. Each of these elements contributes to the diagnosis of liver disease in a patient. The null values in the dataset need to be either replaced or removed. Additionally, this dataset may need to be up-sampled due to the unbalanced classes in the dataset.

Data collection

Data Collection and Preprocessing

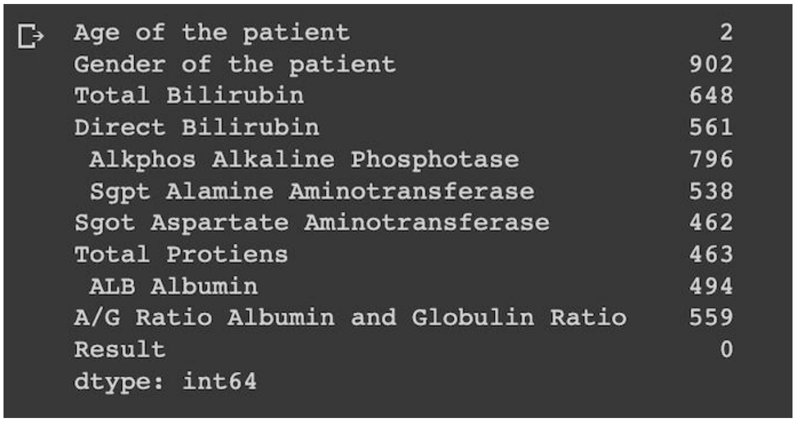
The data for this project is collected from Kaggle. There are 30,691 observations or rows in this data set (Liver Patient Dataset (LPD) train.csv), which has 11 characteristics or columns.

From the 11 variables, 10 are predictor variables and 1 is a responsible variable Result. There are both integer and categorical variables included in the dataset.

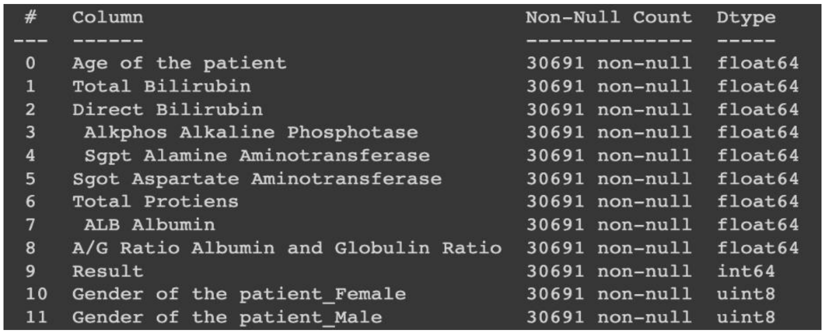


	Age of the patient	Gender of the patient	Total Bilirubin	Direct Bilirubin	Alkphos Alkaline Phosphotase	Sgpt Alamine Aminotransferase	Sgot Aspartate Aminotransferase	Total Protiens	ALB Albumin	A/G Ratio Albumin and Globulin Ratio	Result
0	65.0	Female	0.7	0.1	187.0	16.0	18.0	6.8	3.3	0.90	1
1	62.0	Male	10.9	5.5	699.0	64.0	100.0	7.5	3.2	0.74	1
2	62.0	Male	7.3	4.1	490.0	60.0	68.0	7.0	3.3	0.89	1
3	58.0	Male	1.0	0.4	182.0	14.0	20.0	6.8	3.4	1.00	1
4	72.0	Male	3.9	2.0	195.0	27.0	59.0	7.3	2.4	0.40	1

This dataset contains null values that need to be either updated or removed, which is a challenge. Because using the mean could lead the data to being skewed, we calculated the median of the remaining entries in the columns to replace the null values for numerical variables. Additionally, there are no fewer null values than there are entries to be deleted. When dealing with categorical variables, there is only one column, "Gender of the Patient," and any missing values can be filled in by employing mode imputation.

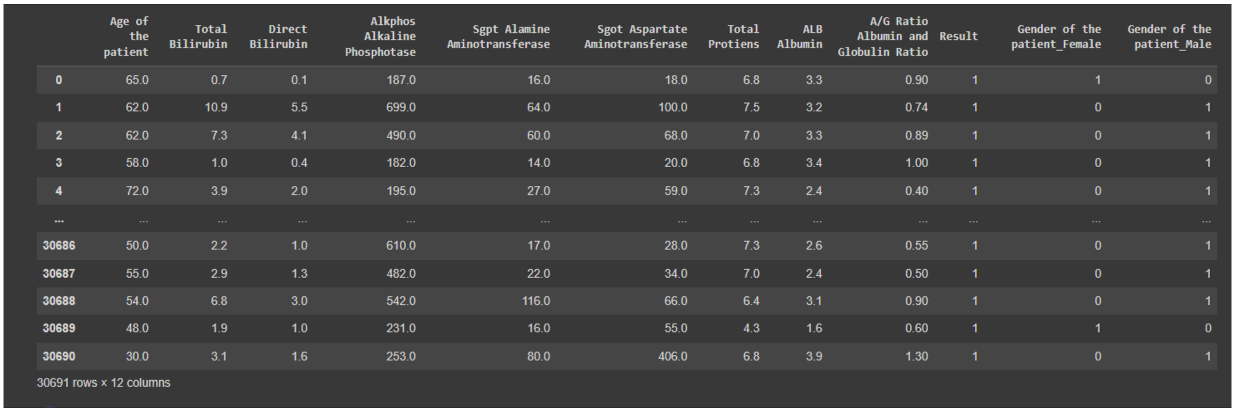


```
Age of the patient                       2
Gender of the patient                  902
Total Bilirubin                        648
Direct Bilirubin                       561
 Alkphos Alkaline Phosphotase          796
 Sgpt Alamine Aminotransferase         538
Sgot Aspartate Aminotransferase        462
Total Protiens                         463
 ALB Albumin                           494
A/G Ratio Albumin and Globulin Ratio   559
Result                                   0
dtype: int64
```



```
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Age of the patient                    30691 non-null  float64
 1   Total Bilirubin                       30691 non-null  float64
 2   Direct Bilirubin                      30691 non-null  float64
 3    Alkphos Alkaline Phosphotase         30691 non-null  float64
 4    Sgpt Alamine Aminotransferase        30691 non-null  float64
 5   Sgot Aspartate Aminotransferase       30691 non-null  float64
 6   Total Protiens                        30691 non-null  float64
 7    ALB Albumin                          30691 non-null  float64
 8   A/G Ratio Albumin and Globulin Ratio  30691 non-null  float64
 9   Result                                30691 non-null  int64
 10  Gender of the patient_Female          30691 non-null  uint8
 11  Gender of the patient_Male            30691 non-null  uint8
```

