**RSIP Second ML 007**

**Project Title**

**Liver Patient Analysis**

**Category: Machine Learning**

**Skills Required:**

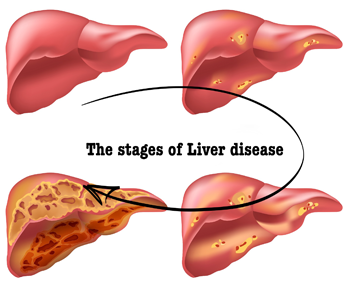
**Python, Python For Data Analysis, Python For Data Visualization, Exploratory Data Analysis, Data Pre-processing Techniques, Machine Learning, Regression Algorithms.**

**Team members:**

**K.Indhumadhi**

**S.Bagavathi@Bavatharani**

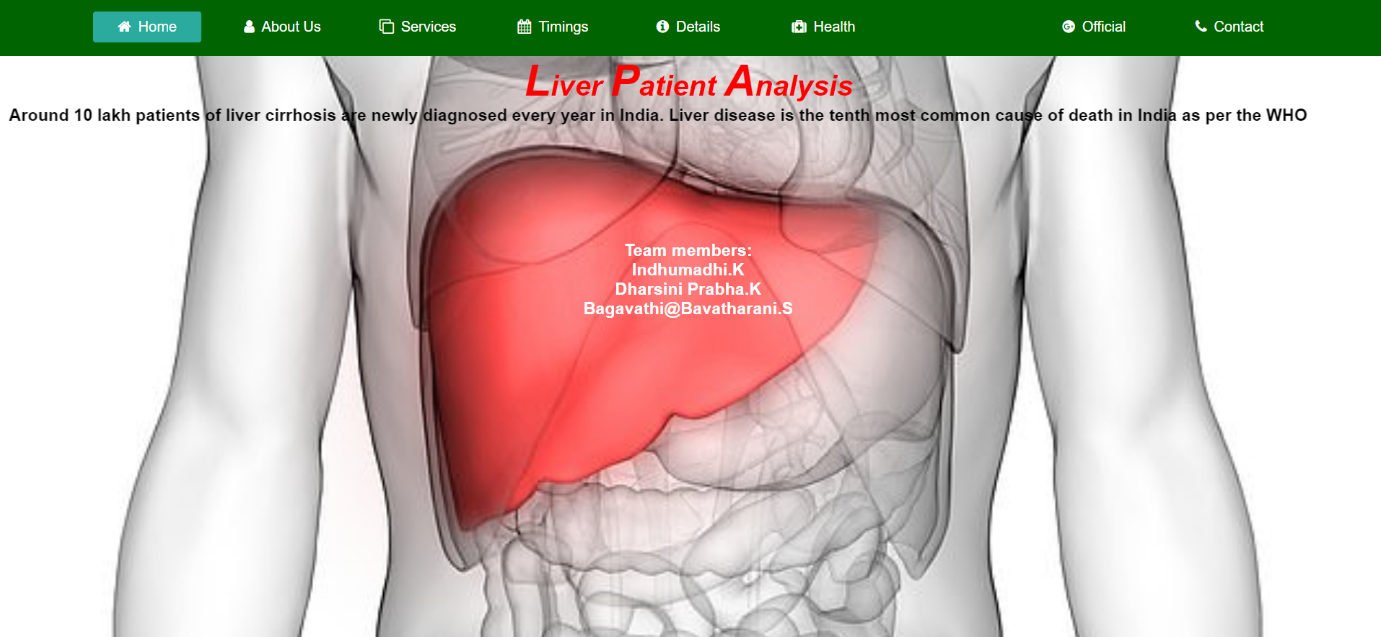
**K.Dharsini Prabha**



**PROJECT DESCRIPTION:**

The human liver is one of the major organs in the body and liver disease can cause many problems in human life. Now a days liver disease is extending markedly due to excessive alcohol consumption, smoking, drinking arsenic contaminated water, obesity, low immunity and by inheritance. Fast and accurate prediction of liver disease allows early and effective treatments. According to the Global Burden of Disease project, in the year of 2010, more than 2 million deaths were due to major liver diseases.

**LIVER PATIENT ANALYSIS:**



**INTRODUCTION:**

Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged. An early diagnosis of liver problems will increase patient’s survival rate. Liver failures are at high rate of risk among Indians. It is expected that by 2025 India may become the World Capital for Liver Diseases. The widespread occurrence of liver infection in India is contributed due to deskbound lifestyle, increased alcohol consumption and smoking. There are about 100 types of liver infections. Therefore, developing a machine that will enhance in the diagnosis of the disease will be of a great advantage in the medical field. These systems will help the physicians in making accurate decisions on patients and also with the help of Automatic classification tools for liver diseases (probably mobile enabled or web enabled), one can reduce the patient queue at the liver experts such as endocrinologists.

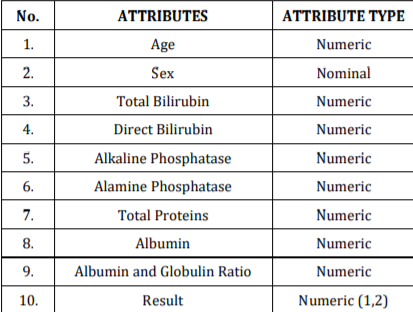
The dataset used is The Indian Liver Patient Dataset (ILPD) which was selected from UCI Machine learning repository for this study. It is a sample of the entire Indian population collected from Andhra Pradesh region and comprises of 585 patient data.

**RELATED WORKS:**

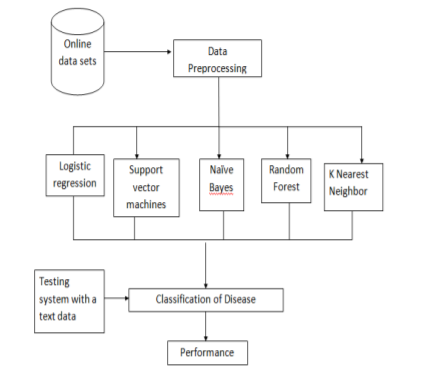
In recent research works, several neural network models have been developed to aid in diagnosis of liver diseases in the medical field by the physicians such as diagnosis support system [3], expert system, intelligent diagnosis system, and hybrid intelligent system. In addition, Christopher N. [4] proposed a system to diagnose medical diseases considering 6 benchmarks which are liver disorder, heart diseases, diabetes, breast cancer, hepatitis and lymph. The authors developed two systems based on WSO and C4.5, an accuracy of 64.60% with 19 rules of liver disorder dataset and 62.89% with 43rules which was obtained from the WSO and C4.5respectively. Ramana [5] also made acritical study on liver diseases diagnosis by evaluating some selected classification algorithms such as naïve Bayes classifier, C4.5, backpropagation neural network, K-NN and support vector. The authors obtained 51.59% accuracy on Naïve Bayes classifier, 55.94% on C4.5 algorithm, 66.66% on BPNN, 62.6% on KNN and 62.6% accuracy on support vector machine.

**IMPLEMENTATION:**

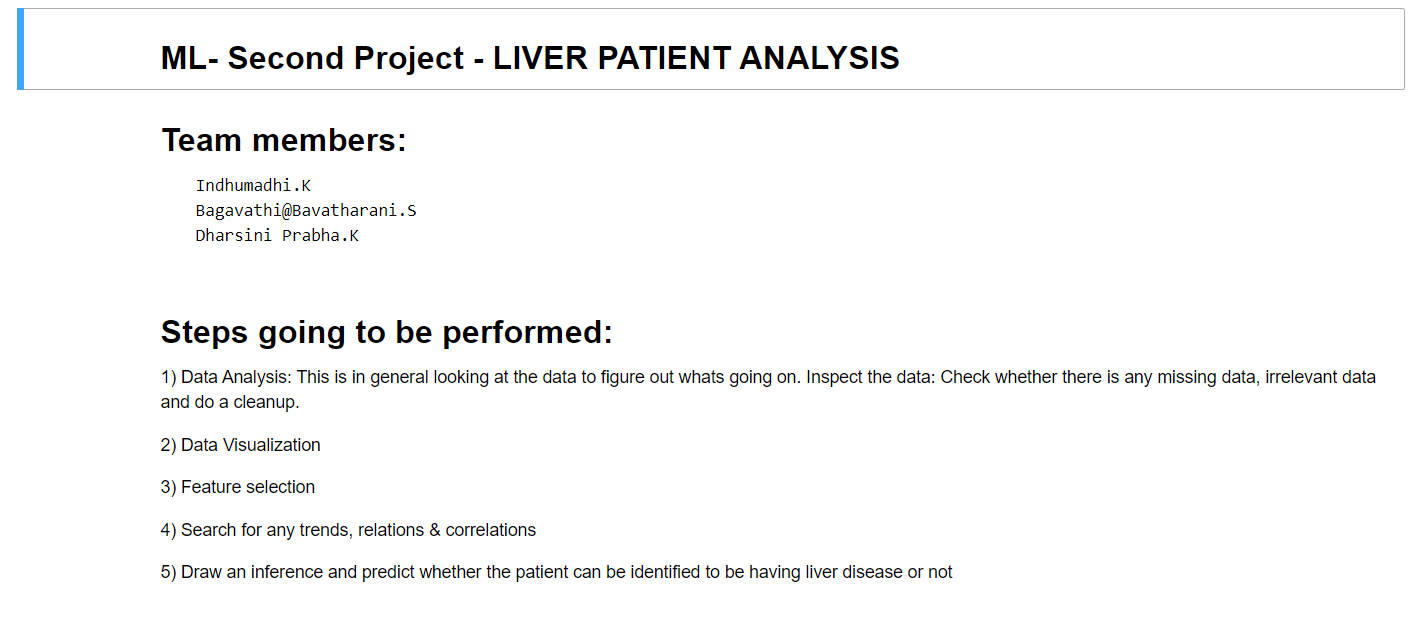
**DATASET** The Indian Liver Patient Dataset comprised of 10 different attributes of 583 patients. The patients were described as either 1 or 2 on the basis of liver disease. The detailed description of the dataset is shown in Table.



