Ahmednagar Jilha Maratha Vidya Prasarak Samaj's

New Arts, Commerce and Science College (Autonomous), Ahmednagar

Department of Statistics

Project report on

Title: Indian Liver Patients Disease Prediction and classification

Submitted by ,

1. Mengade Aishwarya Namdev
2. Shahane Prashant Bharat
3. Bhuse Ashwini Sambhaji

Head of the Department **Under the guidance of**

Dr. A. A Kulkarni **Prof. B. P. Kharat**

Ahmednagar Jilha Maratha Vidya Prasarak Samaj‘s

New Arts, Commerce and Science College (Autonomous) Ahmednagar

Department of Statistics

(M. Sc. II)

CERTIFICATE

This is to certify that____, a student of class M. Sc. II has successfully completed the project “**Indian Liver Patients Disease Prediction and Classification**” during the year 2022-2023.

.

Date:
Place: Ahmednagar

Prof. B.P. Kharat		**Dr. A. A. Kulkarni**
		Head
Teacher-In-Charge	**External Examiner**	Department of Statistics

Acknowledgements

In our efforts towards realisation of our project work, we have drawn on the guidance of many people, whom we wish to acknowledge. We are thankful to our Principal Prof. Dr. B. H. Zaware, Vice Principal Dr. A. E. Athare and Head of the Department of Statistics Dr. A. A Kulkarni for their support and co-operation towards successful completion of this project. We are thankful to this opportunity and express our deep sense of gratitude and whole hearted thanks to our guide **Prof. B. P. Kharat** for giving their priceless guidance, inspiration and encouragement to embark this project. We are also extremely thankful to Dr. B. K. Thorve, Prof. K. B. Mane, Prof. M. D. Rohakale, Prof M. Z. Shaikh and Prof. K. K. Kawale for their support and guidance for this project. Last but not the least we would like to thank all those who helped us directly and indirectly in our endeavour.

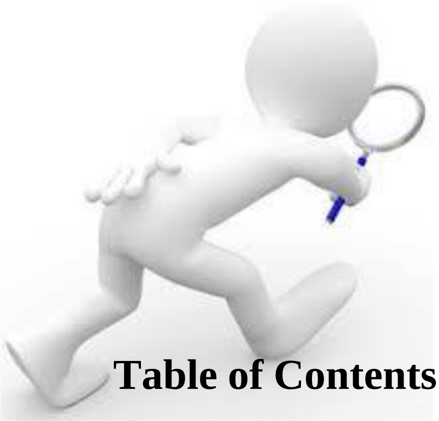


Table of Contents:

shutterstock.com · 66926647

Abstract

The Data Indian Patient Records consists of 583 rows and 11 columns; from which 1 variable is binary, 1 is string and 9 are numerical. The data is taken from https://www.kaggle.com/datasets/uciml/indian-liver-patient-records . Since our data has categorical variables, we have applied logistic regression to see which of the factors are important for risk of liver disease. To check which group of people have risk of liver disease, we have done the survival analysis. For checking the model accuracies we have applied logistic regression, K-Nearest neighborhood, Decision Tree and Random Forest.

Since the data has 441 data points / observations having Gender Male and 142 data points having Gender Female. So we have done Oversampling data Technique to balance the dataset for further Machine LearningTechniques.

Keywords: Survival Statistical Analysis, Logistic Regression, K- Nearest neighborhood, Decision Tree, Random Forest.

Introduction

The Indian Liver Patient Dataset is a collection of medical records and data from Indian patients with liver-related conditions. This dataset provides valuable insights into liver diseases prevalent in India and can aid in understanding the factors affecting liver health in the Indian population. The objective of this report is to analyze the dataset, explore key trends and patterns, and gain insights that can contribute to improved diagnosis, treatment, and preventive measures for liver diseases in India.

Liver diseases pose a significant global health burden, affecting millions of people worldwide. These conditions encompass a wide range of disorders, including hepatitis, cirrhosis, non-alcoholic fatty liver disease (NAFLD), liver cancer, and autoimmune liver diseases. Timely detection and accurate diagnosis are crucial for effective management and the prevention of further complications. What can we do with this dataset?

This dataset can be used to apply a range of machine learning methods, most notably classifier models (logistic regression, Decision Tree, KNN, random forest, etc. We should treat the variable "LiverDisease" as a binary ("Yes" - respondent had heart disease; "No" - respondent had no heart disease). But note that classes are not balanced, so the classic model application approach is not advisable.

How to balance the data set?

Since the data has 441 data points / observations having Gender Male and 142 data points having Gender Female. So we have done Oversampling data Technique to balance the dataset for further Machine LearningTechniques.

.

Data Description:

There are 11 variables (1 binary, 1 string and 9 numerical) with 11 columns and 583 rows.

1) Age : Age of the patients

2) Gender : Sex of the patients

3) Total_Bilirubin : Total Billirubin in mg/dL. Billirubin is a yellowish pigment that is produced during the normal breakdown of red blood cells.

4) Direct_Bilirubin : Conjugated Billirubin in mg/dL

5) Alkaline_Phosphotase : Conjugated Billirubin in mg/dL

6) Alamine_Aminotransferase : ALT in IU/L

7) Aspartate_Aminotransferase : AST in IU/L

8) Total_Protiens : Total Proteins g/dL

9) Albumin : Albumin in g/dL

10) Albumin_and_Globulin_Ratio: A/G ratio

Aims And Objectives

Aim: **To check which are the personal key indicators of Liver disease.**

Objectives:

1) To visualize the data with respect to different factors.

2) To check the relationship among the different personal key indicatorson liver disease by observing the correlation matrix.

3) To apply logistic regression model for checking the significant variables.

4) To find which model gives better accuracy by applying various machine learning techniques.

5) To handle the censored part and to visualize the survival curve of data by applying Kaplan-Meier analysis.

6) To compare survival curves between males and females by applying the log-rank test or the Wilcoxon test.

7) To find which variables affect most to the hazard rate of the model by using the cox proportionality hazards rate model.

Methodology

1.Logistic regression Model fitting:

In statistics, the (binary) logistic model is a statistical model that models the probability of one event (out of two alternatives) taking place by having the log-odds (the logarithm of the odds) for the event be a linear combination of one or more independent variables ("predictors").

In regression analysis, logistic regression is estimating the parameters of a logistic model (the coefficients in the linear combination). Formally, in binary logistic regression there is a single binary dependent variable, coded by an indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling;[2] the function that converts log-odds to probability is the logistic function, hence the name.

Binary variables are widely used in statistics to model the probability of a certain class or event taking place, such as the probability of a team winning, of a patient being healthy, etc.

Logistic regression is used in various fields, including machine learning, most medicalfields, and social sciences.

Logistic regression may be used to predict the risk of developing a given disease (e.g. diabetes; coronary heart disease), based on observed characteristics of the patient (age,sex, body mass index, results of various blood tests, etc.).

2.Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular toolin machine learning.

A decision tree is a flowchart-like structure in which each internal node represents a "test"on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used asa visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of three types of nodes:

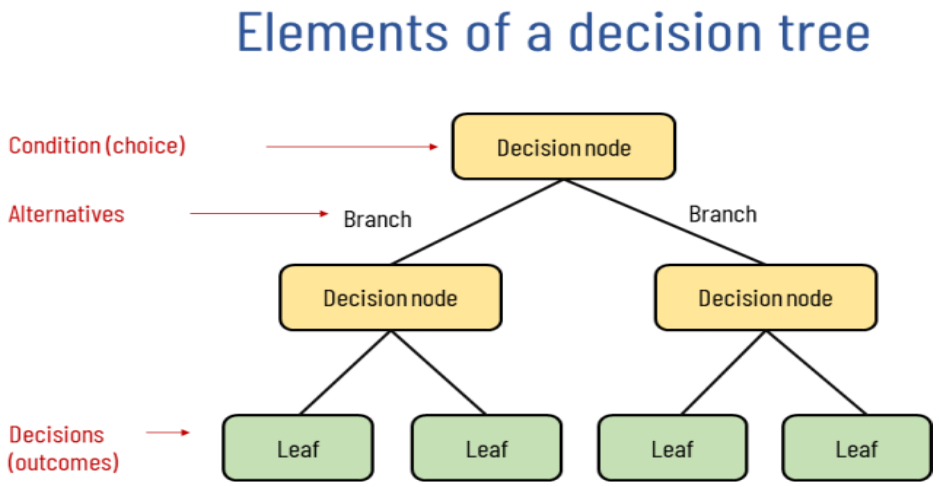
Decision nodes – typically represented by

squaresChance nodes – typically represented by

circles End nodes – typically represented by

triangles

It is important to know the measurements used to evaluate decision trees. The main metrics used are accuracy, sensitivity, specificity, precision, miss rate, false discovery rate, and false omission rate. All these measurements are derived from the number of true positives, false positives, true negatives, and false negatives obtained when running a set of samples through the decision tree classification model. Also, a confusion matrix can bemade to display these results.



3.K-Nearest Neighborhood

In statistics, the k-nearest neighbors algorithm (k-NN) is a non-parametric supervised learning method first developed by Evelyn Fix and Joseph Hodges in 1951,[1] and later expanded by Thomas Cover.[2] It is used for classification and regression. In both cases,the input consists of the k closest training examples in a data set. The output depends onwhether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most commonamong its k nearest neighbors (k is a positive integer, typically small). If k = 1, then theobject is simply assigned to the class of that single nearest neighbor.In k-NN regression, the output is the property value for the object. This value is theaverage of the values of k nearest neighbors.

K-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.

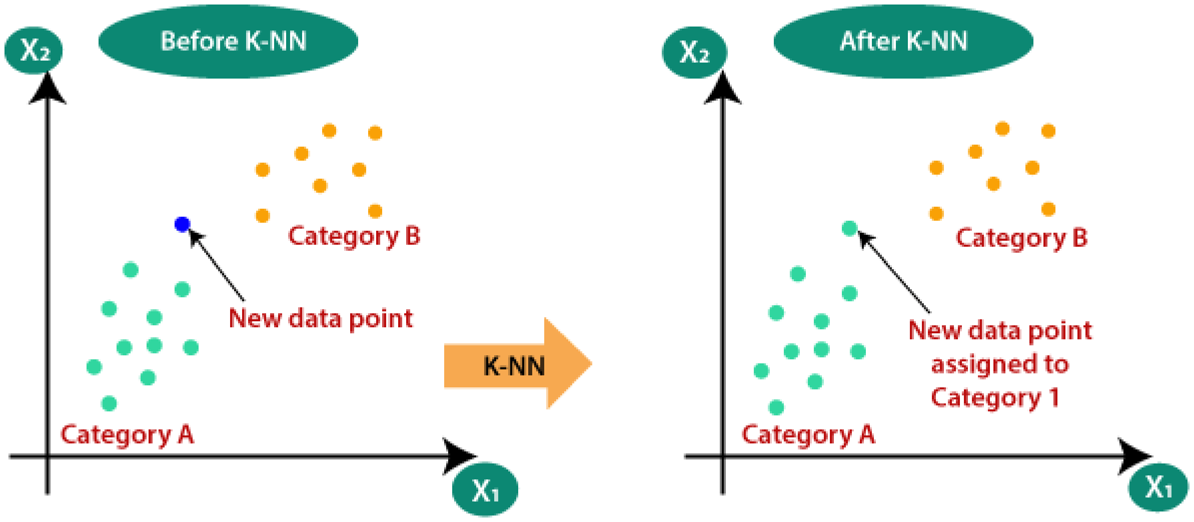
In k-NN regression, the k-NN algorithm[citation needed] is used for estimating continuous variables. One such algorithm uses a weighted average of the k nearest neighbors, weighted by the inverse of their distance. This algorithm works as follows:

Compute the Euclidean or Mahalanobis distance from the query example to the labeledexamples.

Order the labeled examples by increasing distance.

Find a heuristically optimal number k of nearest neighbors, based on RMSE. This is done using cross validation.

Calculate an inverse distance weighted average with the k-nearest multivariate neighbors.



4.Random Forest

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest isthe class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees [citation needed]. However, data characteristics can affect their performance.