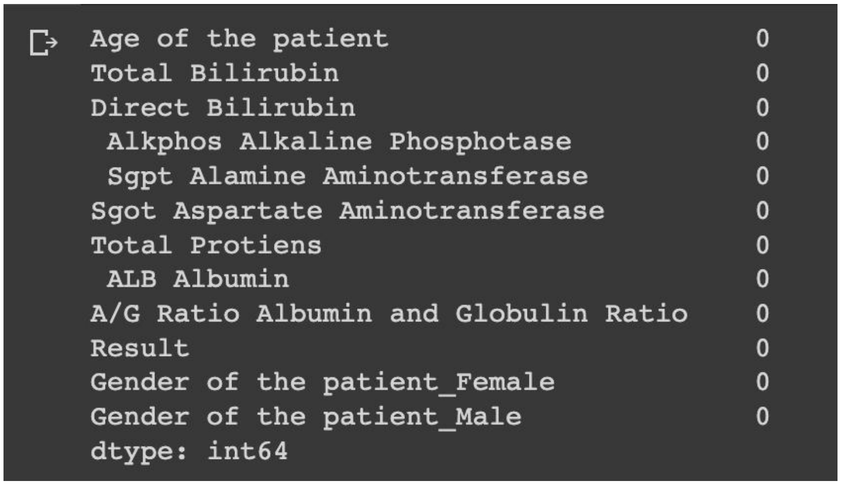
We have also converted the categorical variables into numerical variables for the column "Gender of the Patient".



	Age of the patient	Total Bilirubin	Direct Bilirubin	Alkphos Alkaline Phosphotase	Sgpt Alamine Aminotransferase	Sgot Aspartate Aminotransferase	Total Protiens	ALB Albumin	A/G Ratio Albumin and Globulin Ratio	Result	Gender of the patient_Female	Gender of the patient_Male
0	65.0	0.7	0.1	187.0	16.0	18.0	6.8	3.3	0.90	1	1	0
1	62.0	10.9	5.5	699.0	64.0	100.0	7.5	3.2	0.74	1	0	1
2	62.0	7.3	4.1	490.0	60.0	68.0	7.0	3.3	0.89	1	0	1
3	58.0	1.0	0.4	182.0	14.0	20.0	6.8	3.4	1.00	1	0	1
4	72.0	3.9	2.0	195.0	27.0	59.0	7.3	2.4	0.40	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...
30686	50.0	2.2	1.0	610.0	17.0	28.0	7.3	2.6	0.55	1	0	1
30687	55.0	2.9	1.3	482.0	22.0	34.0	7.0	2.4	0.50	1	0	1
30688	54.0	6.8	3.0	542.0	116.0	66.0	6.4	3.1	0.90	1	0	1
30689	48.0	1.9	1.0	231.0	16.0	55.0	4.3	1.6	0.60	1	1	0
30690	30.0	3.1	1.6	253.0	80.0	406.0	6.8	3.9	1.30	1	0	1

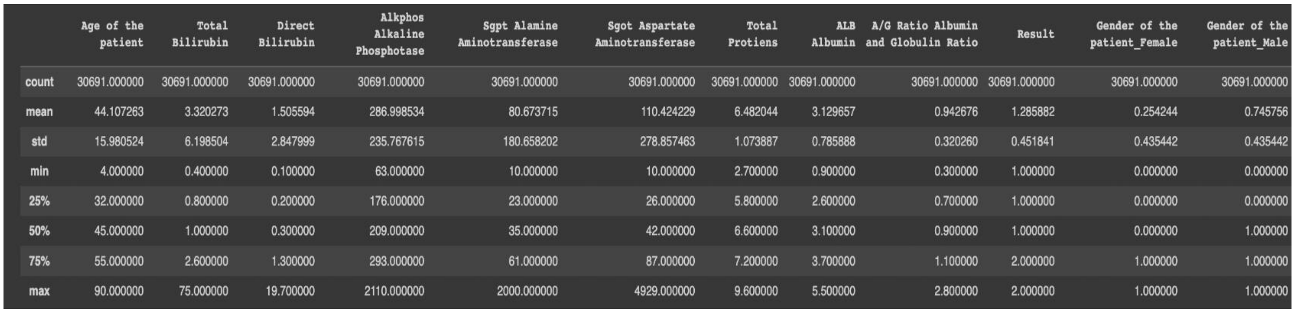
30691 rows × 12 columns

The dataset is preprocessed and ready to be used for the project after the missing values are replaced using the techniques.



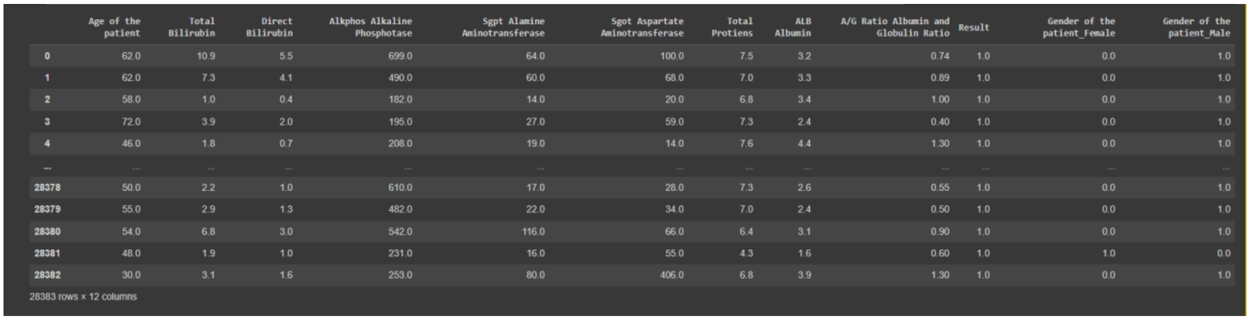
```
Age of the patient                      0
Total Bilirubin                         0
Direct Bilirubin                        0
 Alkphos Alkaline Phosphotase           0
 Sgpt Alamine Aminotransferase          0
Sgot Aspartate Aminotransferase         0
Total Protiens                          0
 ALB Albumin                            0
A/G Ratio Albumin and Globulin Ratio    0
Result                                  0
Gender of the patient_Female            0
Gender of the patient_Male              0
dtype: int64
```

We have calculated the statistics of each variable using describe ().