**OVERALL PROCESS:**



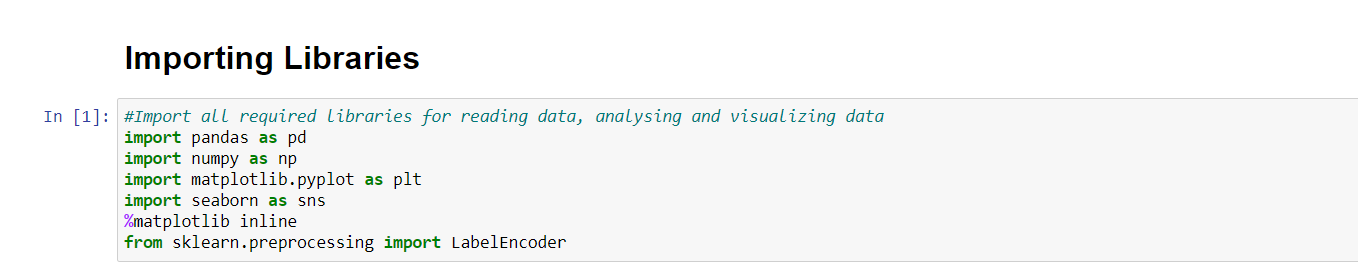
**PROJECT AND DETAILS:**



**DATA PREPARATION AND EXPLORATION:**

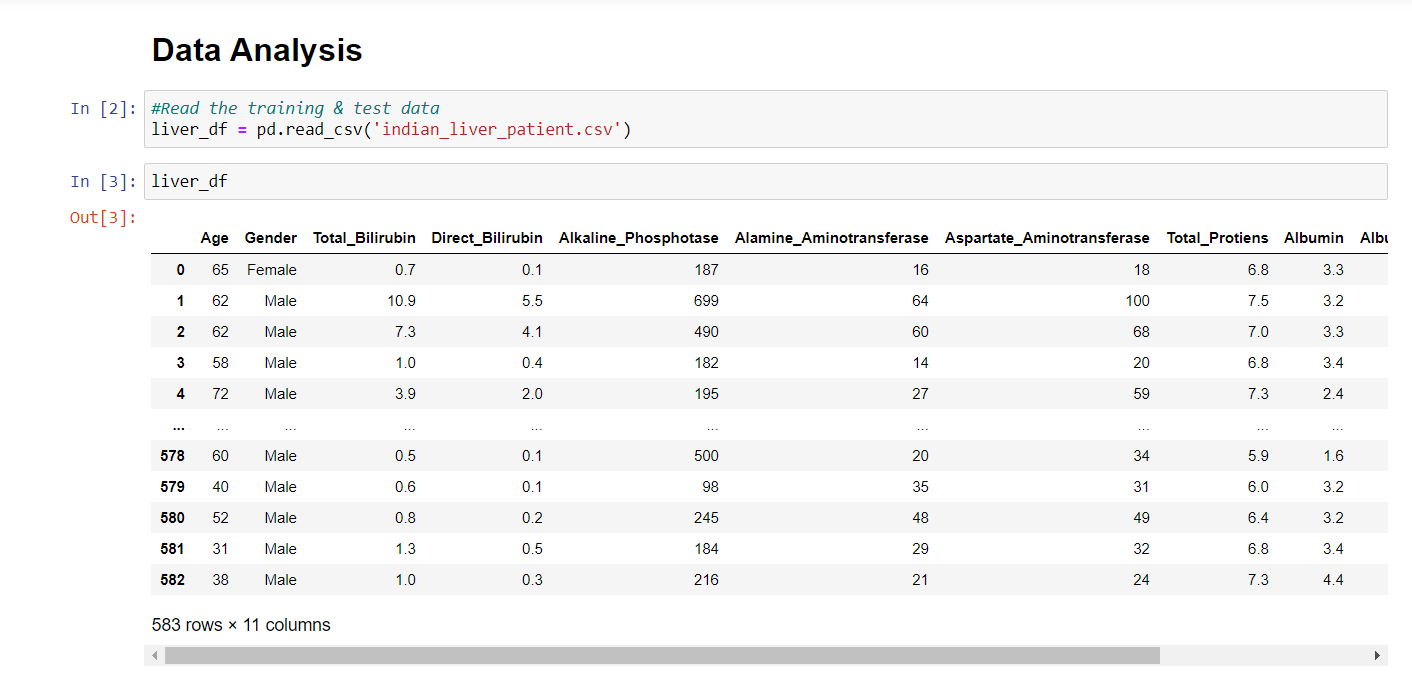
   Importing the Libraries:

     Pandas, Numpy, Matplotlib

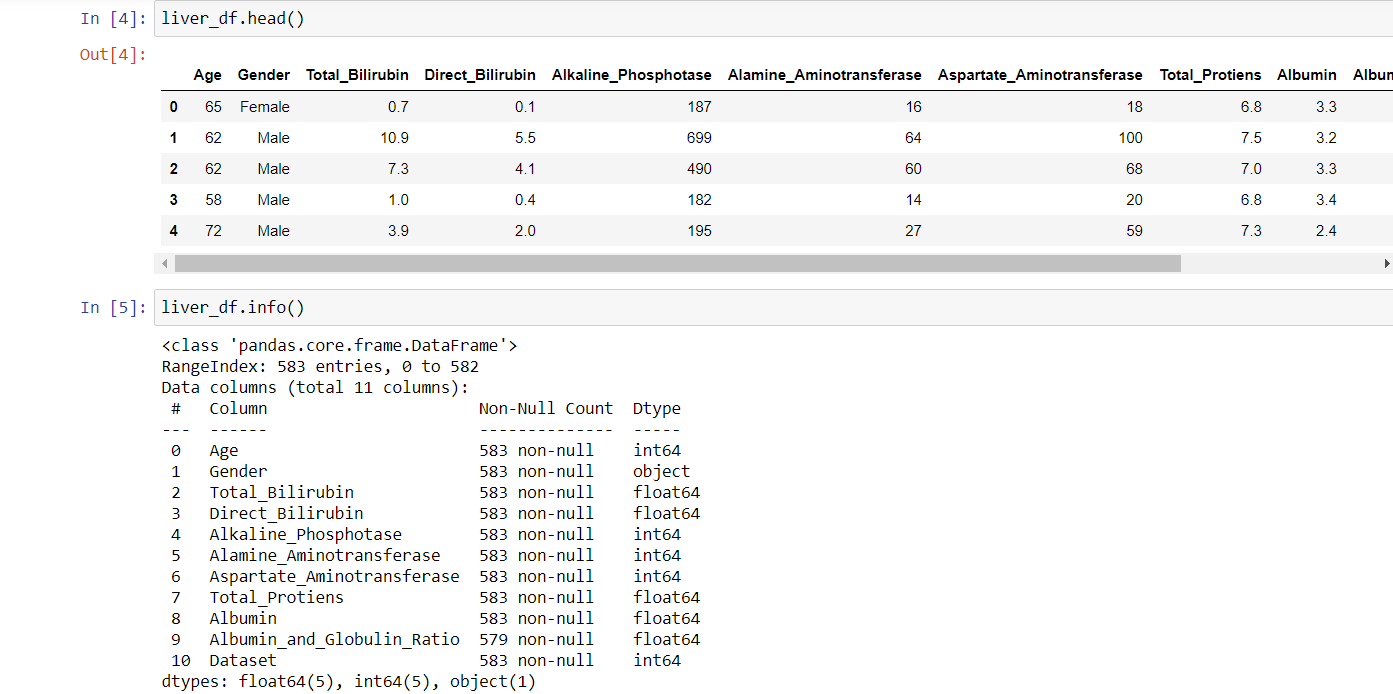


**DATA PREPROCESSING:**

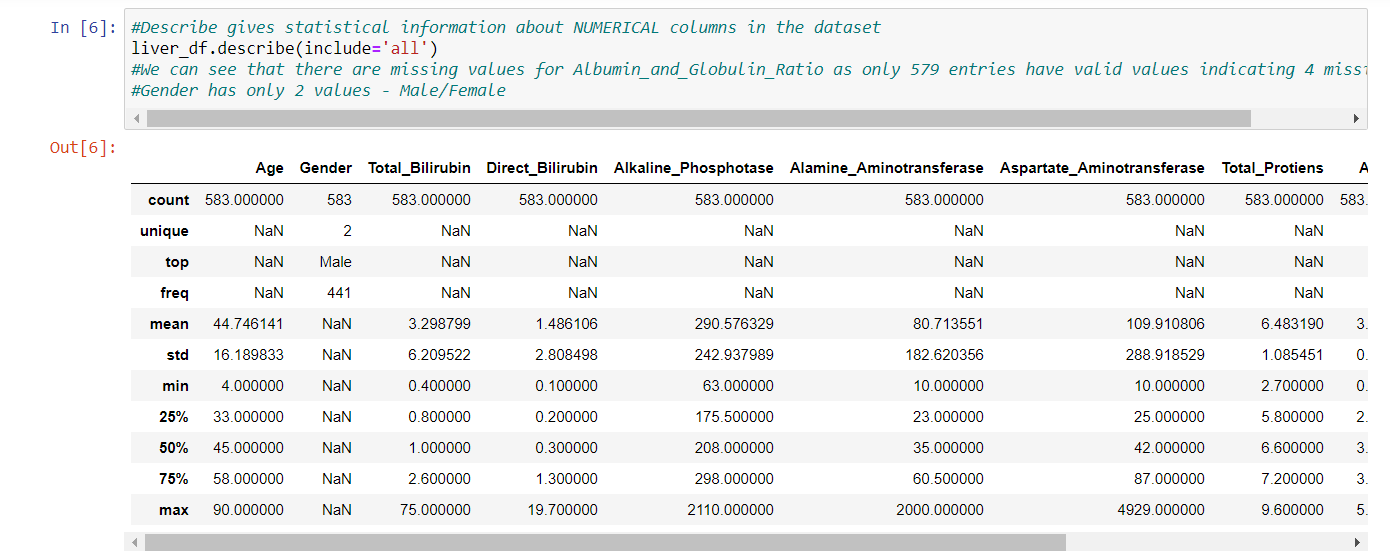
Data pre-processing is an important step of solving every machine learning problem. Most of the datasets used with Machine Learning problems need to be processed / cleaned / transformed so that a Machine Learning algorithm can be trained on it. Most commonly used pre-processing techniques are very few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

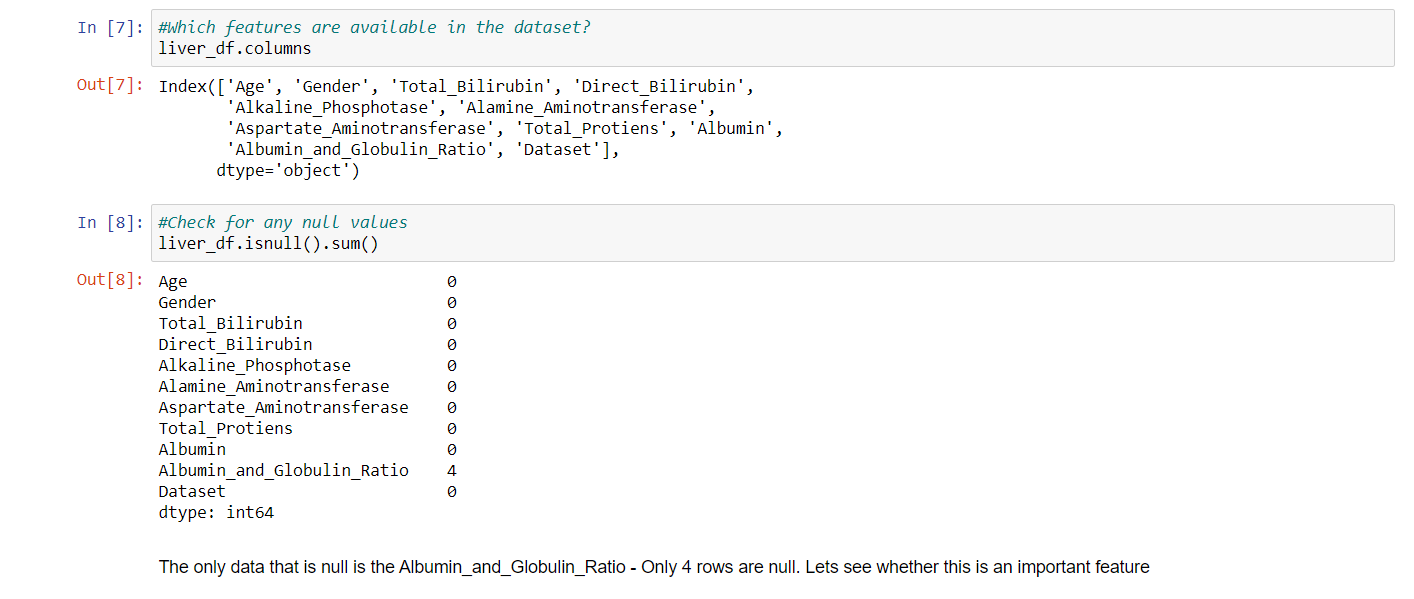


This data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India. The "Dataset" column is a class label used to divide groups into liver patient (liver disease) or not (no disease).



**Here is the observation from the dataset:**  
**1) Only gender is non-numeric veriable. All others are numeric.**  
**2) There are 10 features and 1 output - dataset. Value 1 indicates that the patient has liver disease and 0 indicates the patient does not have liver disease.**

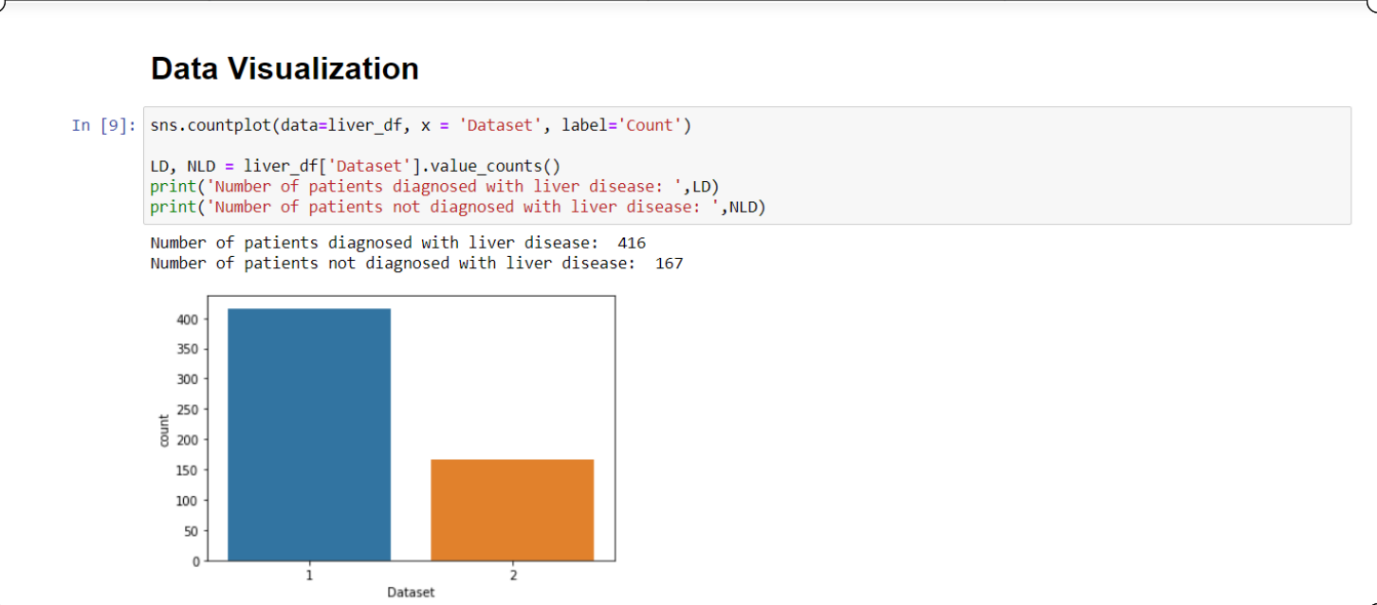




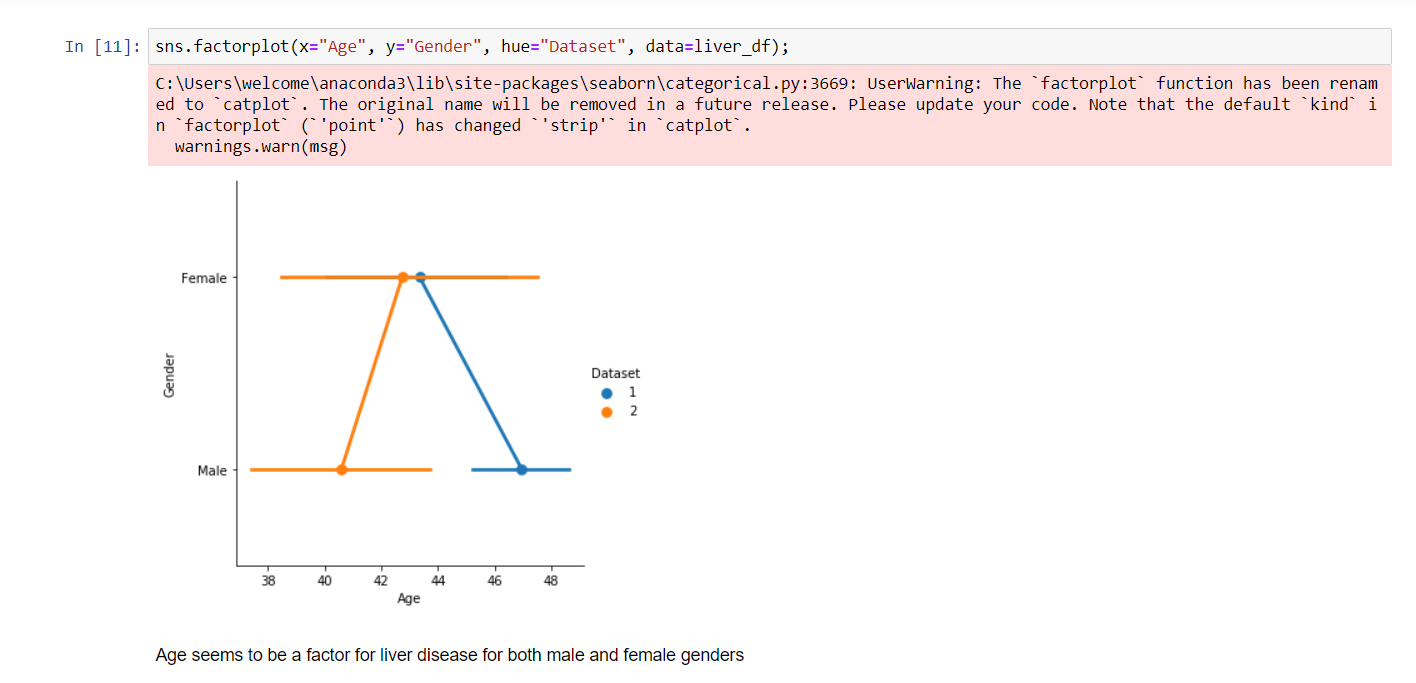
The only data that is null is the Albumin\_and\_Globulin\_Ratio - Only 4 rows are null. Lets see whether this is an important feature