Decision trees are a popular method for various machine learning tasks. Tree learning "come[s] closest to meeting the requirements for serving as an off-the-shelf procedure for data mining", say Hastie et al., "because it is invariant under scaling and various other transformations of feature values, is robust to inclusion of irrelevant features, and produces inspectable models. However, they are seldom accurate".[3]: 352
In particular, trees that are grown very deep tend to learn highly irregular patterns: they overfit their training sets, i.e. have low bias, but very high variance. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance.[3]: 587–588 This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.

A big part of machine learning is classification — we want to know what class (a.k.a. group) an observation belongs to. The ability to precisely classify observations is extremely valuable for various business applications like predicting whether a particular user will buy a product or forecasting whether a given loan will default or not.
The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model worksso well is:

The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don‘t constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are ableto move in the correct direction**.**

5.Survival Analysis

Survival analysis is a statistical methodology used to analyze the time until an event of interest occurs. It is widely used in various fields, including medicine, engineering, social sciences, and finance. The "event" in survival analysis can refer to various outcomes, such as death, failure, relapse, or any other event that can happen over time.

The primary goal of survival analysis is to estimate the survival function, which describes the probability that an event has not occurred by a certain time point. This function provides valuable information about the time it takes for an event to happen. In addition to estimating the survival function, survival analysis also involves comparing survival experiences between different groups or populations, identifying factors that influence survival times, and making predictions about future events.

Survival data typically consists of a time-to-event variable and a status indicator variable. The time-to-event variable represents the duration until the event of interest occurs, and the status indicator variable indicates whether the event has occurred or not. In some cases, the data may also include explanatory variables that are believed to affect the survival time.

Kaplan-Meier estimator is a commonly used non-parametric method in survival analysis to estimate the survival function when there are no covariates. It provides a step-by-step estimate of the survival function based on the observed survival times in the dataset. Additionally, parametric models such as the exponential, Weibull, and Cox proportional hazards models are used to make assumptions about the underlying distribution of survival times and estimate the survival function.

Survival analysis also introduces the concept of censoring, which occurs when the event of interest has not occurred for some individuals by the end of the study or they are lost to follow-up. Censoring is a crucial aspect of survival analysis and requires specific statistical techniques to handle appropriately.

Some commonly used methods in survival analysis include log-rank test, Cox proportional hazards regression, accelerated failure time models, and competing risks analysis, among others. These methods allow researchers to compare survival experiences between different groups, identify risk factors associated with the event of interest, and account for various factors that may affect survival times.

In conclusion, survival analysis provides a powerful framework for analyzing time-to-event data, estimating survival probabilities, and exploring factors that influence the occurrence of events. It has broad applications in various fields and offers valuable insights into understanding and predicting the time until specific events occur

1.The Kaplan-Meier Survival Curve

The Kaplan-Meier survival curve is a graphical representation of the survival probabilities over time in a survival analysis. It is commonly used to analyze and interpret time-to-event data, such as the survival time of patients, failure times of components, or any other event occurrence.

Here's how you can interpret the Kaplan-Meier survival curve:

1) Shape of the Curve: The shape of the curve provides information about the survival experience of the population under study. If the curve declines steeply at the beginning and then levels off, it indicates a high early mortality rate followed by a more stable survival rate. Conversely, a curve that decreases gradually suggests a lower initial mortality rate and a more stable survival rate over time.

2) Time Axis: The horizontal axis represents the time duration of the study. It is divided into intervals or time points where events (deaths, failures) occur or censored observations are present. The vertical axis represents the survival probability.

3) Drops in the Curve: Each drop in the curve represents an event (e.g., death) that occurred at that particular time point. These drops indicate a decrease in the survival probability due to an event.

4) Censored Data: If there are horizontal lines in the curve, it represents censored observations. Censoring occurs when the event of interest has not occurred for some individuals within the study period or they are lost to follow-up. Censored observations provide information about the individuals who were still under observation at the end of the study or who were lost to follow-up.

5) Survival Probability: The y-axis represents the survival probability, which is the probability that an individual or group will survive beyond a certain time point. The survival probability decreases over time as events occur. At any specific time point, the y-coordinate represents the estimated proportion of individuals who have not experienced the event (survived) up to that time.

6) Median Survival Time: The median survival time is the time point at which 50% of the individuals have experienced the event of interest. It can be determined by finding the time at which the survival probability is 0.5.

7) Comparisons between Groups: If there are multiple survival curves, it allows for comparisons between different groups (e.g., treatment vs. control). Differences in survival curves indicate potential differences in the survival experience between the groups being compared.

By analyzing the shape, drops, censored data, and survival probabilities in the Kaplan-

Meier survival curve, you can gain insights into the survival experience and make comparisons between different groups or populations.

2.The Log-Rank Test

The log-rank test is a statistical test commonly used in survival analysis to compare the survival distributions between two or more groups. It is particularly useful when analyzing time-to-event data, where the event of interest is typically death or failure.

Here are some reasons why the log-rank test is used in survival analysis:

1. Comparison of survival curves: The log-rank test allows for a formal comparison of survival curves between different groups. It assesses whether there are significant differences in the survival experience among the groups based on the observed events and censoring information.

2. Non-parametric approach: The log-rank test is a non-parametric test, meaning it does not make assumptions about the shape of the survival distributions or the underlying probability distributions. It is robust and widely applicable in various scenarios.

3. Censoring: In survival analysis, censoring occurs when the event of interest has not occurred for some individuals within the study period or they are lost to follow-up. The log-rank test takes into account censored observations and adjusts the test statistic accordingly, providing valid results even in the presence of censoring.

4. Hypothesis testing: The log-rank test allows researchers to test the null hypothesis that there are no differences in survival between the groups being compared. If the p-value associated with the log-rank test is below a predefined significance level (e.g., 0.05), it suggests that there are significant differences in survival between the groups.

5. Widely used and accepted: The log-rank test is a widely accepted method in survival analysis and has been extensively used in medical and epidemiological research. It is also implemented in various statistical software packages, making it readily accessible and easy to apply.

It is important to note that the log-rank test assumes proportional hazards, meaning that the hazard ratio comparing the groups being compared remains constant over time. If this assumption is violated, alternative methods such as the stratified log-rank test or Cox proportional hazards regression may be more appropriate.

3.The Cox-Proportionality Hazard rate model

The Cox proportional hazards model does not rely on specific hypotheses in the same way as traditional hypothesis tests. Instead, it assumes a proportional hazards assumption, which is an underlying assumption of the model. However, it is common to formulate hypotheses about the individual regression coefficients in the Cox model.

In the Cox proportional hazards model, the hypothesis typically focuses on the relationship between a specific covariate (independent variable) and the hazard rate (risk of event occurrence). The null and alternative hypotheses can be stated as follows:

Null Hypothesis (H_0): The specific covariate has no effect on the hazard rate. The regression coefficient associated with the covariate is equal to zero.

Vs

Alternative Hypothesis (H_A): The specific covariate has an effect on the hazard rate. The regression coefficient associated with the covariate is not equal to zero.

By testing these hypotheses, we can determine whether a particular covariate has a statistically significant impact on the hazard rate, and thus plays a role in predicting the occurrence of the event of interest.

To test these hypotheses, statistical methods such as Wald tests, likelihood ratio tests, or score tests can be used. These tests assess the significance of the coefficient estimate by comparing it to the null hypothesis of no effect.

It's important to note that the Cox proportional hazards model assumes that the hazard ratios remain constant over time, implying that the covariate effects are proportional. Violation of this assumption may affect the validity of the results, and alternative models or methods may be more appropriate in such cases.

1. coef: The coefficient represents the estimated effect of the variable on the hazard rate. For example, an age coefficient of 0.030 indicates that, on average, for each additional unit of age, the hazard rate (risk of event) increases by 0.030.

2. exp(coef): The exponentiated coefficient represents the hazard ratio. It is the multiplicative factor by which the hazard rate changes for a one-unit increase in the corresponding variable. For instance, an exp(coef) value of 1.030 for age implies that the hazard rate increases by approximately 3% for each additional year of age.

3. se(coef): The standard error represents the variability or uncertainty in the estimation of the coefficient. A smaller standard error indicates a more precise estimate.

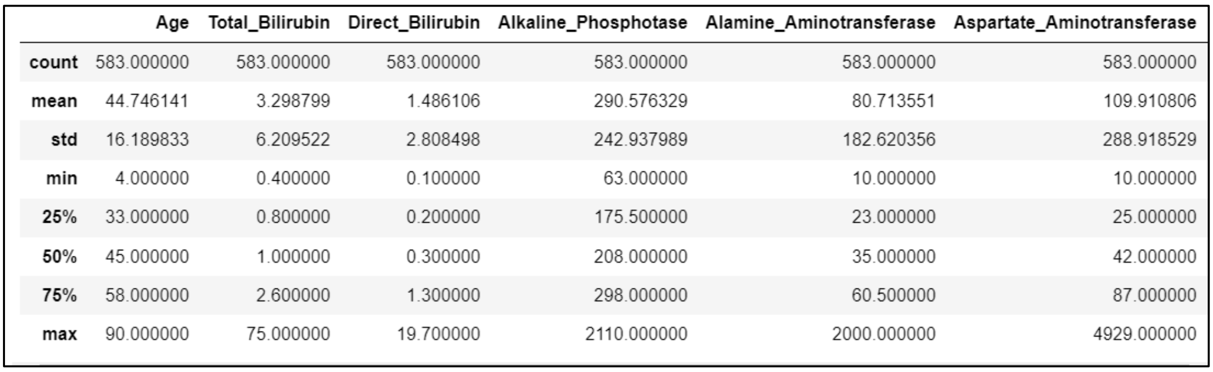
4. z: The z-value is the coefficient divided by its standard error. It measures the number of standard deviations the estimated coefficient is away from zero. A larger absolute z-value indicates a more significant association between the variable and the hazard rate.

5. p: The p-value represents the statistical significance of the variable's association with the hazard rate. It indicates the probability of observing such an extreme or more extreme effect if the null hypothesis (no association) were true. A smaller p-value (e.g., < 0.05) suggests that the variable is significantly associated with the hazard rate.

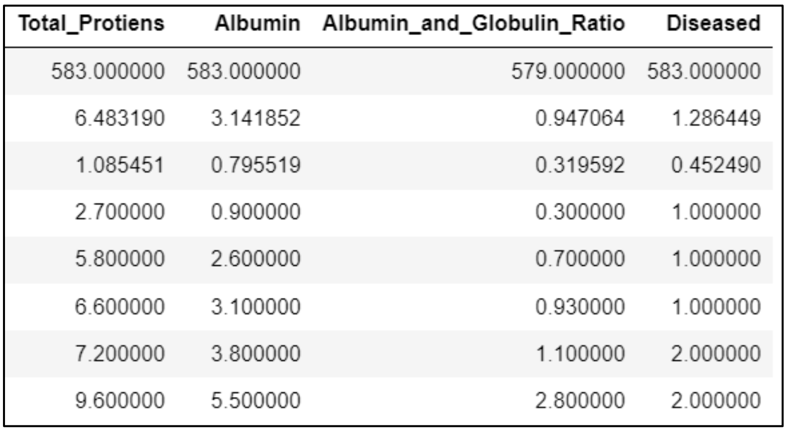
6. -log2(p): The negative log base 2 of the p-value, also known as the log-likelihood ratio statistic, provides a measure of the strength of evidence against the null hypothesis. A larger absolute -log2(p) value suggests stronger evidence against the null hypothesis.