	Age of the patient	Total Bilirubin	Direct Bilirubin	Alkphos Alkaline Phosphotase	Sgpt Alamine Aminotransferase	Sgot Aspartate Aminotransferase	Total Protiens	ALB Albumin	A/G Ratio Albumin and Globulin Ratio	Result	Gender of the patient_Female	Gender of the patient_Male
count	30691.000000	30691.000000	30691.000000	30691.000000	30691.000000	30691.000000	30691.000000	30691.000000	30691.000000	30691.000000	30691.000000	30691.000000
mean	44.107263	3.320273	1.505594	286.998534	80.673715	110.424229	6.482044	3.129657	0.942676	1.285882	0.254244	0.745756
std	15.980524	6.198504	2.847999	235.767615	180.658202	278.857463	1.073887	0.785888	0.320260	0.451841	0.435442	0.435442
min	4.000000	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000	0.900000	0.300000	1.000000	0.000000	0.000000
25%	32.000000	0.800000	0.200000	176.000000	23.000000	26.000000	5.800000	2.600000	0.700000	1.000000	0.000000	0.000000
50%	45.000000	1.000000	0.300000	209.000000	35.000000	42.000000	6.600000	3.100000	0.900000	1.000000	0.000000	1.000000
75%	55.000000	2.600000	1.300000	293.000000	61.000000	87.000000	7.200000	3.700000	1.100000	2.000000	1.000000	1.000000
max	90.000000	75.000000	19.700000	2110.000000	2000.000000	4929.000000	9.600000	5.500000	2.800000	2.000000	1.000000	1.000000

Outlier Detection and Removal



	Age of the patient	Total Bilirubin	Direct Bilirubin	Alkphos Alkaline Phosphotase	Sgpt Alamine Aminotransferase	Sgot Aspartate Aminotransferase	Total Protiens	ALB Albumin	A/G Ratio Albumin and Globulin Ratio	Result	Gender of the patient_Female	Gender of the patient_Male
0	62.0	10.9	5.5	699.0	64.0	100.0	7.5	3.2	0.74	1.0	0.0	1.0
1	62.0	7.3	4.1	490.0	60.0	68.0	7.0	3.3	0.89	1.0	0.0	1.0
2	58.0	1.0	0.4	182.0	14.0	20.0	6.8	3.4	1.00	1.0	0.0	1.0
3	72.0	3.9	2.0	195.0	27.0	59.0	7.3	2.4	0.40	1.0	0.0	1.0
4	46.0	1.8	0.7	208.0	19.0	14.0	7.6	4.4	1.30	1.0	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...
28378	50.0	2.2	1.0	610.0	17.0	28.0	7.3	2.6	0.55	1.0	0.0	1.0
28379	55.0	2.9	1.3	482.0	22.0	34.0	7.0	2.4	0.50	1.0	0.0	1.0
28380	54.0	6.8	3.0	542.0	116.0	66.0	6.4	3.1	0.90	1.0	0.0	1.0
28381	48.0	1.9	1.0	231.0	16.0	55.0	4.3	1.6	0.60	1.0	1.0	0.0
28382	30.0	3.1	1.6	253.0	80.0	406.0	6.8	3.9	1.30	1.0	0.0	1.0

28383 rows × 12 columns

We have removed the outliers by calculating the values of the numeric columns which are 4 standard deviations (cut off) from the mean. We have then deleted those rows which are having values lower or higher than the cutoff we have defined.

Splitting the dataset

We split the data into train and test by using sklearn library method train_test_split() and get the shape of the train data as (19868,9) and test data as (8515,9).

Then we implemented the Standard scaler method of standardizing data such the transformed feature has 0 mean and standard deviation of 1. The standard scaler scales all the data and brings all data to equal units.

The challenge of working with imbalanced datasets is that of poor performance on, the minority class, which is by far the most important.

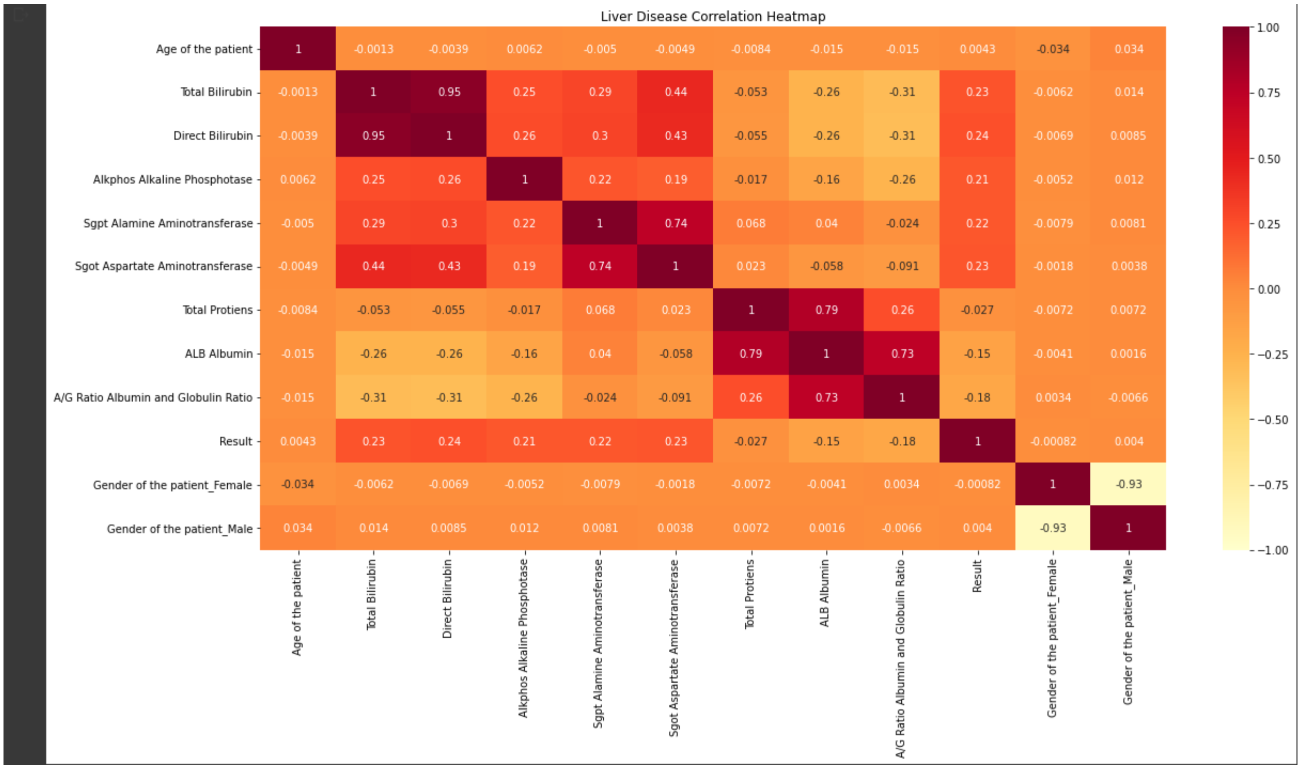
Oversampling the minority class is one way to deal with unbalanced datasets. Duplicating examples from the minority class is the simplest method, but these examples don't provide any new insight into the model. Instead, fresh examples can be created by synthesizing the old ones.

This is a type of data augmentation for the minority class and is referred to as the Synthetic Minority Oversampling Technique or SMOTE. We have used this function to balance the classes.

Dimensional Reduction and Variable Selection

We need to transform data from a high-dimensional space into a low-dimensional space with the intention of keeping the low-dimensional representation as close as possible to the inherent dimension of the original data, this in turn is called as dimension reduction. Working with high-dimensional spaces can be undesirable for different reasons such as the fact that the data analysis is typically computationally uncontrollable, and that the raw data are frequently sparse because of the curse of dimensionality.