**DATA VISUALIZATION:**

**Data visualization** is the representation of **data** or information in a graph, chart, or other visual format. It communicates relationships of the **data** with images. This is important because it allows trends and patterns to be more easily seen.

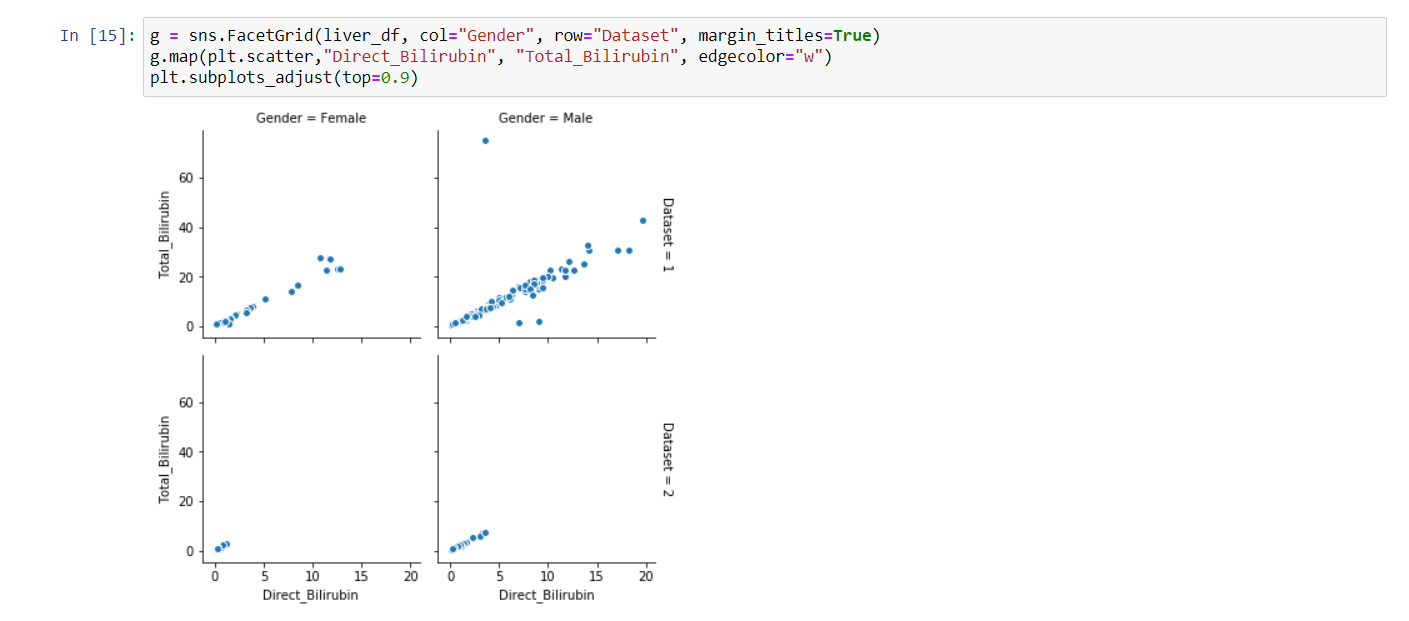


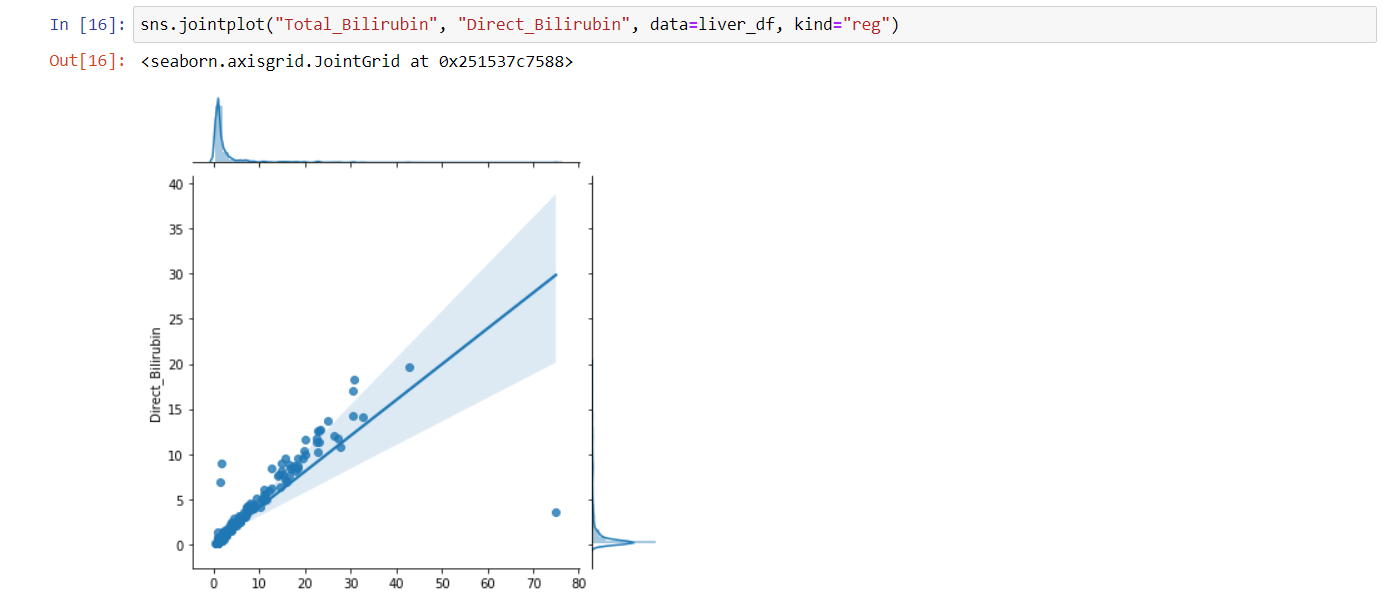




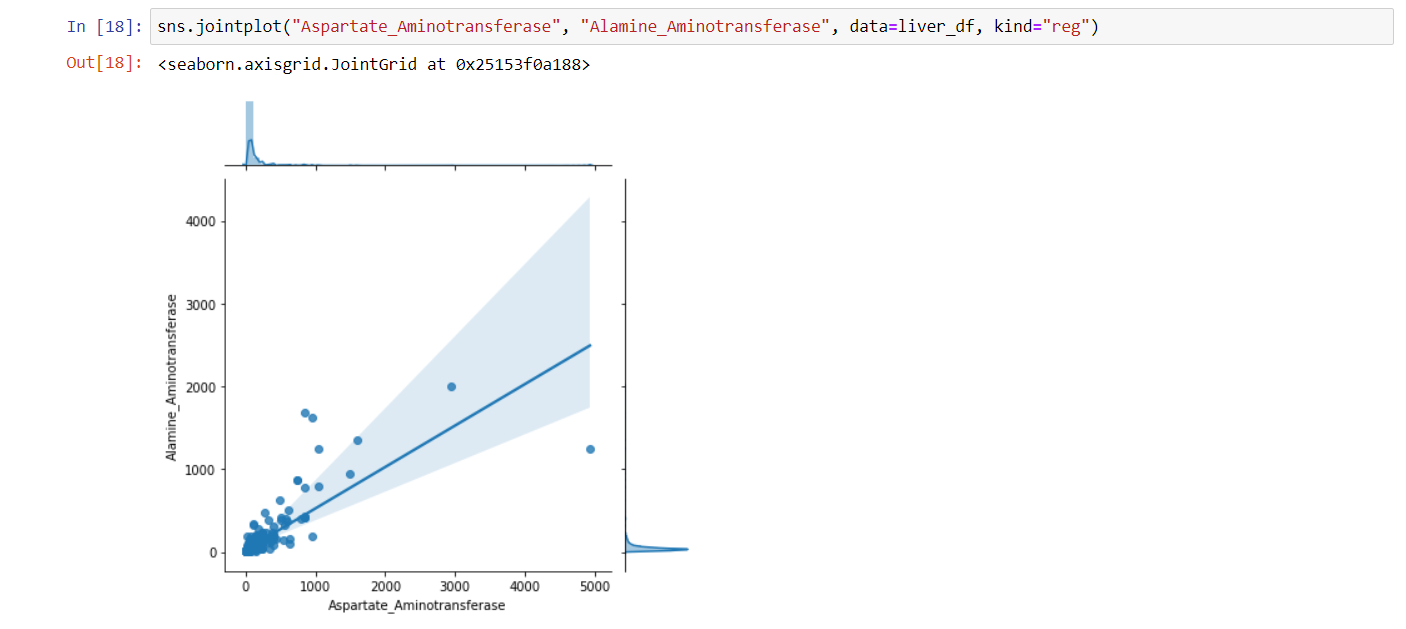