Interpreting the summary output involves considering the sign, magnitude, statistical significance (p-value), and direction (hazard ratio) of the coefficients. It allows you to assess the impact of each variable on the hazard rate and identify significant predictors in the Cox proportional hazards model

Descriptive Statistics



	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase
count	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
std	16.189833	6.209522	2.808498	242.937989	182.620356	288.918529
min	4.000000	0.400000	0.100000	63.000000	10.000000	10.000000
25%	33.000000	0.800000	0.200000	175.500000	23.000000	25.000000
50%	45.000000	1.000000	0.300000	208.000000	35.000000	42.000000
75%	58.000000	2.600000	1.300000	298.000000	60.500000	87.000000
max	90.000000	75.000000	19.700000	2110.000000	2000.000000	4929.000000



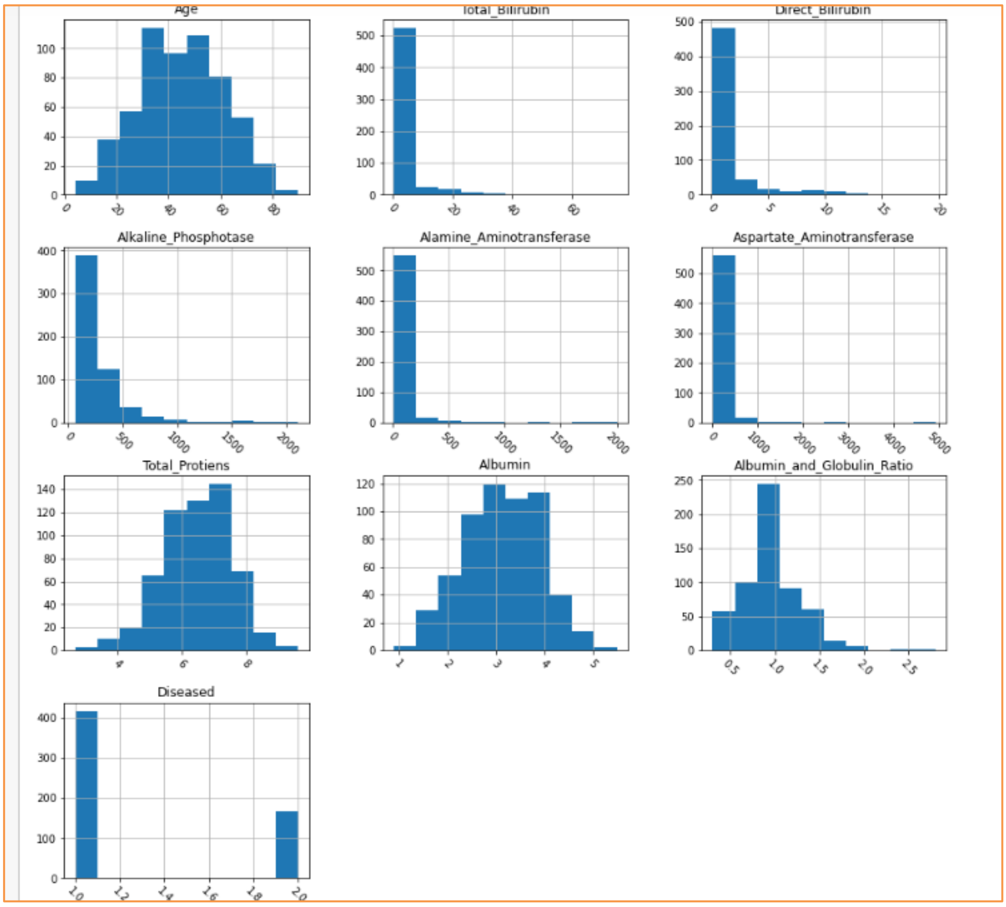
Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Diseased
583.000000	583.000000	579.000000	583.000000
6.483190	3.141852	0.947064	1.286449
1.085451	0.795519	0.319592	0.452490
2.700000	0.900000	0.300000	1.000000
5.800000	2.600000	0.700000	1.000000
6.600000	3.100000	0.930000	1.000000
7.200000	3.800000	1.100000	2.000000
9.600000	5.500000	2.800000	2.000000

Conclusion:

1. It seems that Aspartate_Aminotransferase is outlier because it has very high value of max value compere to other mean values
2. Output value has 1 for liver disease and 0 for no liver disease so let make it 0 for no disease for our convenient.

Data Visualizaton

1.Counts of Different Features with respect to Liver disease:



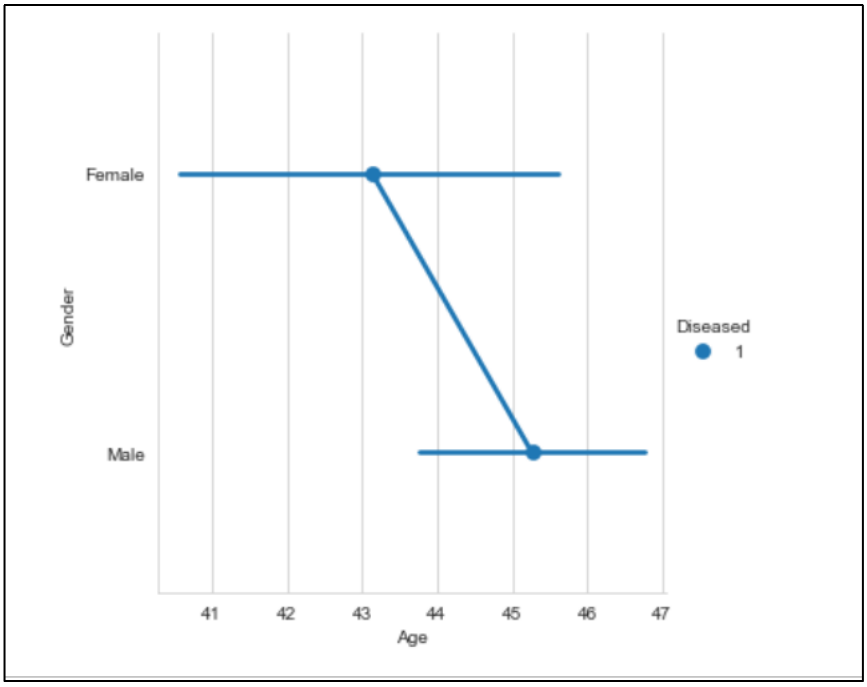
Conclusion:

1.We can clearly see in the output of the graph that, it is an imbalanced dataset, any patients diagnosed with liver disease are higher compared to the ones who are not diagnosed.

2.Age is normally distributed. i.e. At any certain age the disease might be occurred.

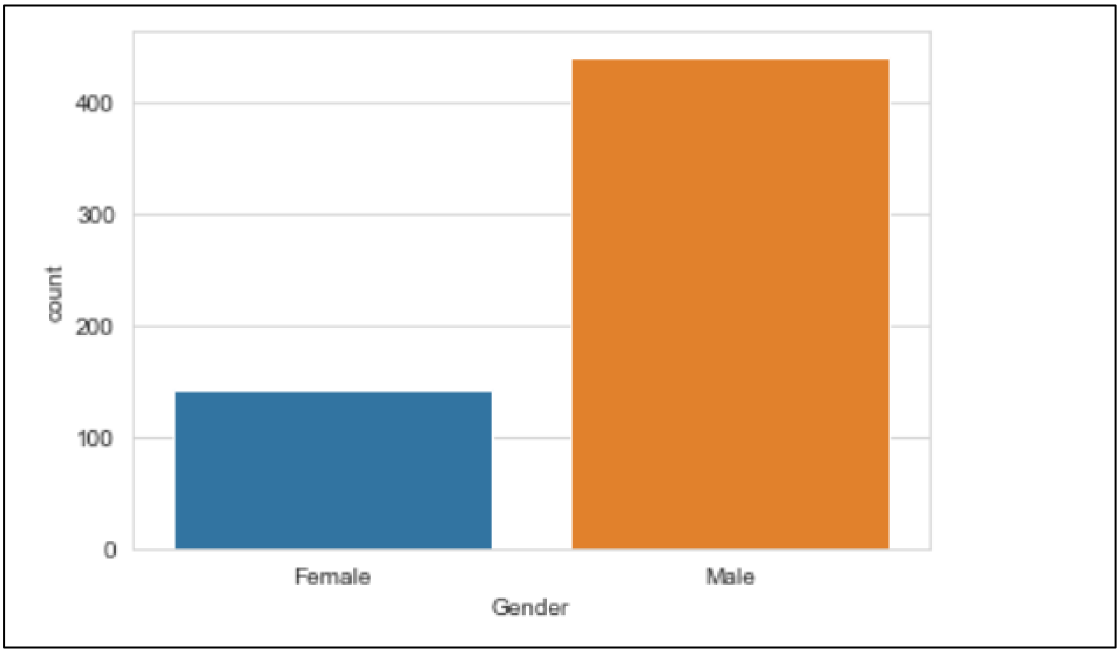
3. the amount of Total_protien and amount of Albumin are normally distributed means that no dependency of this factor on event of diseses.i.e at Any amount of Total_protien and amount of Albumin the disease might be occurred

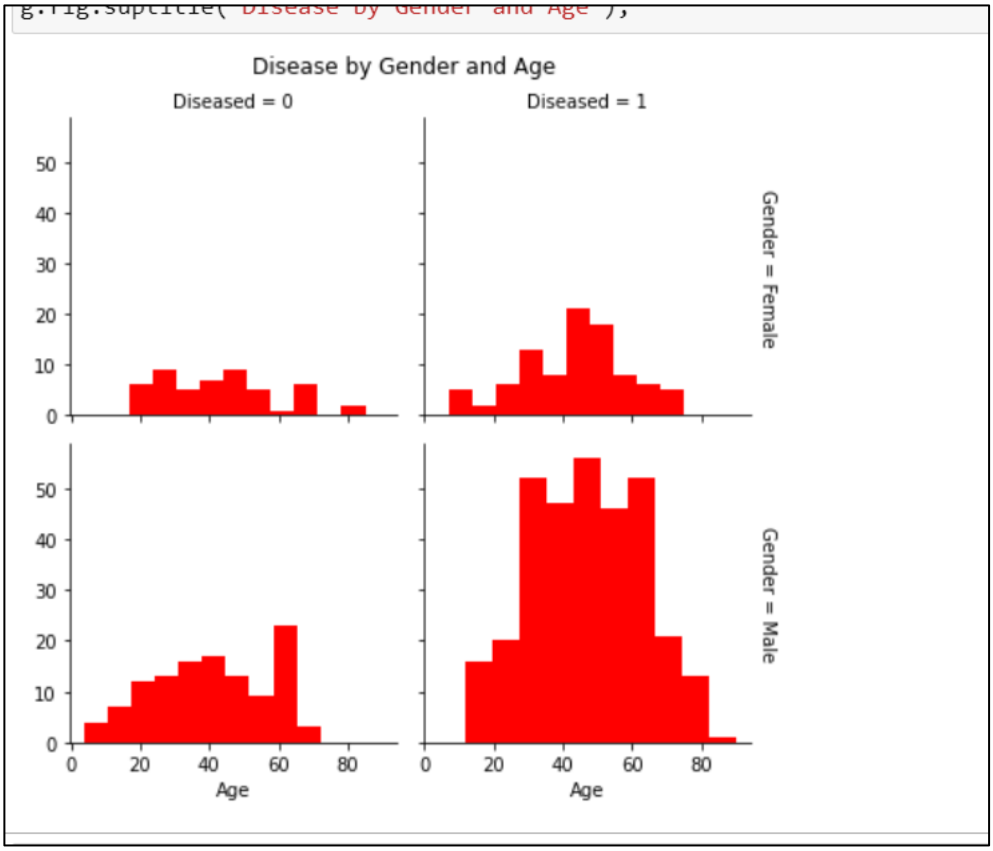
2.Factor Plot For Categorical Variables:



Conclusion: It seems to be a factor for liver disease for both male and female gender.