We have built a heatmap to visualize highly correlated columns.

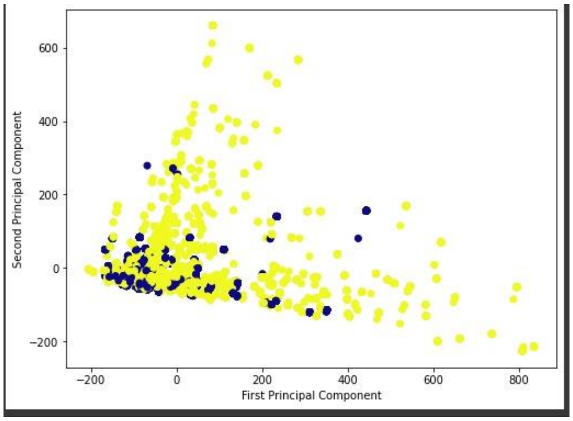


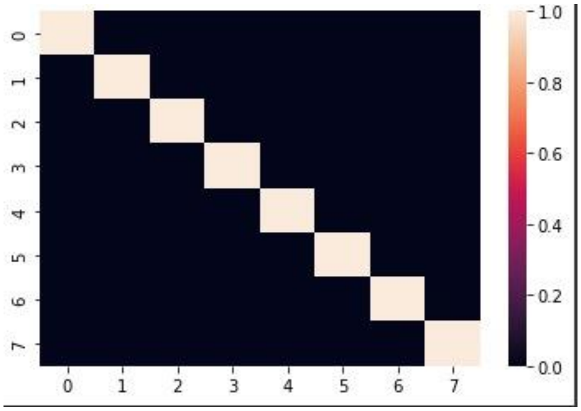
	Age of the patient	Total Bilirubin	Direct Bilirubin	Alkphos Alkaline Phosphotase	Sgpt Alamine Aminotransferase	Sgot Aspartate Aminotransferase	Total Protiens	ALB Albumin	A/G Ratio Albumin and Globulin Ratio	Result	Gender of the patient_Female	Gender of the patient_Male
Age of the patient	1	-0.0013	-0.0039	0.0062	-0.005	-0.0049	-0.0084	-0.015	-0.015	0.0043	-0.034	0.034
Total Bilirubin	-0.0013	1	0.95	0.25	0.29	0.44	-0.053	-0.26	-0.31	0.23	-0.0062	0.014
Direct Bilirubin	-0.0039	0.95	1	0.26	0.3	0.43	-0.055	-0.26	-0.31	0.24	-0.0069	0.0085
Alkphos Alkaline Phosphotase	0.0062	0.25	0.26	1	0.22	0.19	-0.017	-0.16	-0.26	0.21	-0.0052	0.012
Sgpt Alamine Aminotransferase	-0.005	0.29	0.3	0.22	1	0.74	0.068	0.04	-0.024	0.22	-0.0079	0.0081
Sgot Aspartate Aminotransferase	-0.0049	0.44	0.43	0.19	0.74	1	0.023	-0.058	-0.091	0.23	-0.0018	0.0038
Total Protiens	-0.0084	-0.053	-0.055	-0.017	0.068	0.023	1	0.79	0.26	-0.027	-0.0072	0.0072
ALB Albumin	-0.015	-0.26	-0.26	-0.16	0.04	-0.058	0.79	1	0.73	-0.15	-0.0041	0.0016
A/G Ratio Albumin and Globulin Ratio	-0.015	-0.31	-0.31	-0.26	-0.024	-0.091	0.26	0.73	1	-0.18	0.0034	-0.0066
Result	0.0043	0.23	0.24	0.21	0.22	0.23	-0.027	-0.15	-0.18	1	-0.00082	0.004
Gender of the patient_Female	-0.034	-0.0062	-0.0069	-0.0052	-0.0079	-0.0018	-0.0072	-0.0041	0.0034	-0.00082	1	-0.93
Gender of the patient_Male	0.034	0.014	0.0085	0.012	0.0081	0.0038	0.0072	0.0016	-0.0066	0.004	-0.93	1

From the heatmap, we can say that the columns, Direct Bilirubin and Total Bilirubin, Sgpt Alamine Aminotransferase and Sgot Aspartate Aminotransferase, Total Proteins and ALB albumin, Gender of the patient Female, Gender of the patient Male are highly correlated, but we will be dropping Total Proteins and Gender of the patient Male as it's not a good correlation with the target variable from the dataset i.e. reducing the dimension of the dataset.

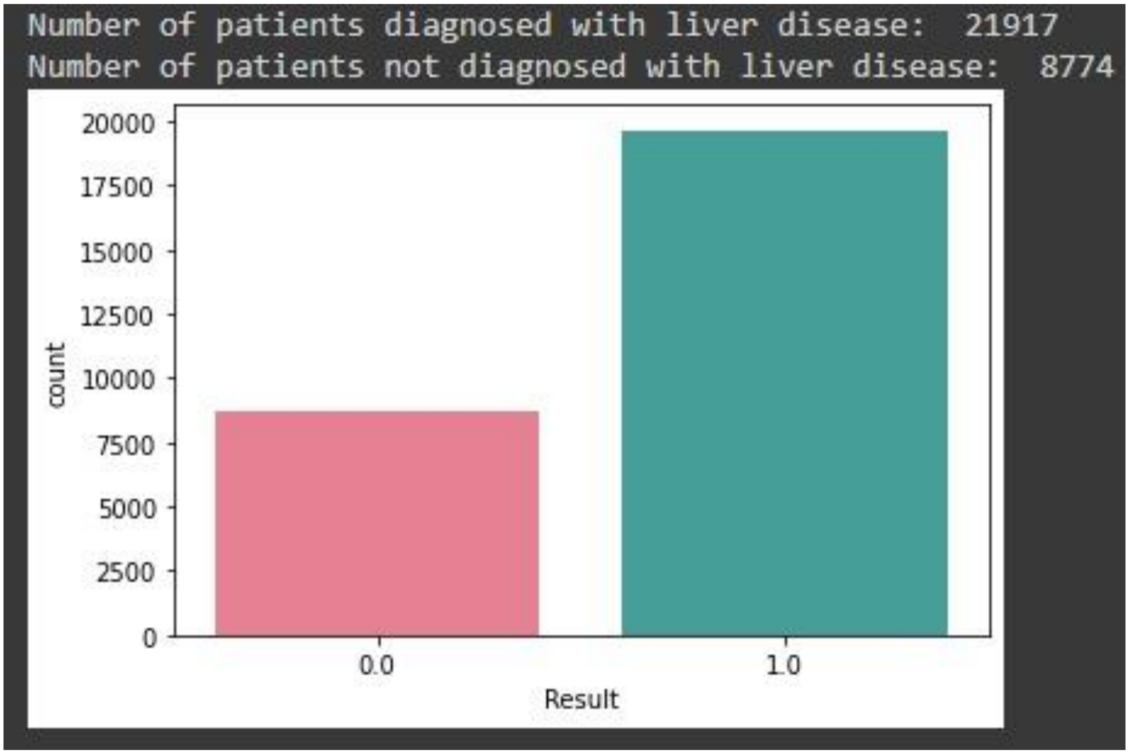
PCA

We will be incorporating unsupervised learning method for lowering the dimensionality of data using principal component analysis. While minimizing information loss, it simultaneously improves interpretability and makes data easier to plot in 2D and 3D and aids in identifying the dataset's most important properties. Large set of information is retained in a smaller set of dimensions. One of the limitations is that the Accuracy naturally suffers as a data set's number of variables is reduced.