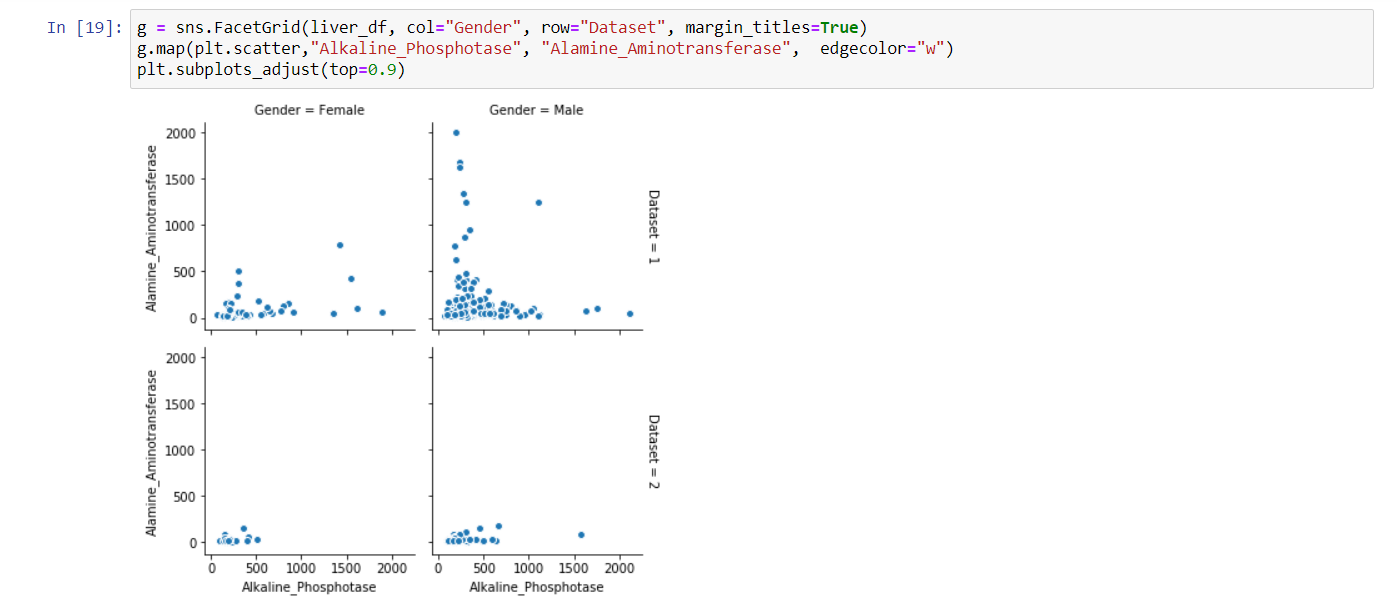




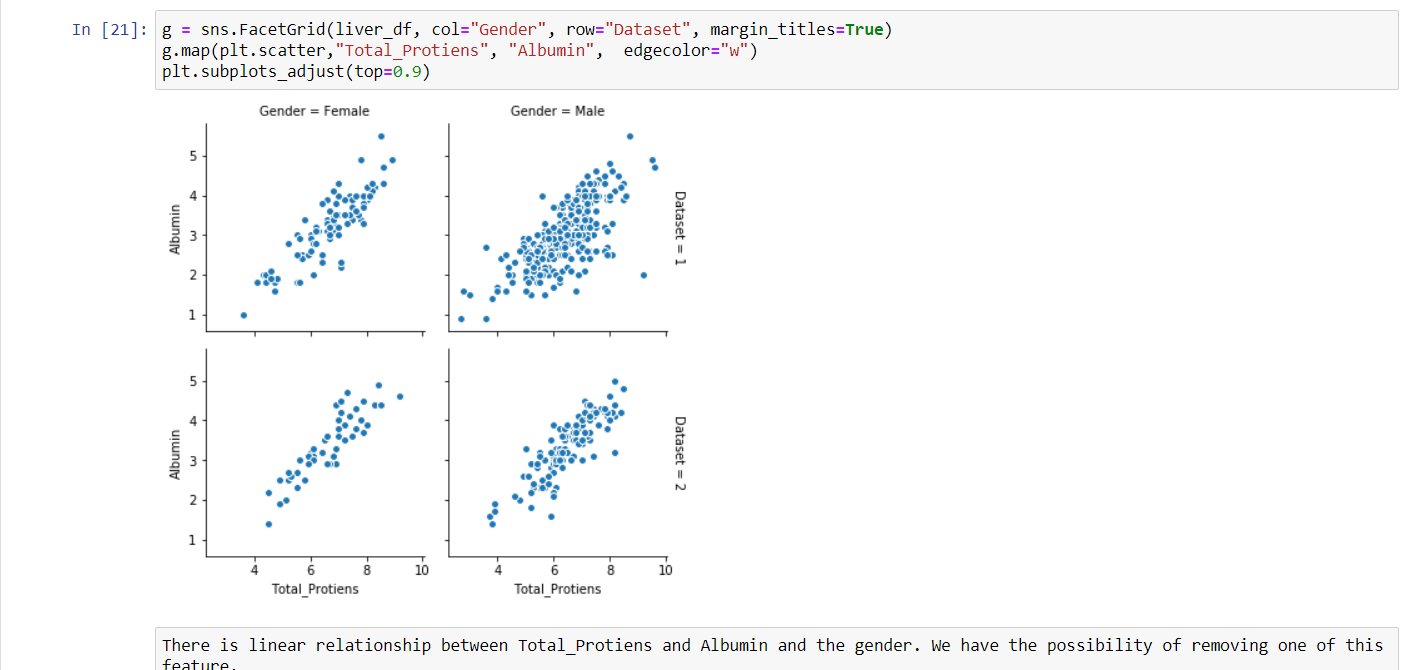


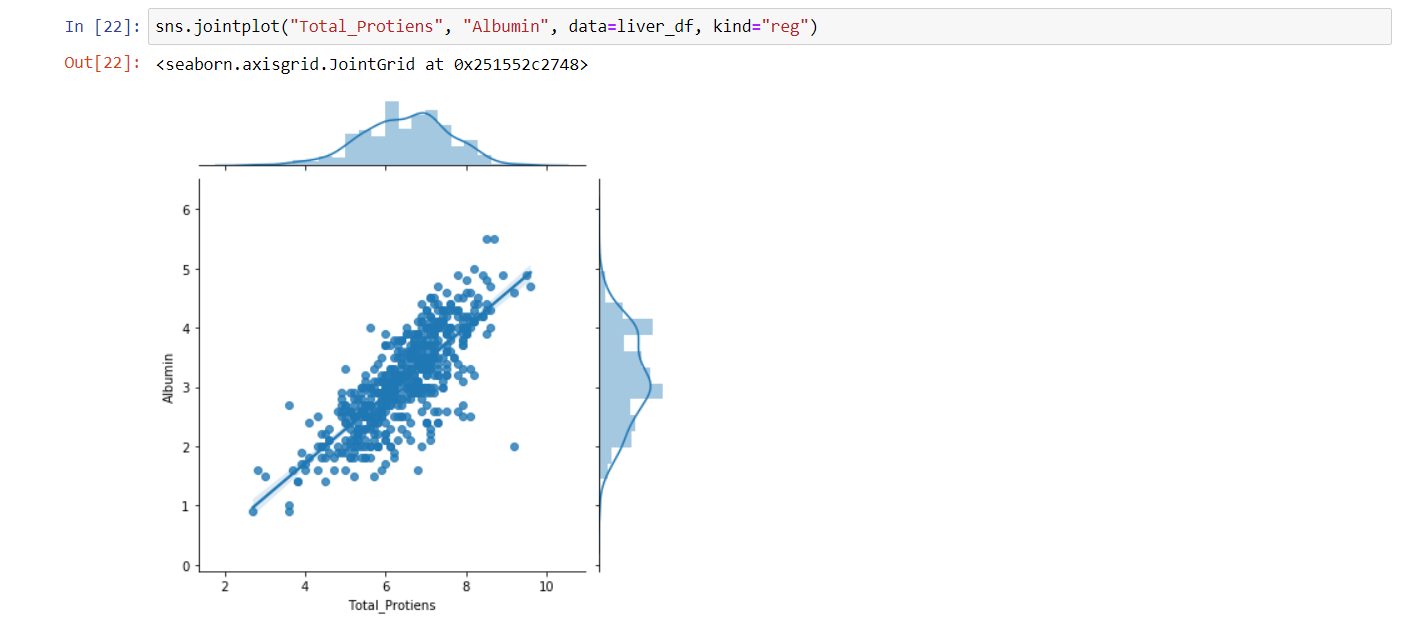




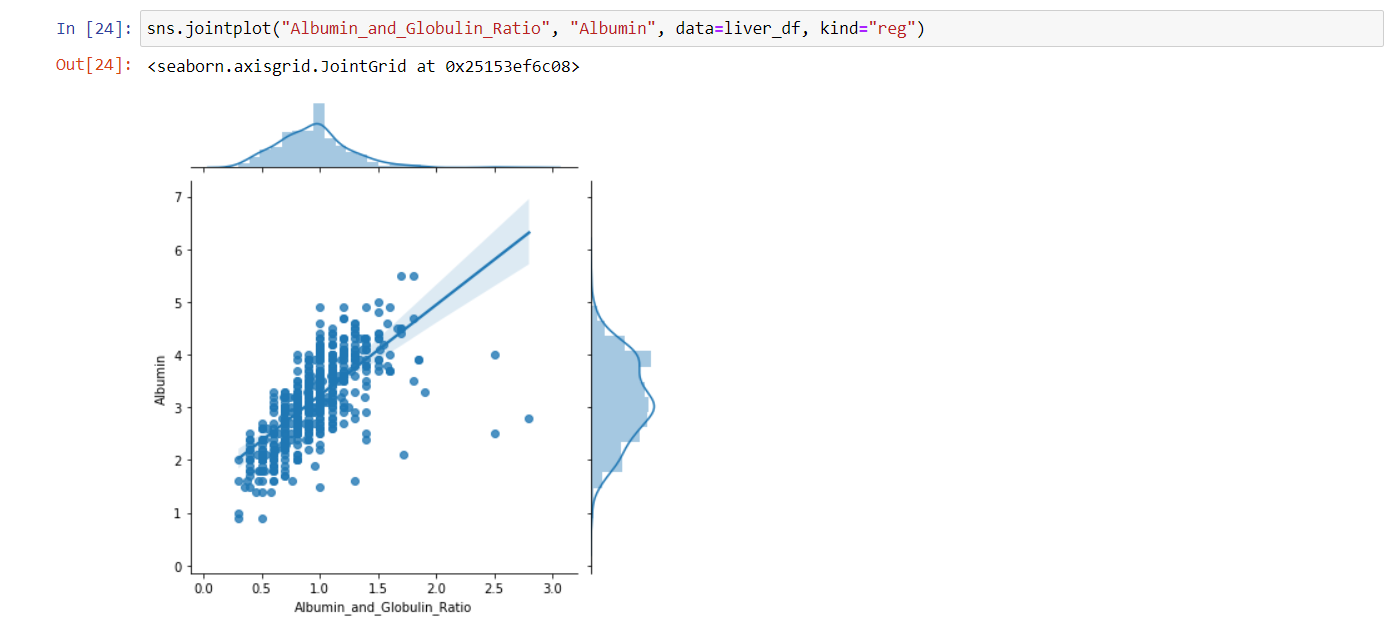


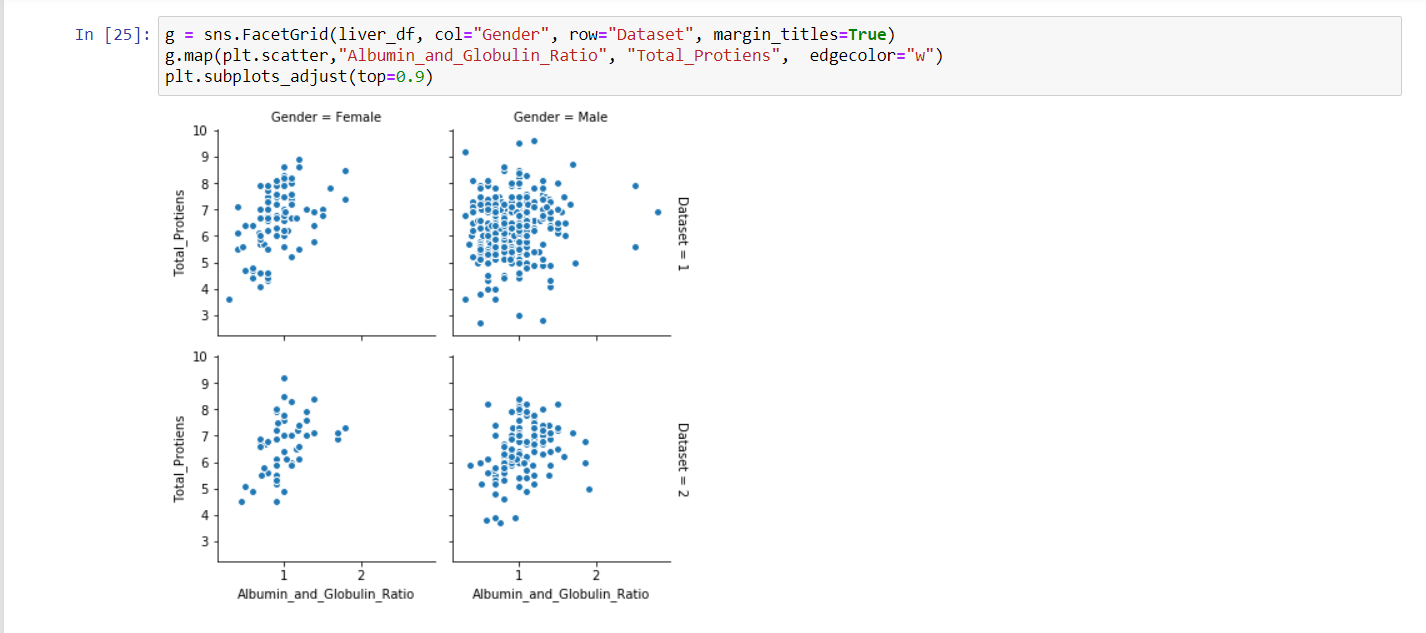