3.Sex Wise Distribution (with respect to liver Disease)



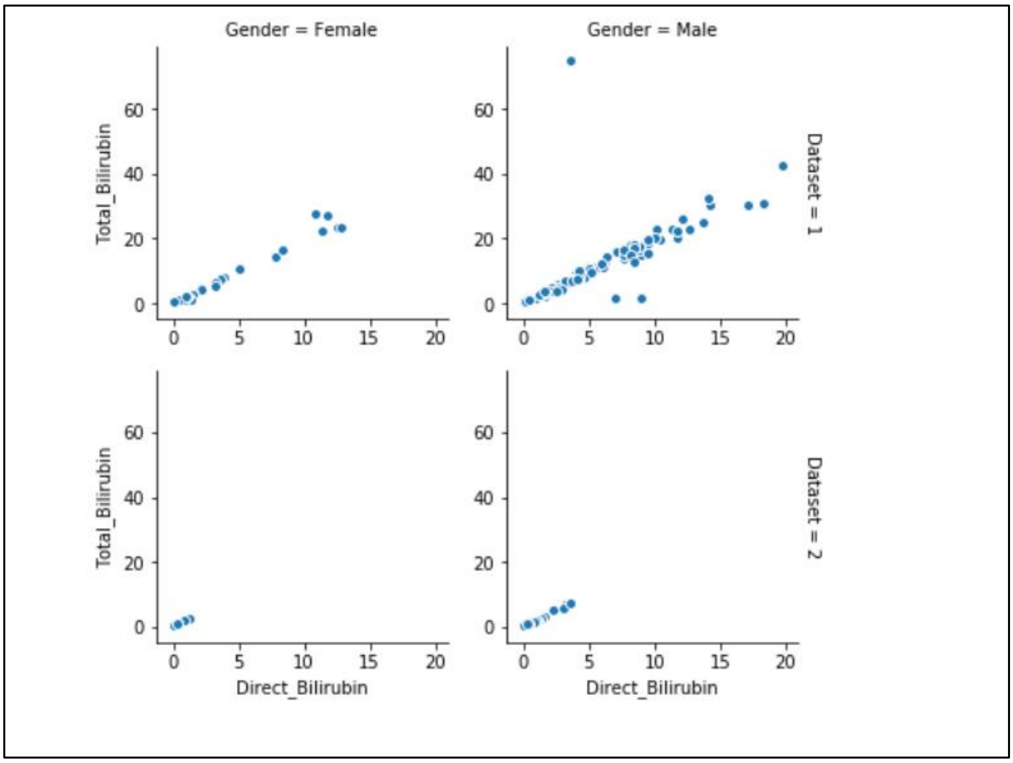


```
g.fig.suptitle('Disease by Gender and Age');
```

Conclusion: We can clearly see in both the graph that, number of patient suffering from liver disease are higher in males than in females.

4.Graphical Relationship between different features with respect to Liver disease

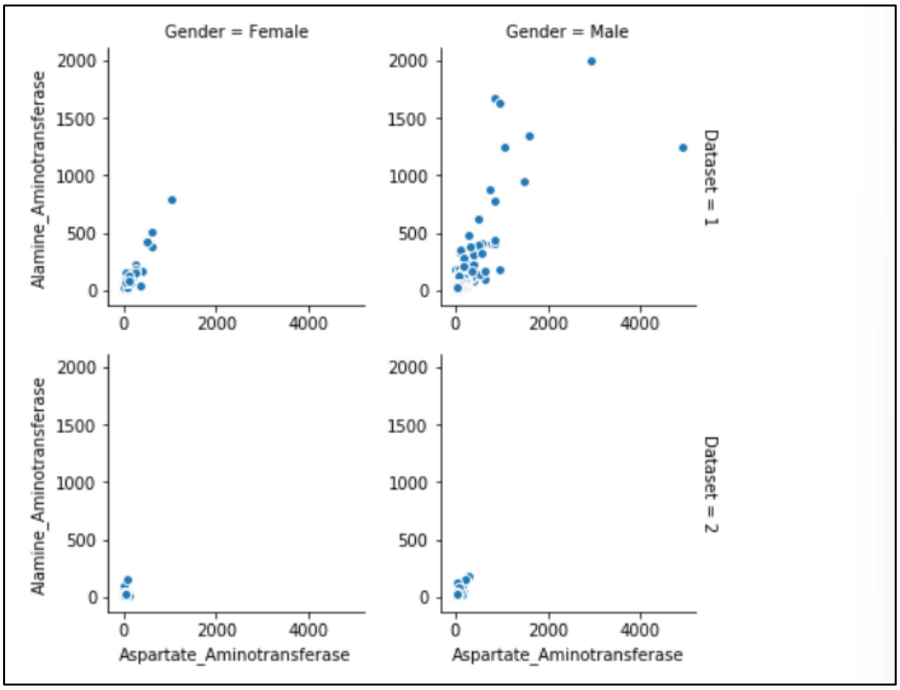
1.Rleation between Total_Bilirubin and Direct_Bilirubin:



Conclusion:

There seems to be direct relationship between Total_Bilirubin and Direct_Bilirubin. We have the possibility of removing one of this feature.

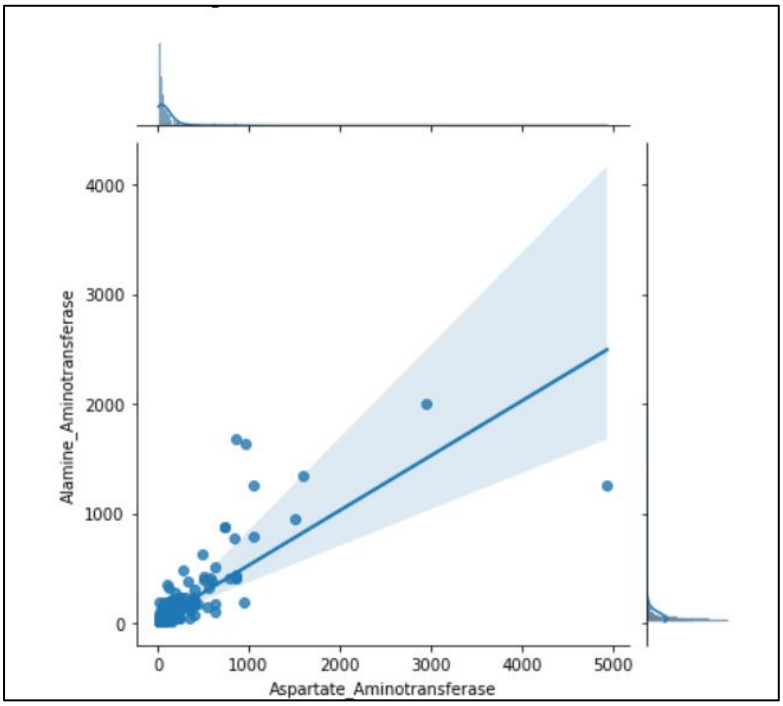
2.Rleation between Aspartate_Aminotransferase and Alamine_Aminotransferase:



Conclusion:

There is linear relationship between Aspartate_Aminotransferase and Alamine_Aminotransferase and the gender. We have the possibility of removing one of this feature.

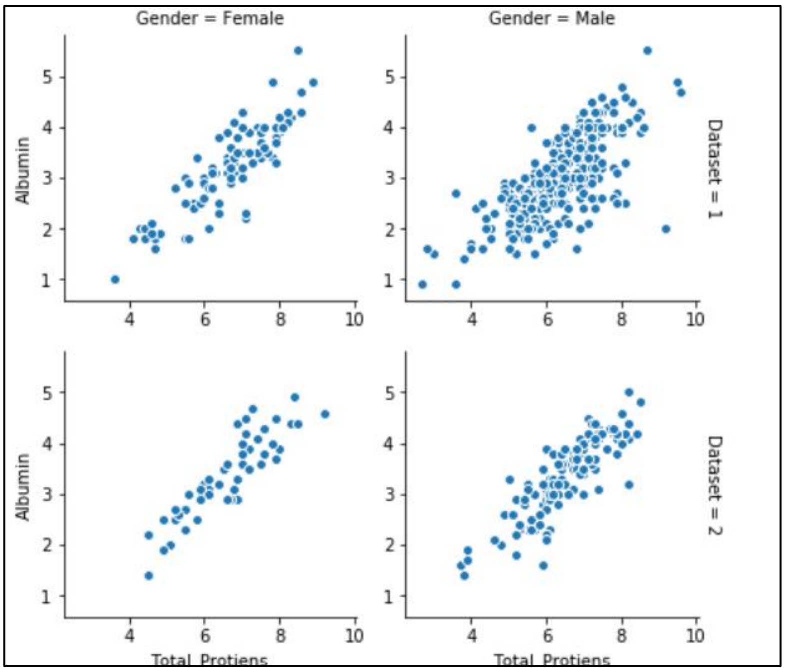
3. Relation between Alamine and Aspartate -Aminotransferase



Conclusion:

In this scatter plot we are plotting two significant features(**Alamine and Aspartate -Aminotransferase**) along with **Gender** as a form of hue and it clearly shows that males are highly effective concerning these two features the most.

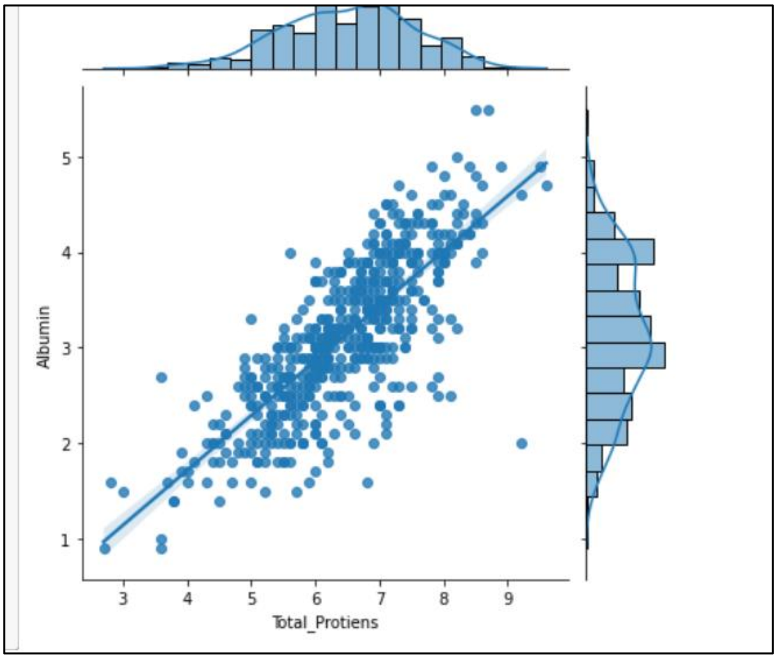
4.Relation between Albumin and Total_protiens



Conclusion:

There is linear relationship between Albumin and Total_protiens the gender. We have the possibility of removing one of this feature.

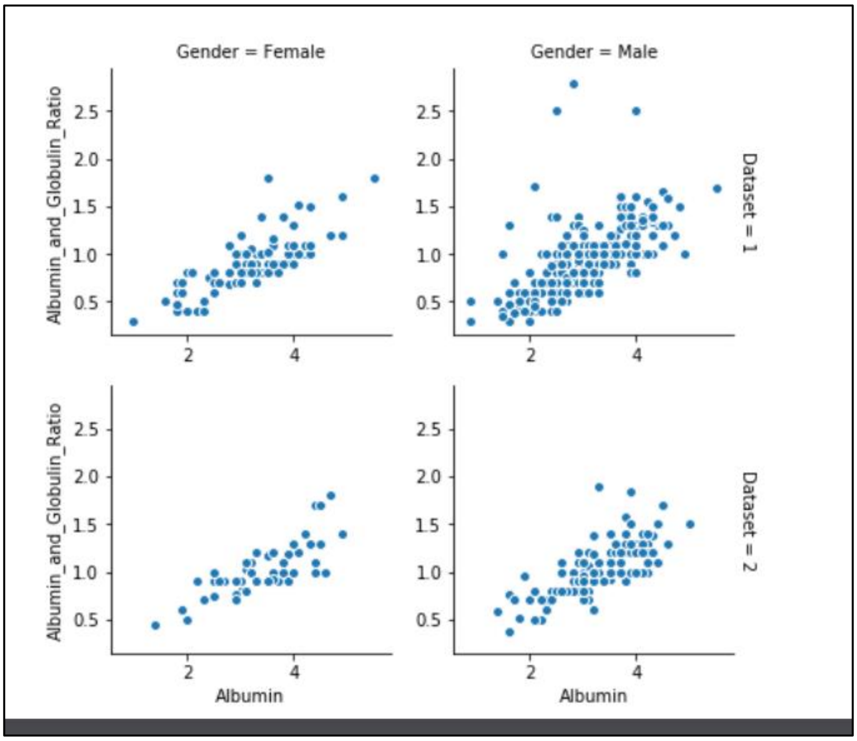
5.Relation Between **Total_protiens** and **Albumin**