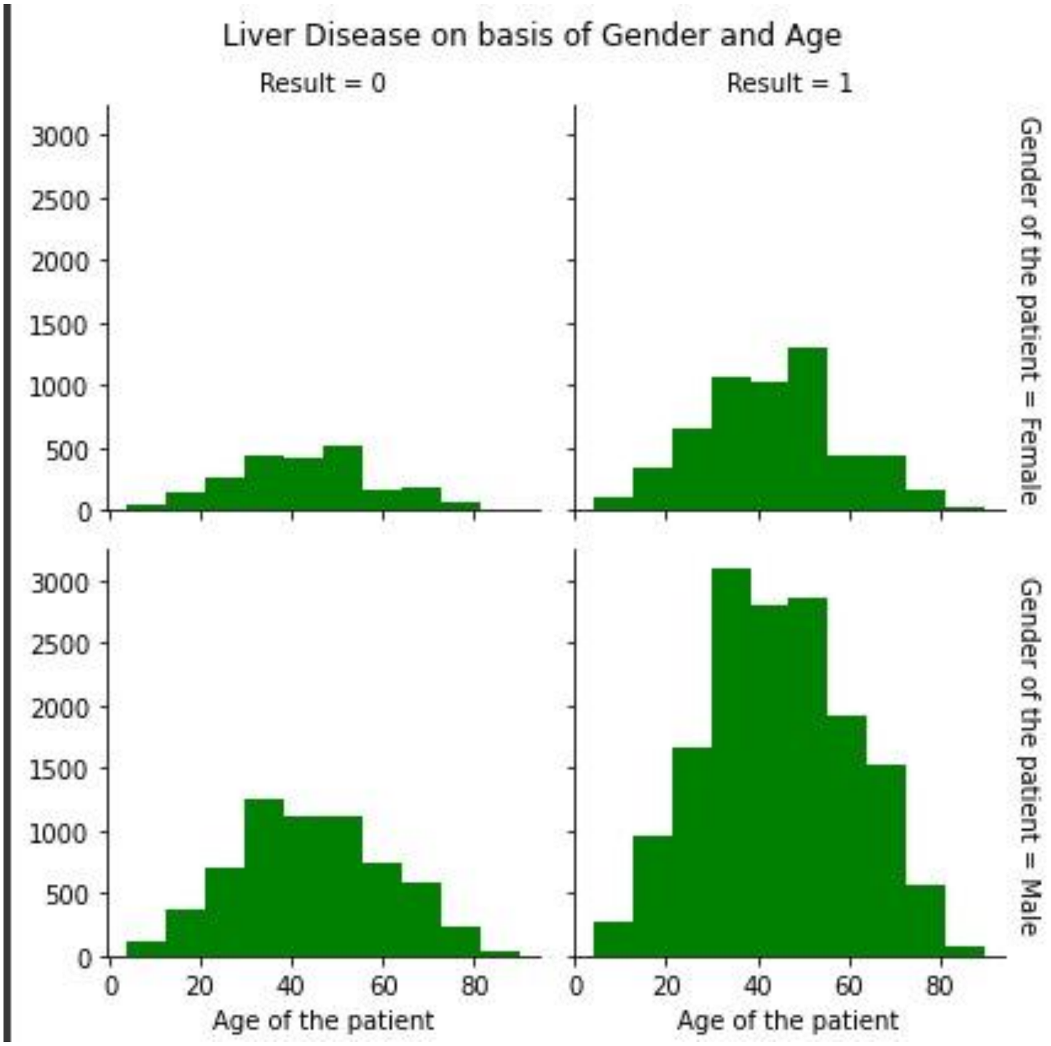


Data Visualization and Exploration

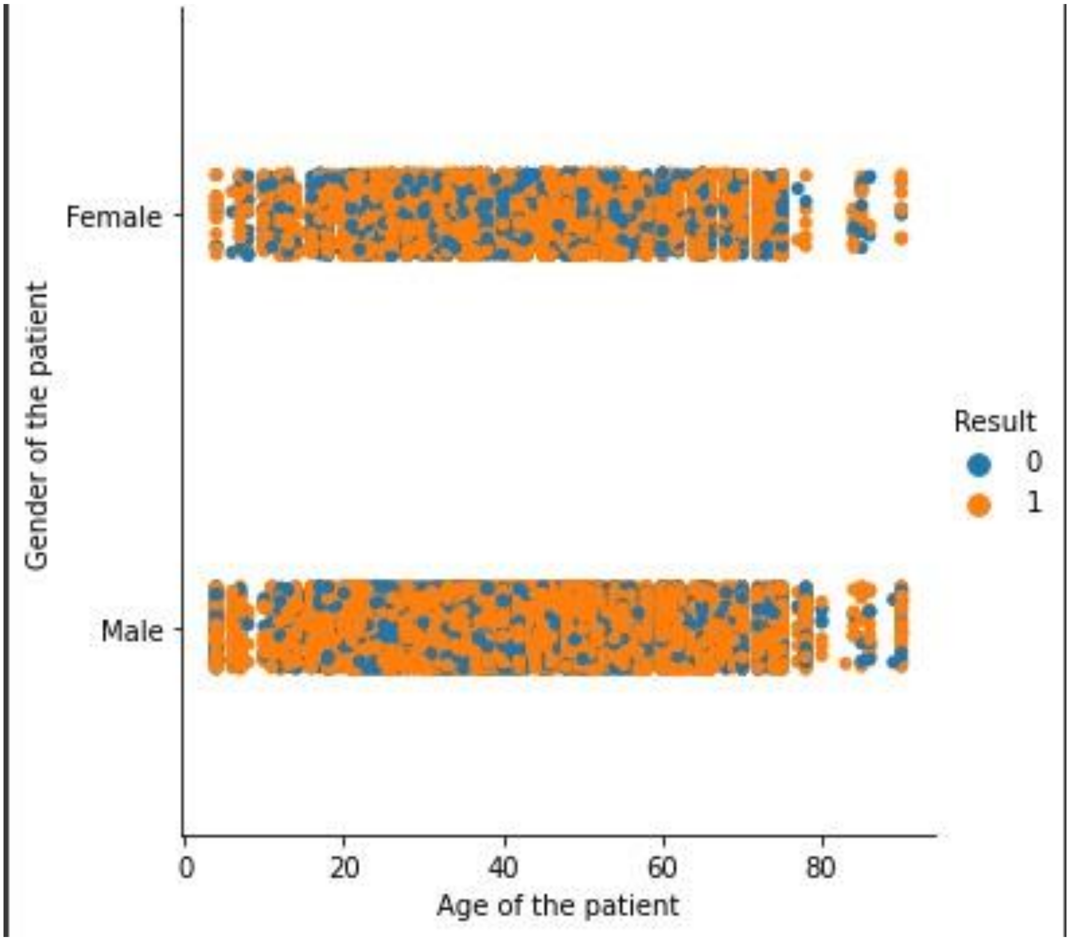


```
Number of patients diagnosed with liver disease:  21917
Number of patients not diagnosed with liver disease:  8774
```