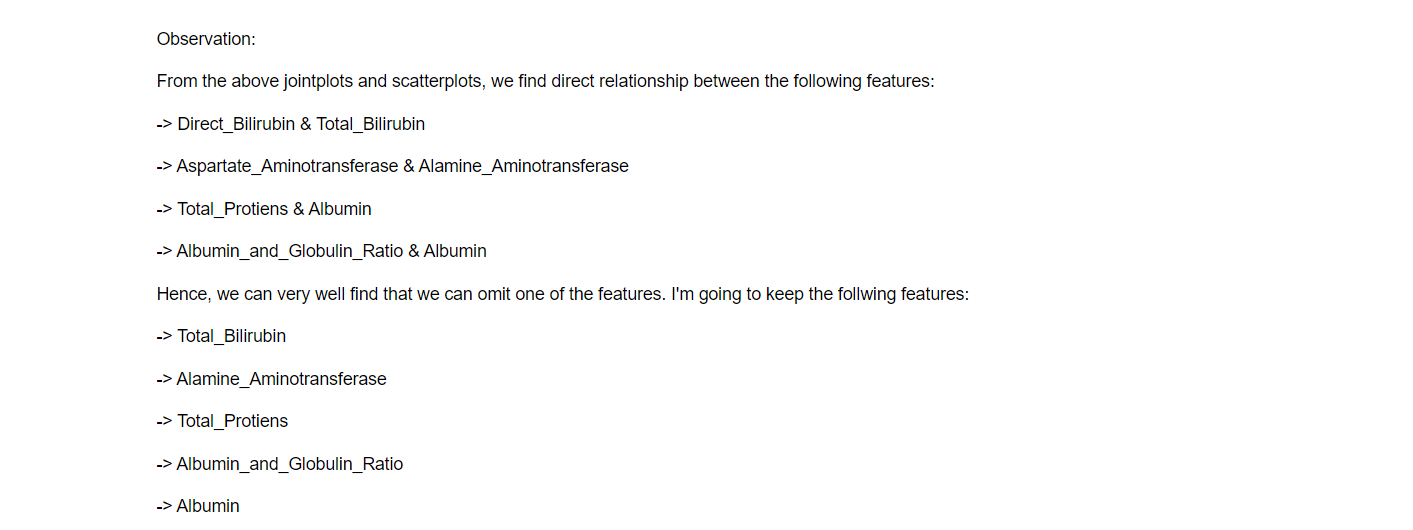


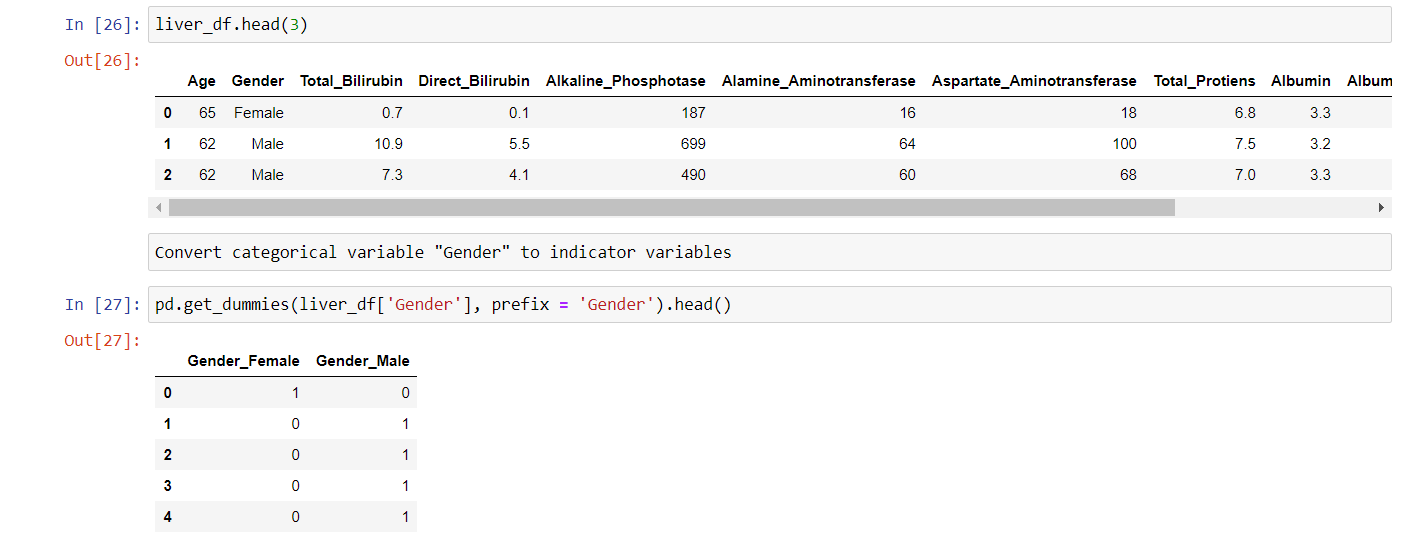


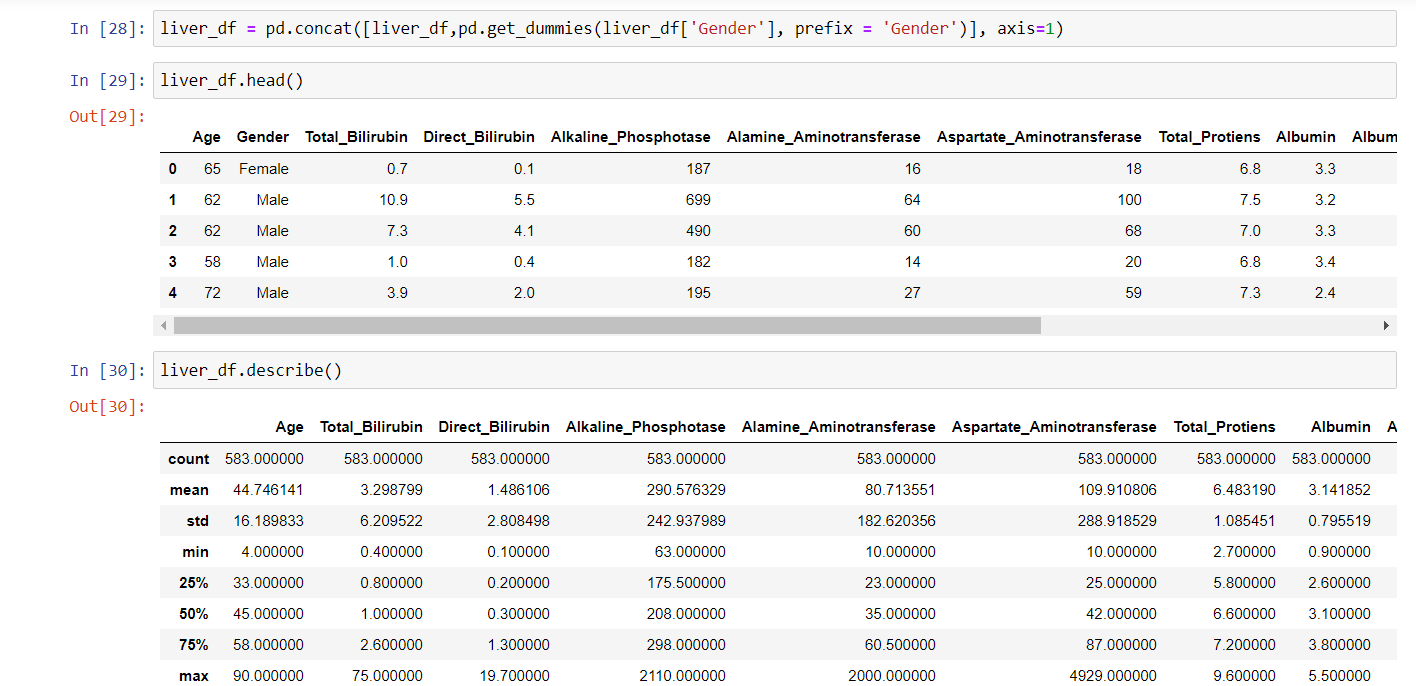


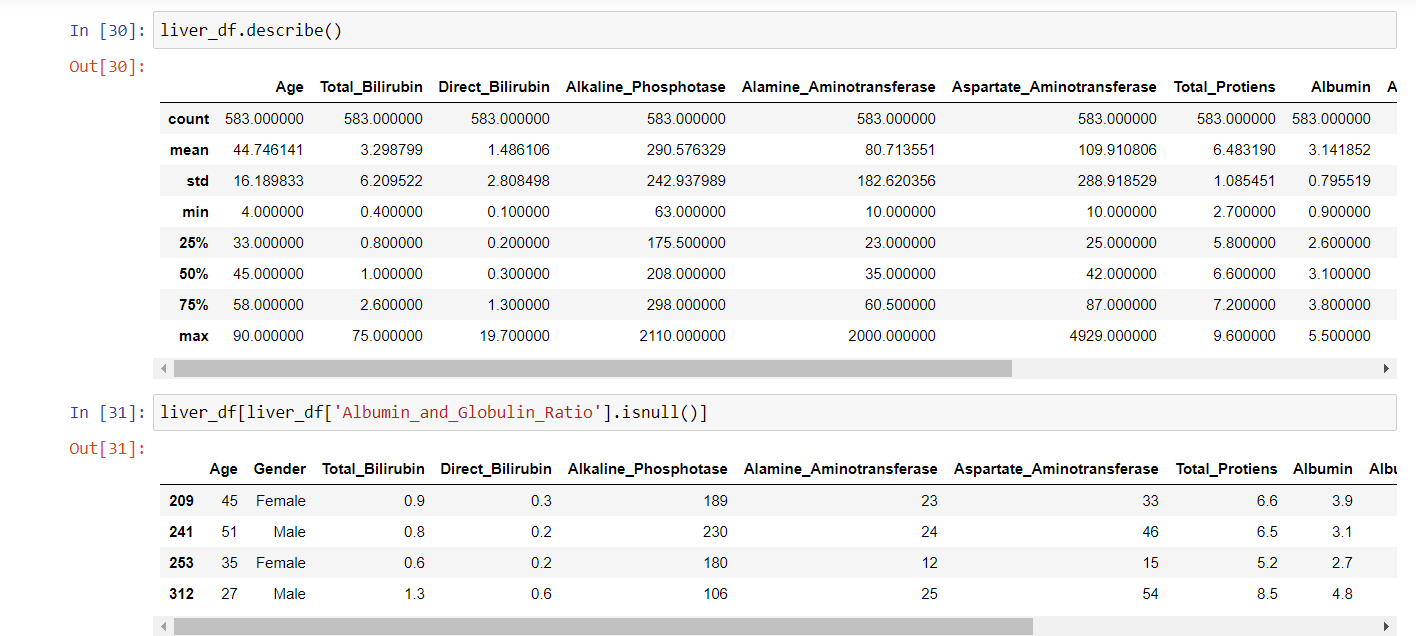