Conclusion:

Now with the help of the above plot we can find out that, **Total_protiens** and **Albumin** features are in positive **regressive nature,** with some **outliers.**

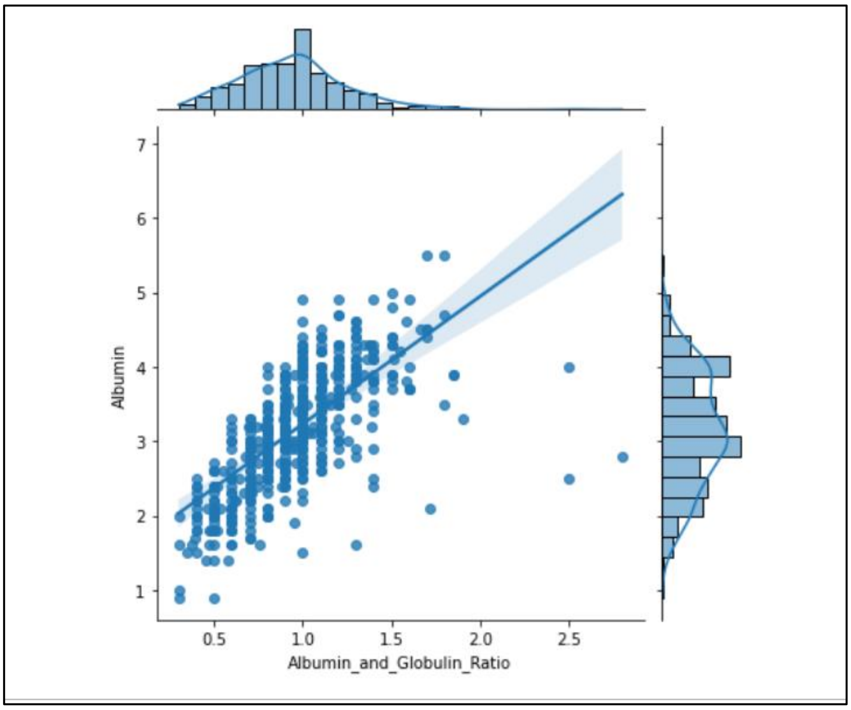
6.. **Relation between Alamine and Aspartate -Aminotransferase**



Conclusion:

There is linear relationship between Albumin_and_Globulin_Ratio and Albumin. We have the possibility of removing one of this feature.

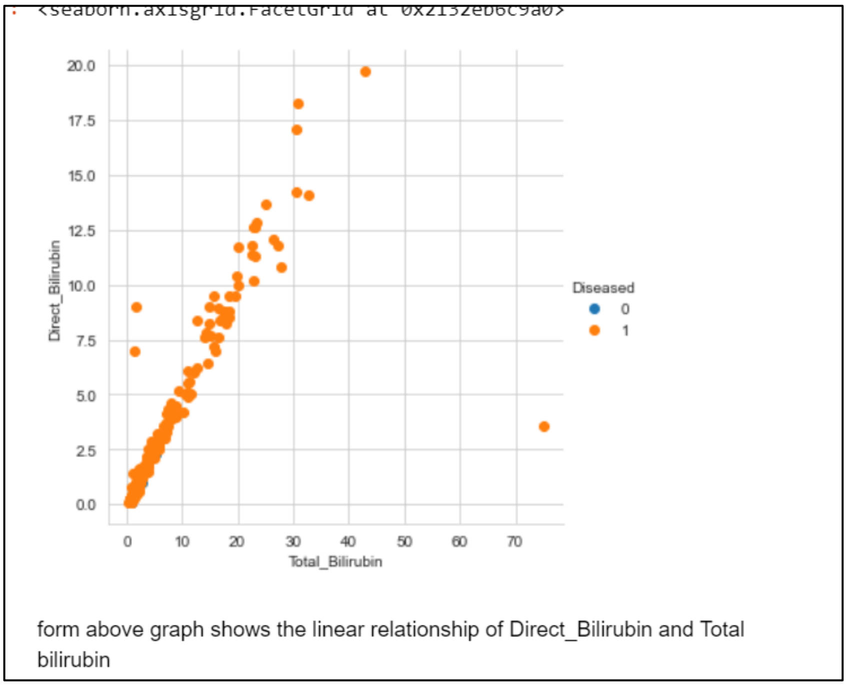
7.Relation between **Albumin** and **Albumin_and_Globulin_Rat**



Conclusion:

After plotting **Albumin** and **Albumin_and_Globulin_Ratio** we conclude that they both share **normal distribution** and have a direct relationship like some other features in the dataset. There is linear relationship between Albumin and globulin ratio and albumin .We have the possibility of removing one of this feature.

8.Relation Between **Total_bilirubin** vs **Direct_Bilrubin**



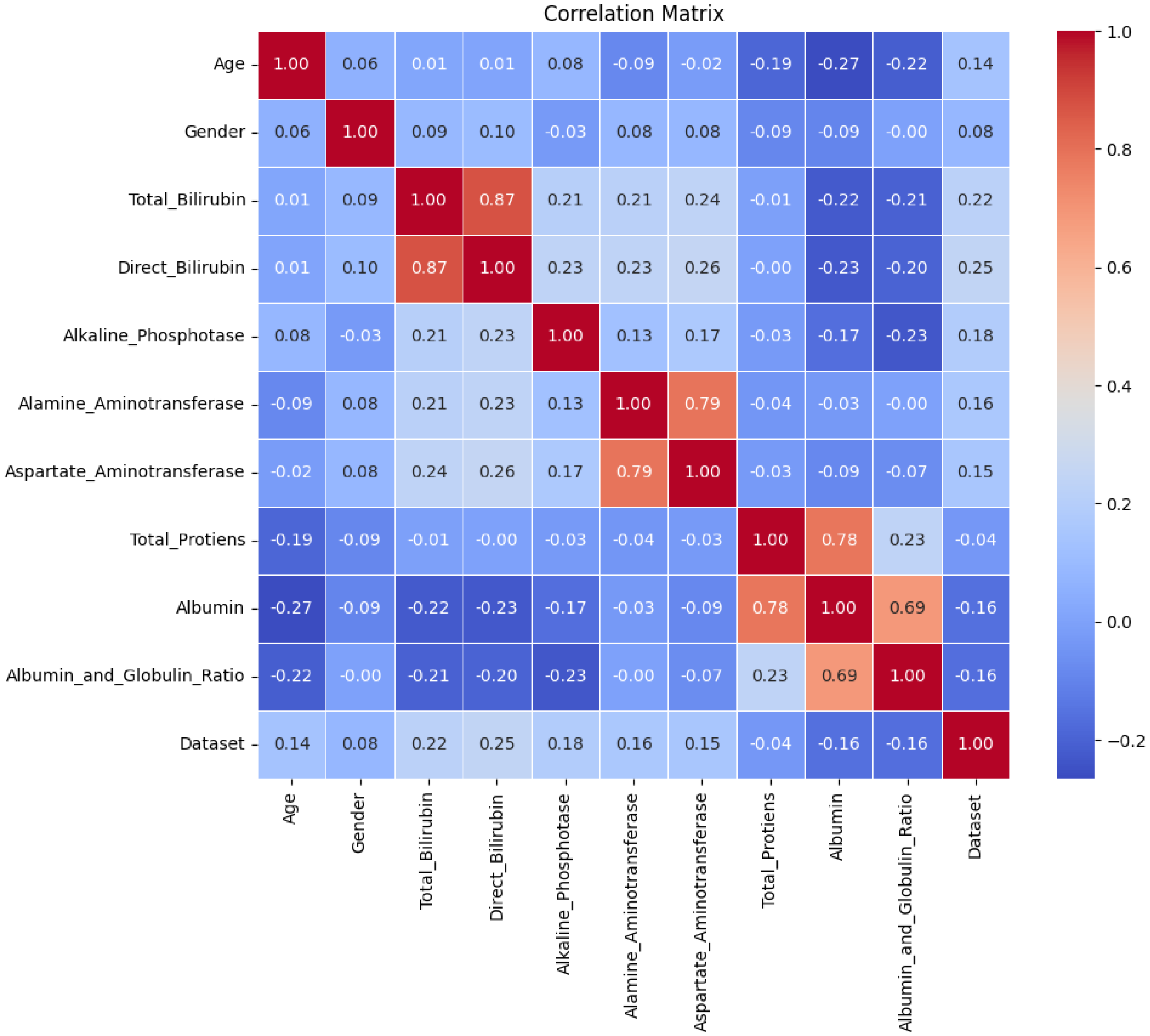
```
<seaborn.axisgrid.FacetGrid at 0x2132eb6c9a0>
```

form above graph shows the linear relationship of Direct_Bilirubin and Total bilirubin

Conclusion:

Here in this plotted **Total_bilirubin** vs **Direct_Bilrubin** and got the insight that both of the features have a **direct relationship** with each other.

Correlation matrix



Conculsion:

We can see that there are correlations between
“direct_bilirubin” and “total_bilirubin”(strong),
“aspartate_aminotransferase” and “alamine_aminotransferase” (a little strong),
“albumin” and “total_protiens” (a little strong),
“ratio_albumin_and_globulin_ratio” and “albumin” (a little strong).

- Corr(“direct_bilirubin” and “total_bilirubin”) = 0.874481
- Corr(“aspartate_aminotransferase” and “alamine_aminotransferase”) = 0.791862
- Corr(“albumin” and “total_protiens”) = 0.783112
- Corr(“ratio_albumin_and_globulin_ratio” and “albumin”) = 0.689632

Here we can observe that so many variables have high correlation with each other so it may affect the result of target variable.

Multicollinearity affects the coefficients and p-values, but it does not influence the predictions, precision of the predictions, and the goodness-of-fit statistics, that‘s why we headed towards machine learning techniques without removing the multicollinearity.

DATA ANALYSIS

Encoding:

Machine learning models require all input and output variables to be numeric.

This means that if your data contains categorical data, you must encode it to numbers before you can fit and evaluate a model.

The two most popular techniques are an **Ordinal Encoding** and a **One-Hot Encoding**.

Ordinal Encoding

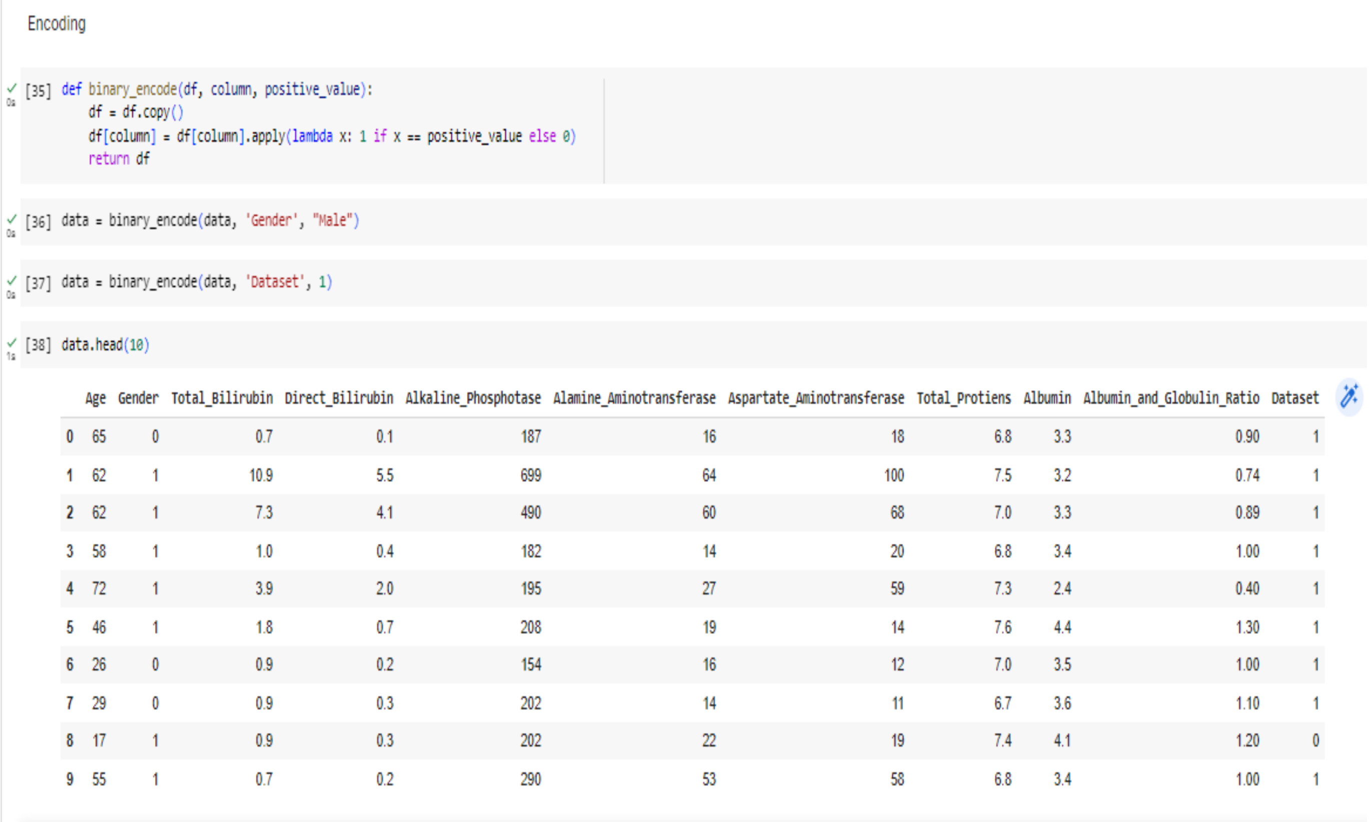
In ordinal encoding, each unique category value is assigned an integer value.