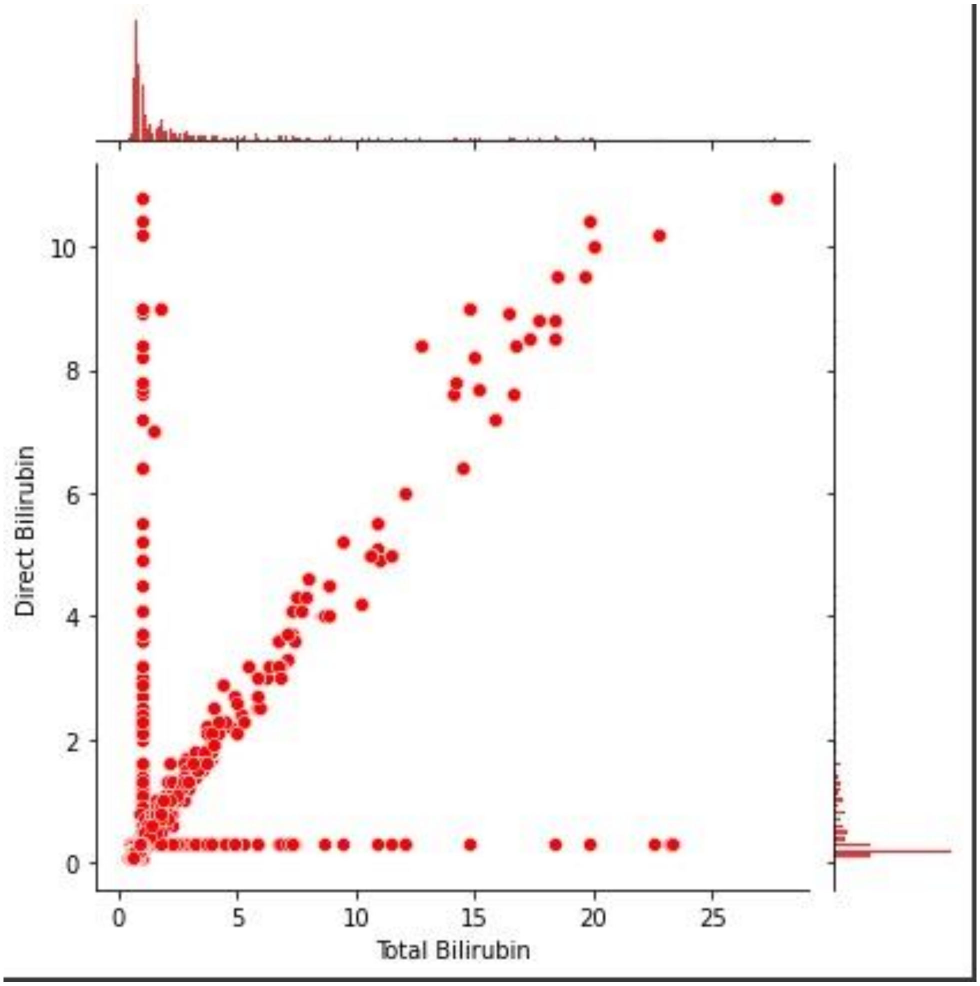
The above bar graph depicts the number of patients diagnosed with or without liver disease. The result 1 depicts the patient having liver disease and the result 0 depicts the number of patients not diagnosed with liver disease. We can observe that more patients are diagnosed with the liver disease.



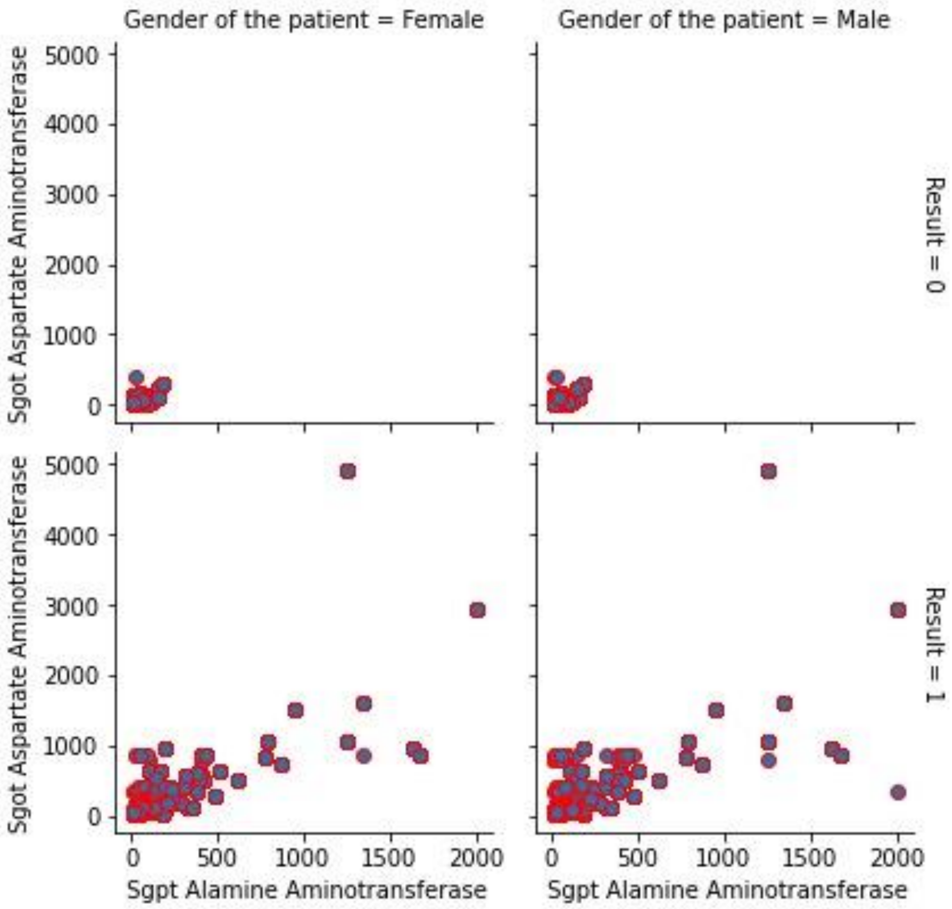
The above histogram depicts the Liver Disease on Basis of Gender and Age, and we can say that in both genders the age between 30 and 70 the chances of liver disease are higher.