One-Hot encoding

A one-hot encoding can be applied to the ordinal representation. This is where the integer encoded variable is removed and one new binary variable is added for each unique integer value in the variable.

In our project we will apply the One-Hot encoding for further project data analysis.

One-Hot Encoding:



Encoding

```
[35] def binary_encode(df, column, positive_value):
         df = df.copy()
         df[column] = df[column].apply(lambda x: 1 if x == positive_value else 0)
         return df
```

```
[36] data = binary_encode(data, 'Gender', "Male")
```

```
[37] data = binary_encode(data, 'Dataset', 1)
```

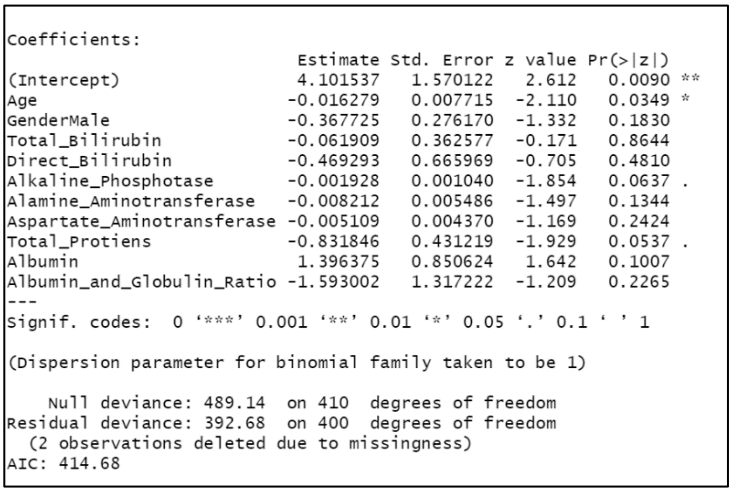
```
[38] data.head(10)
```

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

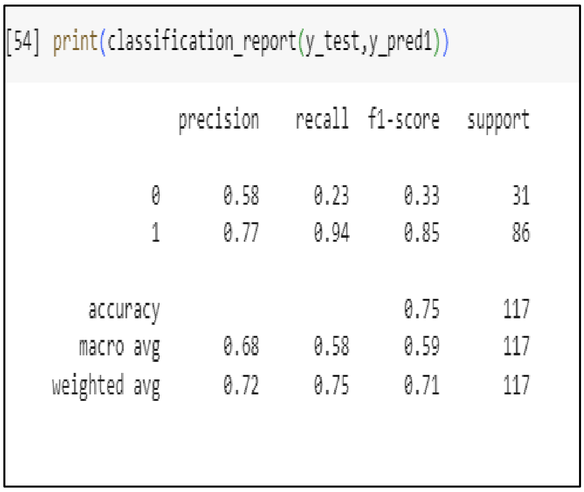
➢ Logistic Regression Analysis (in Python)(without outliers)

❖ Logistic regression equation:

y= 4.101537-0.016279*Age-0.367725*Gender-0.061909*Total_Bilirubin-0.469293*Direct_Bilirubin-0.001928*Alkaline_Phosphotase -0.008212*Alamine_Aminotransferase-0.005109*Aspartate_Aminotransferase-0.831846*Total_Protiens+1.396375*Albumin+1.593002*Albumin_and_Globulin_Ratio



```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                4.101537   1.570122   2.612   0.0090 **
Age                       -0.016279   0.007715  -2.110   0.0349 *
GenderMale                -0.367725   0.276170  -1.332   0.1830
Total_Bilirubin           -0.061909   0.362577  -0.171   0.8644
Direct_Bilirubin          -0.469293   0.665969  -0.705   0.4810
Alkaline_Phosphotase      -0.001928   0.001040  -1.854   0.0637 .
Alamine_Aminotransferase  -0.008212   0.005486  -1.497   0.1344
Aspartate_Aminotransferase -0.005109   0.004370  -1.169   0.2424
Total_Protiens            -0.831846   0.431219  -1.929   0.0537 .
Albumin                    1.396375   0.850624   1.642   0.1007
Albumin_and_Globulin_Ratio -1.593002   1.317222  -1.209   0.2265
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 489.14  on 410  degrees of freedom
Residual deviance: 392.68  on 400  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 414.68
```

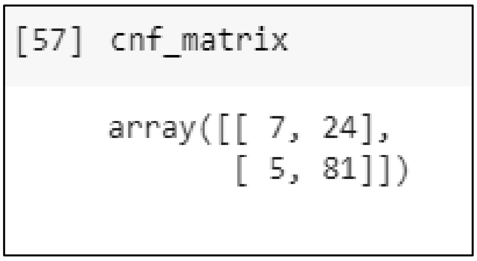


```
[54] print(classification_report(y_test,y_pred1))

              precision    recall  f1-score   support

           0       0.58      0.23      0.33        31
           1       0.77      0.94      0.85        86

    accuracy                           0.75       117
   macro avg       0.68      0.58      0.59       117
weighted avg       0.72      0.75      0.71       117
```

Conclusion: only Amine is the factor which shows positive correlation with the presence or absence of the disease. Hence as amount of amine in the body increases the chance if getting the liver disease is also increases.

From the above table, we can see that the accuracy of our logistic regression model is about 75%. It means that by using the logistic regression model we can predict 75 correct results out of 100 cases.

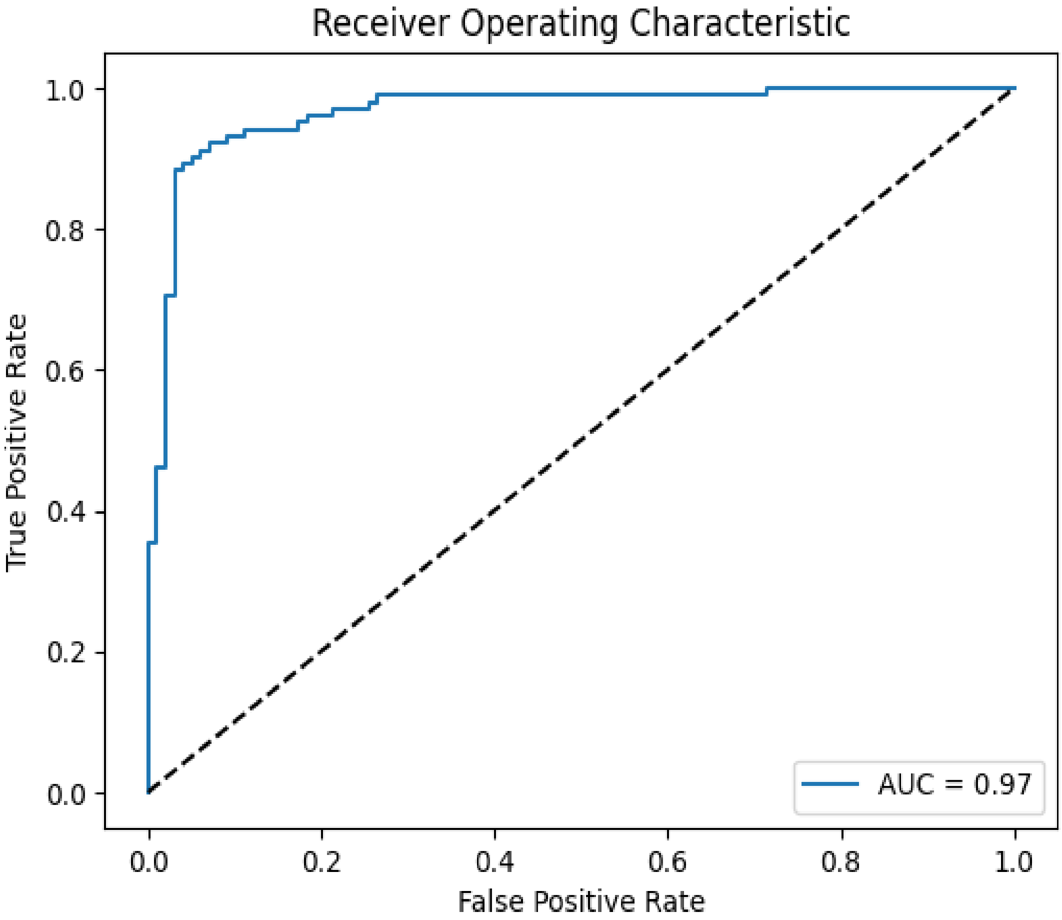
➢ Confusion Matrix



```
[57] cnf_matrix

     array([[ 7, 24],
            [ 5, 81]])
```

ROC curve and AUC curve in Logistic regression model

Logistic Regression is a statistical method that we use to fit a regression model when the response variable is binary. To assess how well a logistic regression model fits a dataset, we can look at the following two metrics:
Sensitivity: The probability that the model predicts a positive outcome for an observation when indeed the outcome is positive. This is also called the ―true positive rate.‖ Specificity: The probability that the model predicts a negative outcome for an observation when indeed the outcome is negative. This is also called the ―true negative rate.‖ One way to visualize these two metrics is by creating a ROC curve, which stands for ―receiver operating characteristic‖ curve. This is a plot that displays the sensitivity and specificity of a logistic regression model.



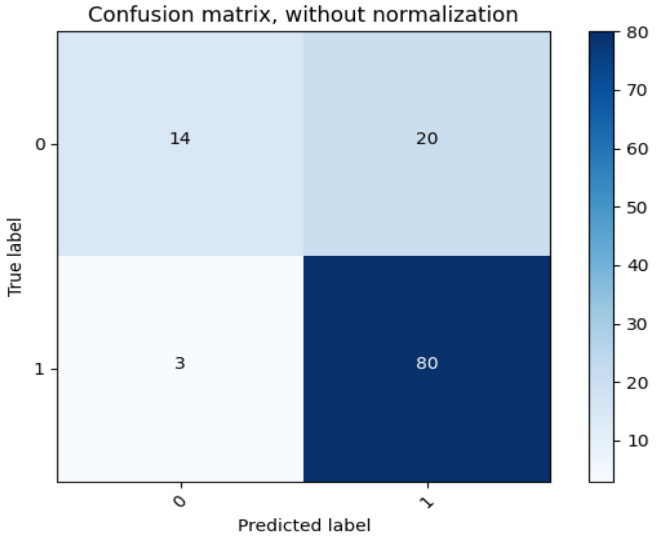
Conclusion:

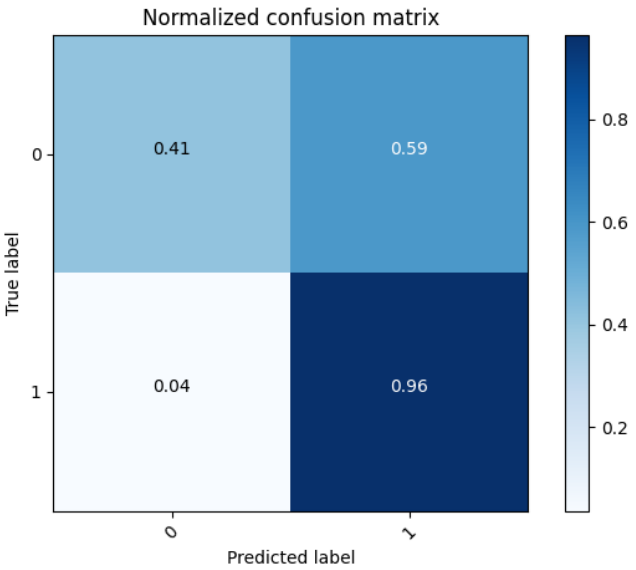
The more that the curve hugs the top left corner of the plot, the better the model does at classifying the data into categories

The closer AUC is to 1, the better the model. We have developed a model with an AUC equal to 0.97 is better than a model that makes random classifications.

➢ Logistic Regression Analysis (In Python) (Original Data):

```
Confusion matrix, without normalization
[[14 20]
 [ 3 80]]
Normalized confusion matrix
[[0.41 0.59]
[0.04 0.96]]
```

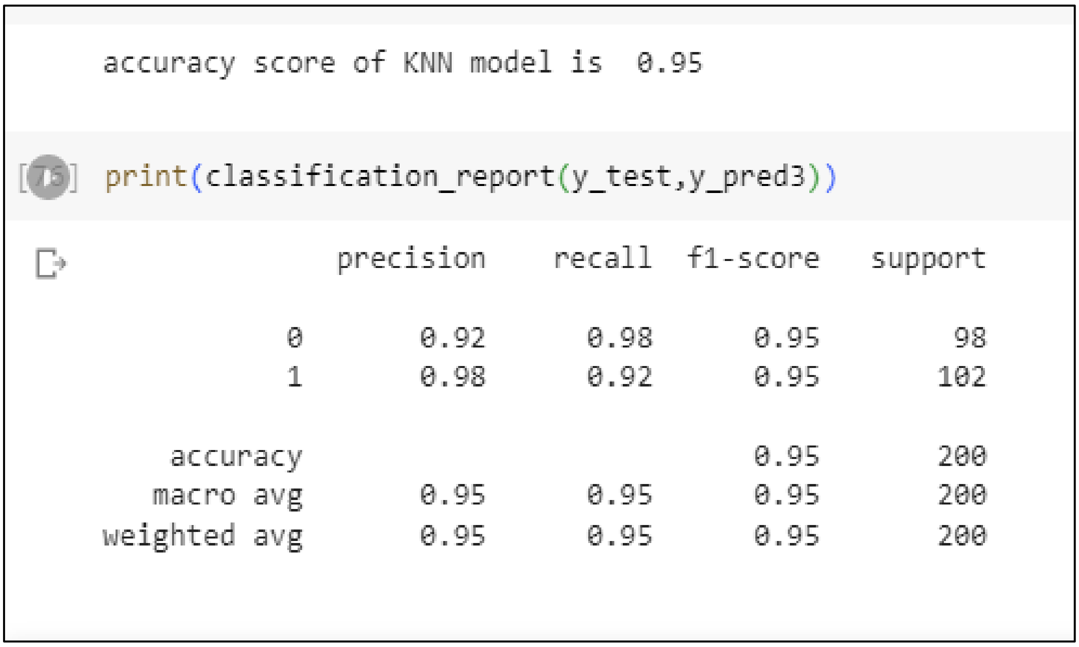

Confusion matrix, without normalization


Normalized confusion matrix

Conclusion:

1) Accuracy score of logistic regression model is **75%**.
2) Here True negative rate is 96% which means logistic regression model has predicted 96% of the data correctly.

➢ K – Nearest Neighbors Algorithm:



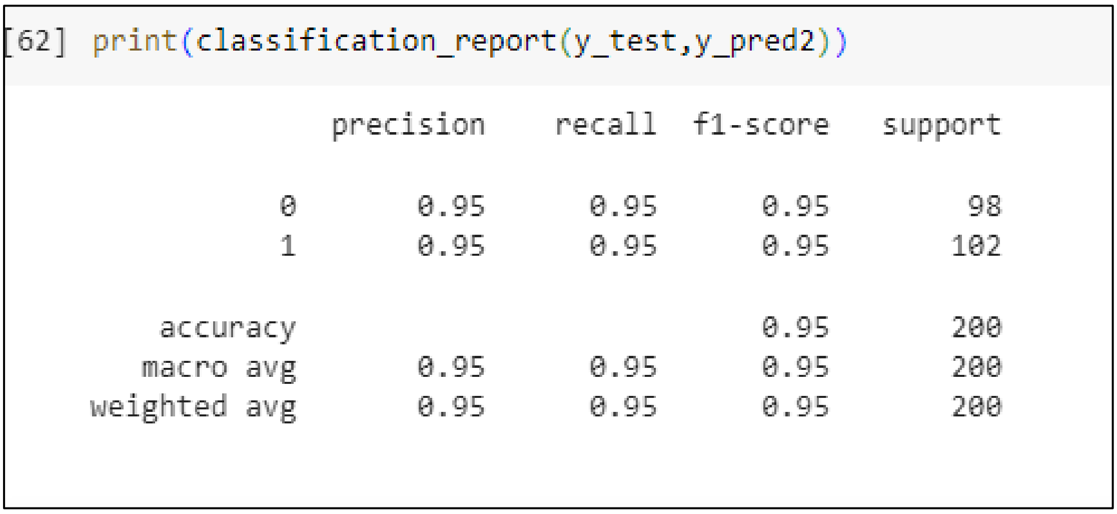
```
accuracy score of KNN model is  0.95
```

```
[75] print(classification_report(y_test,y_pred3))
```

```
              precision    recall  f1-score   support

           0       0.92      0.98      0.95        98
           1       0.98      0.92      0.95       102

    accuracy                           0.95       200
   macro avg       0.95      0.95      0.95       200
weighted avg       0.95      0.95      0.95       200
```

Conclusion: From the above table, we can see that the accuracy of our logistic regression model is about 95%. It means that by using the logistic regression model we can predict 95 correct results out of 100 cases.

➢ Decision Tree Algorithm :

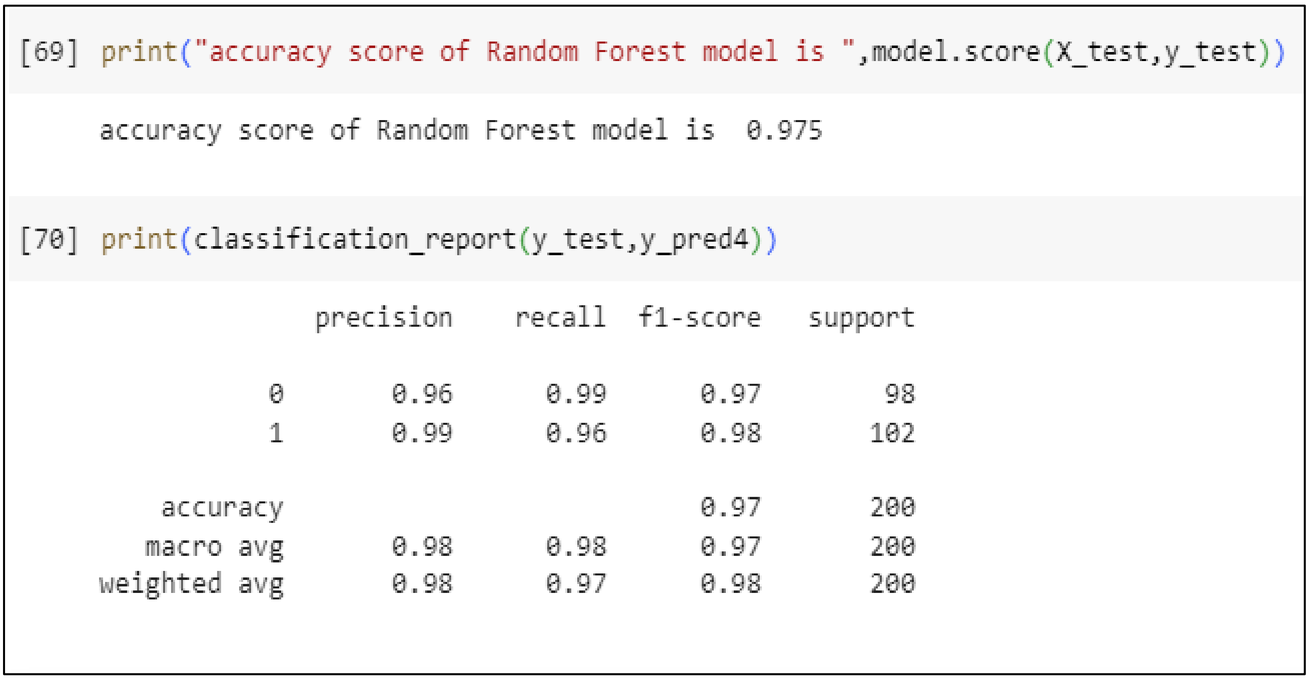


```
[62] print(classification_report(y_test,y_pred2))

              precision    recall  f1-score   support

           0       0.95      0.95      0.95        98
           1       0.95      0.95      0.95       102

    accuracy                           0.95       200
   macro avg       0.95      0.95      0.95       200
weighted avg       0.95      0.95      0.95       200
```

Conclusion: From the above table, we can see that the accuracy of our logistic regression model is about 95%. It means that by using the logistic regression model we can predict 95 correct results out of 100 cases.

Note: A high accuracy indicates that the decision tree classifier is performing well in terms of correctly predicting the class labels for the instances in the dataset. For example, an accuracy of 0.95 means that the model correctly predicts the class labels for 95% of the instances.

➢ Random Forest Algorithm:



```
[69] print("accuracy score of Random Forest model is ",model.score(X_test,y_test))

     accuracy score of Random Forest model is  0.975
```

```
[70] print(classification_report(y_test,y_pred4))

                   precision    recall  f1-score   support

                0       0.96      0.99      0.97        98
                1       0.99      0.96      0.98       102

         accuracy                           0.97       200
        macro avg       0.98      0.98      0.97       200
     weighted avg       0.98      0.97      0.98       200
```