The above categorical plot shows the different categories such as the gender of the patient and age of the patient. The blue color indicates the result 0 whereas the orange color indicates the result 1. This implies that male patients are having higher chances of liver disease as compared to female counterparts.



From the above scatter plot, we can see that the less the total Bilirubin, the higher the Direct Bilirubin. The least direct bilirubin is seen when the total bilirubin is at its highest. Total Bilirubin consists of direct as well as indirect Bilirubin.



From the above plot, we can say that the patient diagnosed with liver disease for all genders, has more Sgpt Alamine Aminotransferase and more Sgot Aspartate Aminotransferase than the ones not diagnosed with the disease.

Model exploration, selection and visualization

One of the most important steps in the Data Mining cycle is to select the appropriate model which will give better accuracy, precision, and F1 scores than other models. In our case, as we are dealing with a binary classification problem, we will be incorporating the following models:

1. Logistic Regression
2. K-NN
3. Random Forest Classifier
4. Decision Tree Classifier
5. PCA

Logistic Regression

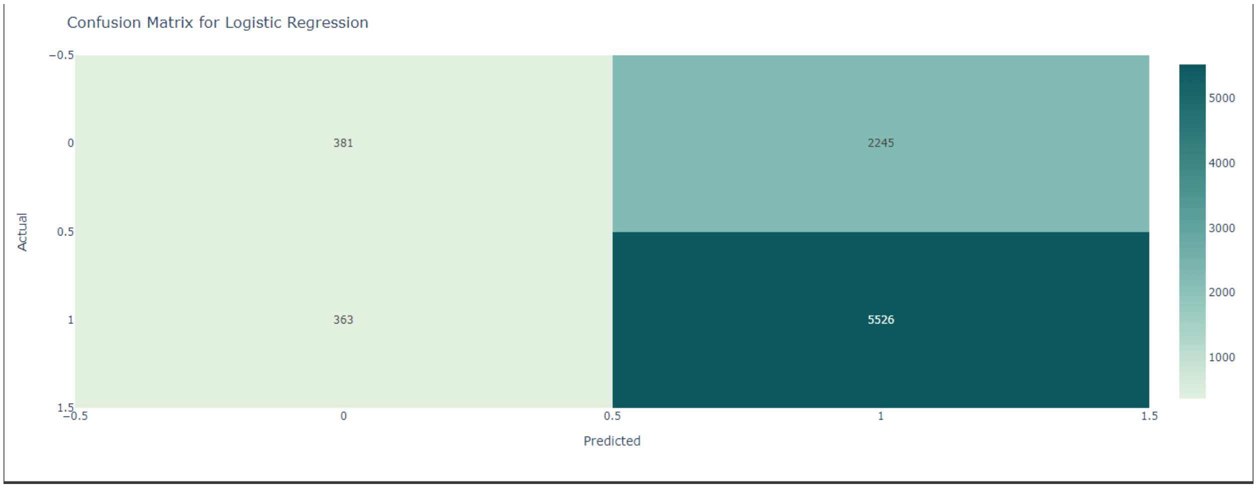
The logistic model in statistics is a statistical model that depicts the likelihood that an event will occur by making the event's log odds a linear combination of one or more independent variables. Logistic regression is used in regression analysis. Predictive analytics and categorization frequently make use of this kind of statistical model, also referred to as a logit model. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. This model gave the *least accuracy* score among all the models used.

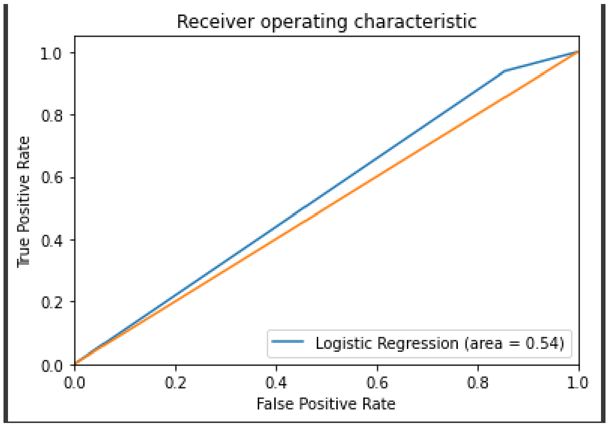
Accuracy: 69.37 %

Precision: 0.7111

Recall: 0.9383

F1 Score: 0.8090





K-NN

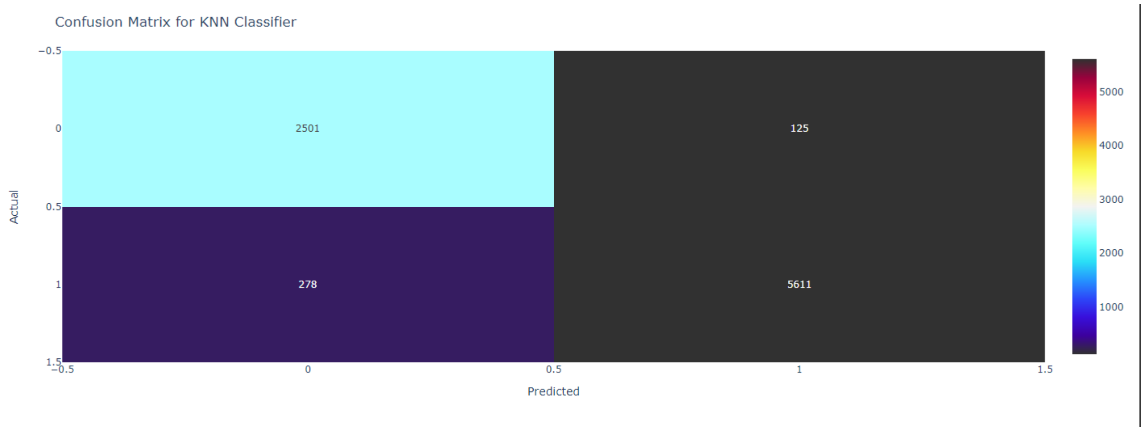
The K-nearest neighbors algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point. Although it can be applied to classification or regression issues, it is commonly employed as a classification algorithm because it relies on the idea that comparable points can be discovered close to one another. It is simple and flexible. One disadvantage of K-NN is that it uses more memory and data storage than other classifiers. Both in terms of time and money, this can be expensive.

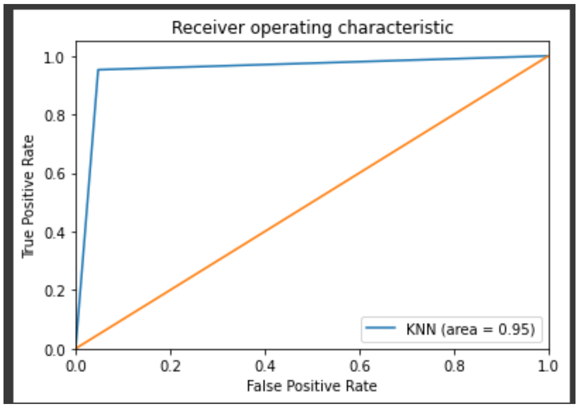
Accuracy: 95.26 %

Precision: 0.9782

Recall: 0.9527

F1 Score: 0.9653





<u>**Random Forest Classifier**</u>

A classification system made up of several decision trees is called the random forest. It is a supervised machine learning algorithm which attempts to produce an uncorrelated forest of trees whose forecast by committee is more accurate than that of any individual tree by using bagging and feature randomness when generating each individual tree. They are frequently employed in classification and regression issues. The Random Forest Algorithm's ability to handle data sets with both continuous variables, as in regression, and categorical variables, as in classification, is one of its most crucial qualities. In terms of classification issues, it delivers superior outcome. One limitation is that Random forests have been observed to overfit for some of the datasets.

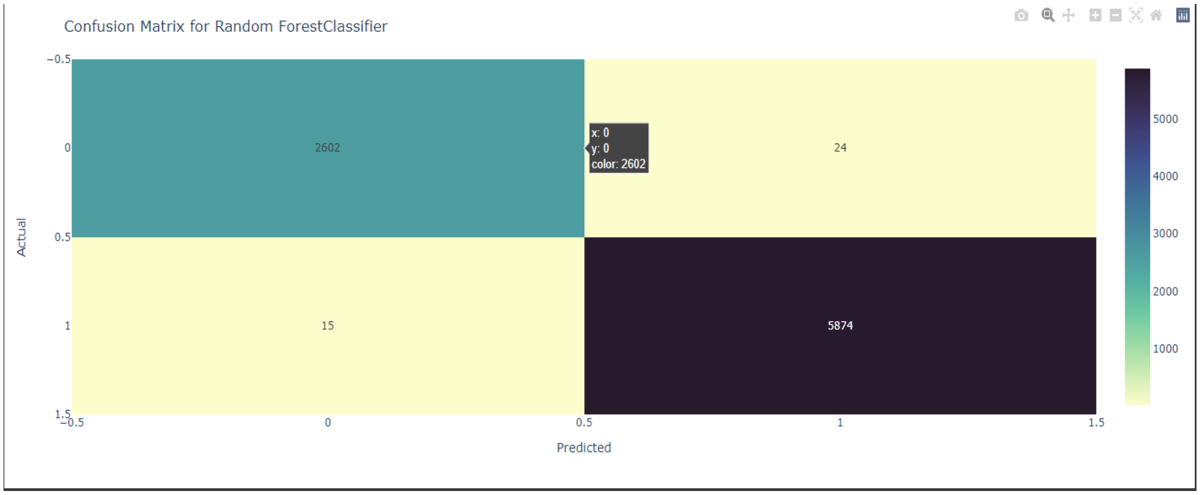
This model Performed very well and gave the **highest accuracy** score among others.

Accuracy: 99.54 %

Precision: 0.9959

Recall: 0.9974

F1 Score: 0.9966



Decision Tree Classifier

A supervised learning method used in statistics, data mining, and machine learning is the decision tree. A classification or regression decision tree is employed in this formalization as a prediction model to draw conclusions from a set of observations.

Classification trees are tree models where the target variable can take a discrete range of values. In these tree structures, the leaves correspond to class labels, and the branches to the attributes that combine to form those class labels.

Because they are straightforward and understandable, decision trees are among the most widely used machine learning methods.

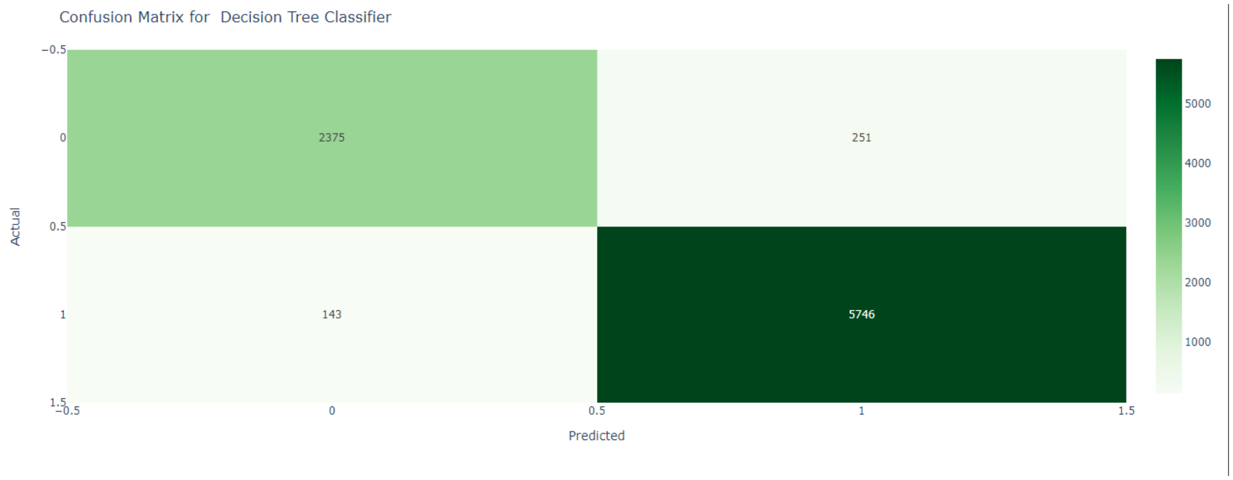
One issue with Decision tree is that if the size of the data is too big, then one single tree may grow a lot of nodes which might result in complexity and leads to overfitting.

Accuracy: 95.37 %

Precision: 0.9581

Recall: 0.9757

F1 Score: 0.9668



Model Performance Evaluation

When compared to the other models we utilized for this project, the random forest classifier has performed exceptionally well. Compared to other models, the accuracy attained is the highest. The accuracy of Logistic Regression is the lowest among all other models.

Models	Accuracy	Precision	Recall	F1 Score
Logistic	69.37 %	0.7111	0.9383	0.8090
K-NN	95.26 %	0.9782	0.9527	0.9653
Random Forest	99.54 %	0.9959	0.9974	0.9966
Decision Tree	95.37 %	0.9581	0.9757	0.9668

Impact of the Project Outcomes

Given that liver illness is fatal, liver disease prediction is crucial in the medical field. Our research focuses on predicting liver illness while considering several variables, including age, gender, and protein levels. Because it will ease the stress on the doctors, our project will be highly helpful in the medical domain.