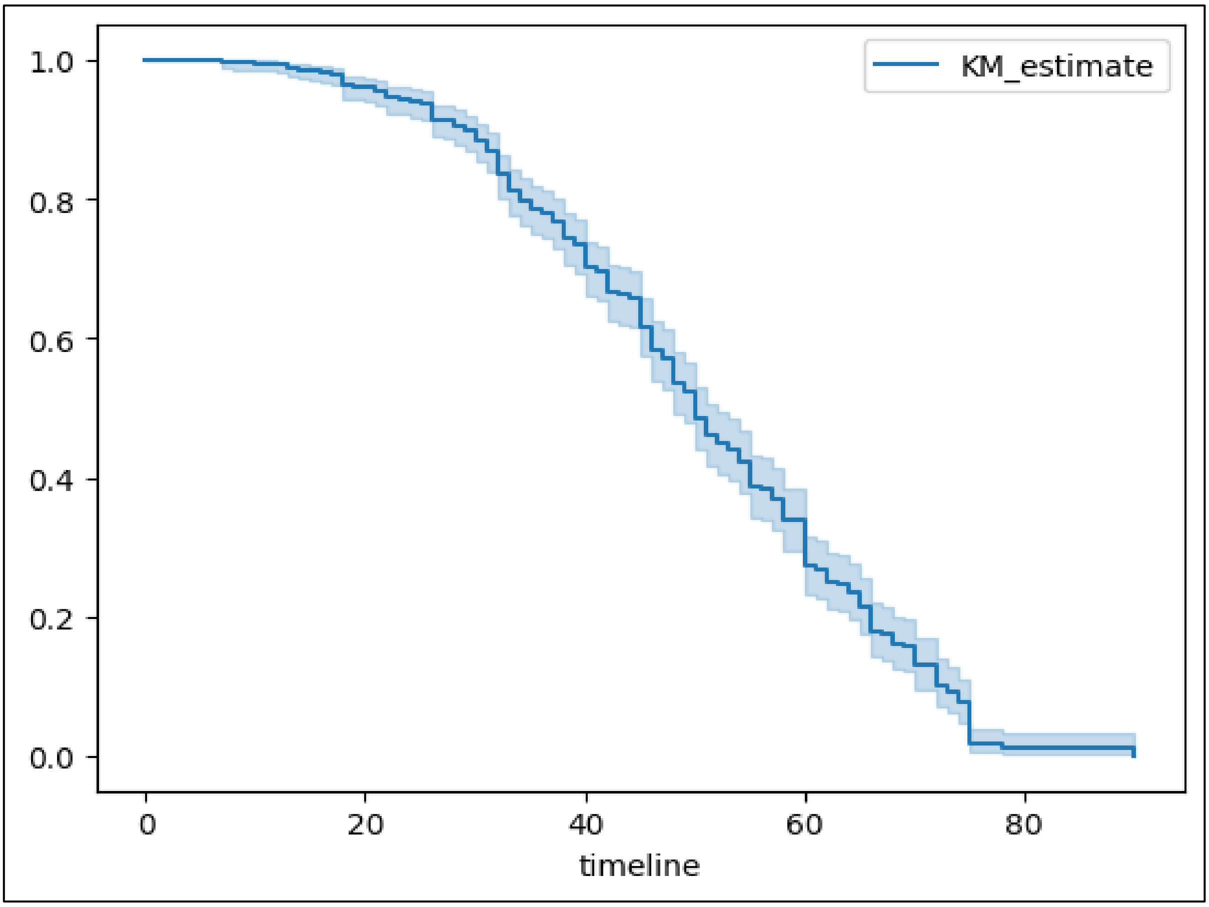
Conclusion: From the above table, we can see that the accuracy of our logistic regression model is about 97%. It means that by using the logistic regression model we can predict 97 correct results out of 100 cases.

Note: accuracy provides a general measure of the performance of a random forest classifier, it should be interpreted in conjunction with other metrics and domain-specific considerations. Understanding the limitations of accuracy and considering additional evaluation metrics will help provide a more robust assessment of the model's performance.

➢ Kaplan – Meier Survival Curve:

Following graph is the curve of the Kaplan-Meier curve which shows the survival distribution of the test. Here we take the age as a time variable and the dataset (i.e. is the liver disease present or not) as a indicator random variable.

For our data the function “lifelines.KaplanMeierFitter:"KM_estimate", fitted with 583 total observations out of which 167 observations were right-censored observations.



Conclusion: From the above graph it is clear that the chances of survival of liver patients decrease as age of the person increases.

➢ Log – Rank Test:

In this test we have divided the entire dataset in two independent groups. One group consisting of all the males and the other group contains all females.

According to the groups we tested the following hypothesis:

❖ Testing of Hypothesis :-

H_0 : The survival distribution of males and females is identical.

Versus,

H_1 : The survival distribution of males and females is different.

❖ Test statistic and p-value :-

The test statistic using the log-rank test is 0.04945.

The p-value using the log-rank test is 0.8240.

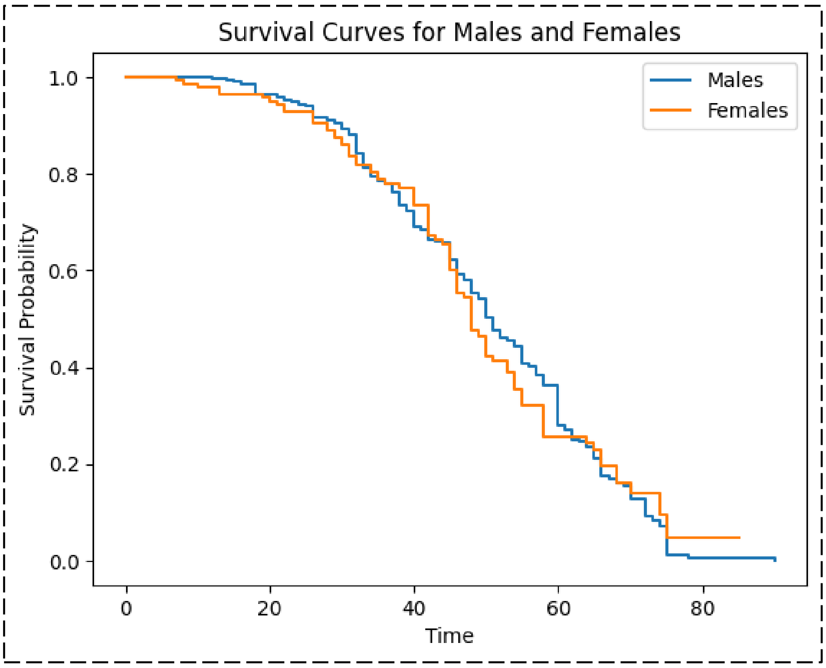
❖ Conclusion :-

Since the p-value is greater than 0.05 level of significance. We have enough evidence in support of the null hypothesis. So we will accept the null hypothesis.

❖ Interpretation :-

The survival distribution of both the groups of males and the females is identical.

❖ Visualization of the survival curves of two groups :-



➢ Cox Proportionality Hazard rate Model :-

In the cox proportionality hazard rate model we check whether which variable affect the target variable more as compared to the other variables.

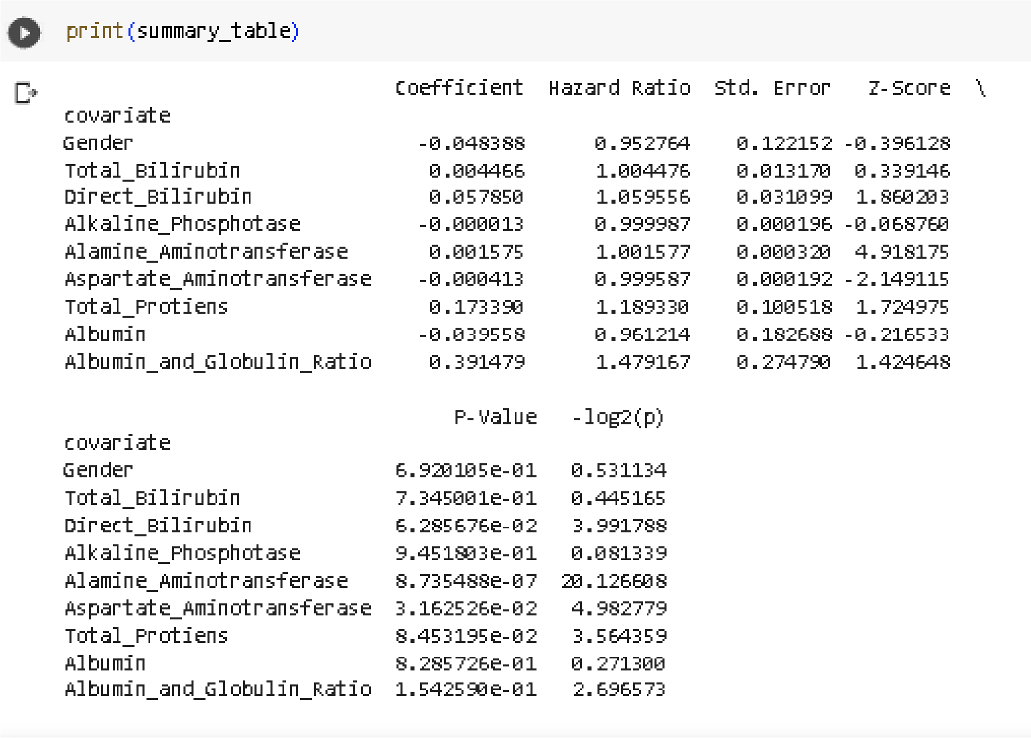
➢ Testing of Hypothesis :-

H_0 : The specific covariate(variable) has no effect on the hazard rate. The regression coefficient associated with the covariate is equal to zero.

Versus,

H_1 : The specific covariate(variable) has an effect on the hazard rate. The regression coefficient associated with the covariate is not equal to zero.

➢ Output :-



```
print(summary_table)
```

```
                            Coefficient  Hazard Ratio  Std. Error   Z-Score  \
covariate
Gender                        -0.048388      0.952764    0.122152 -0.396128
Total_Bilirubin                0.004466      1.004476    0.013170  0.339146
Direct_Bilirubin               0.057850      1.059556    0.031099  1.860203
Alkaline_Phosphotase          -0.000013      0.999987    0.000196 -0.068760
Alamine_Aminotransferase       0.001575      1.001577    0.000320  4.918175
Aspartate_Aminotransferase    -0.000413      0.999587    0.000192 -2.149115
Total_Protiens                 0.173390      1.189330    0.100518  1.724975
Albumin                       -0.039558      0.961214    0.182688 -0.216533
Albumin_and_Globulin_Ratio     0.391479      1.479167    0.274790  1.424648

                                 P-Value   -log2(p)
covariate
Gender                      6.920105e-01   0.531134
Total_Bilirubin             7.345001e-01   0.445165
Direct_Bilirubin            6.285676e-02   3.991788
Alkaline_Phosphotase        9.451803e-01   0.081339
Alamine_Aminotransferase    8.735488e-07  20.126608
Aspartate_Aminotransferase  3.162526e-02   4.982779
Total_Protiens              8.453195e-02   3.564359
Albumin                     8.285726e-01   0.271300
Albumin_and_Globulin_Ratio  1.542590e-01   2.696573
```

➢ Conclusion :-

At 5% level of significance, the variables “Direct_Bilirubin”, “Alamine_Aminotransferase”, “Aspartate_Aminotransferase” and “Total_Protiens” has an effect on hazard rate.

Conclusions:

1) By the data visualization we can predict that the liver disease is independent of age. Also the males have more liver disease probability than the females
2) In Logistic Regression only Amine is the factor which shows positive correlation with the presence or absence of the disease. Hence as amount of amine in the body increases the chance if getting the liver disease is also increases.
3) By applying various techniques of machine learning, we get highest accuracy of the Random Forest model as 97% on original data and precision is very less since, count of people having liver disease in our data is very low.
4) Since the data was imbalanced, we have balanced the data by applying oversampling technique. In Oversampled data Random Forest model has highest accuracy of 97 %.
5) From the log-rank survival analysis plot it is clear that the survival probabilities of the males and the females are identical.
6) From cox-proportionality hazard rate model it is clear that, at 5% level of significance, the variables “Direct_Bilirubin”, “Alamine_Aminotransferase”, “Aspartate_Aminotransferase” and “Total_Protiens” has an effect on hazard rate.

Limitations:

1) As the data is imbalanced, it does not consist of the balanced number of counts of having males and females and also the person having a disease or not.

2) After balancing the data, that is after applying under sampling technique, the accuracy of the model decreased, indicating that in the data like clinical data cannot be treated with balancing the observations by using under sampling method. We can deal with the real time data so that the accuracy of the model ismaintained.

3) We could predict only 70% of correct cases without doing the oversampling. So this indicates that the dataset we taken has limited number of observations we need more observations for predicting the accurate results.

Future Scope of the Study

1) Liver disease is one of the major diseases in most of the areas nowadays.
2) As the data consists of personal key indicators of liver disease or factors for risk of liver disease, the personal factors which depends on individual to individual can be taken into consideration for the risk of liver disease.
3) The patients/ individuals who don't have liver disease may be given indication for future if those factors are seen in that individual which are important for risk of liver disease.

➢ Software Used:

1) R-Software and R-Studio
2) MS-Excel
3) Python
4) Google Colab
5) MS-Word

References

1) www.google.com
2) www.kaggle.com
3) www.wikipedia.co.in
4) Geeks for Greeks – Codes of Python .
5) Introduction to LinearRegression Analysis(Bookby-

   Douglas Montagomery,Elizabeth A.peck,and G.Geoffery Vining).