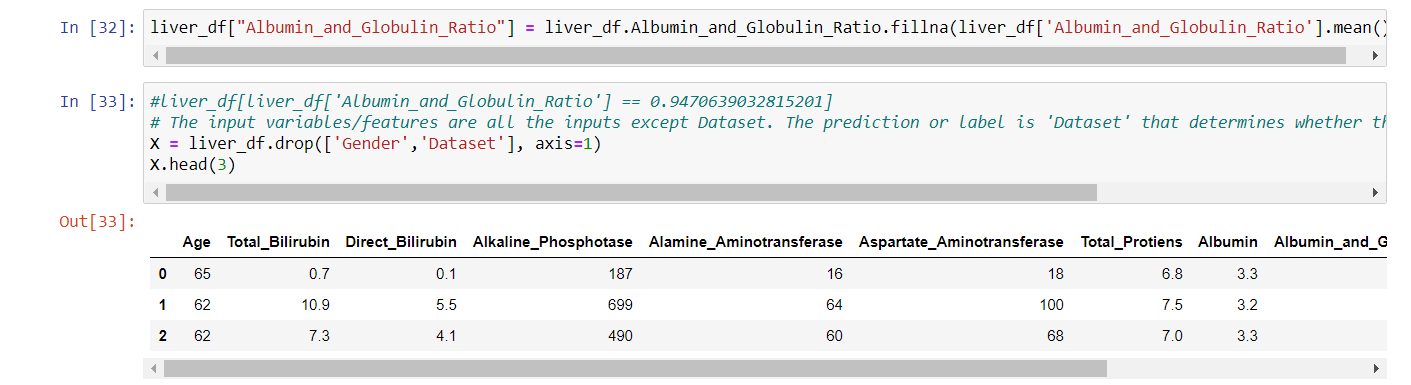


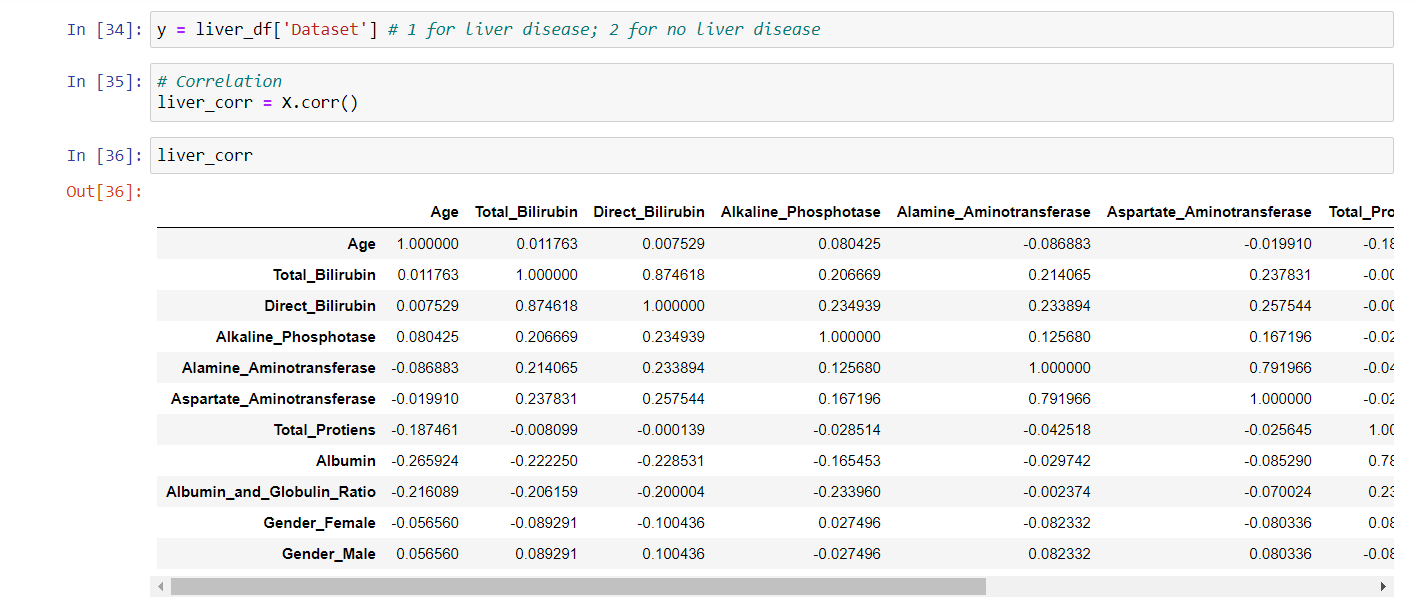


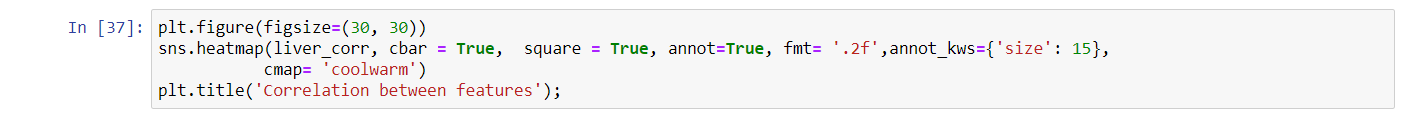


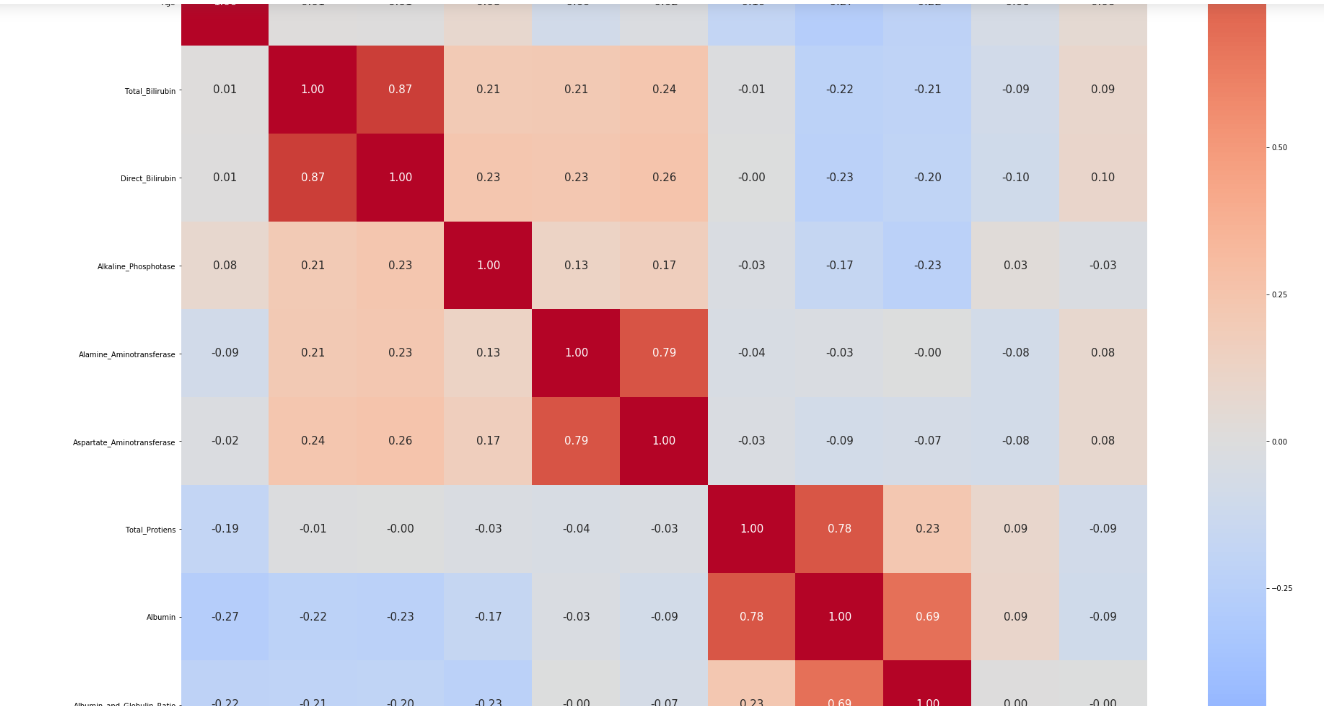


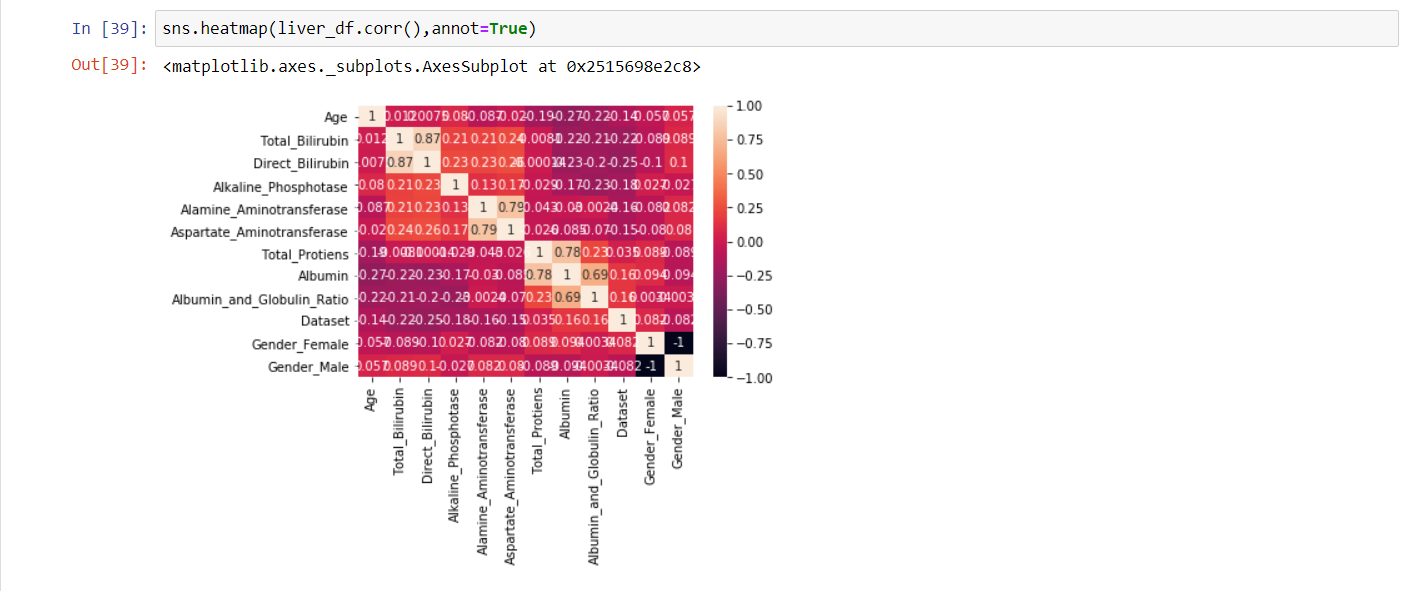






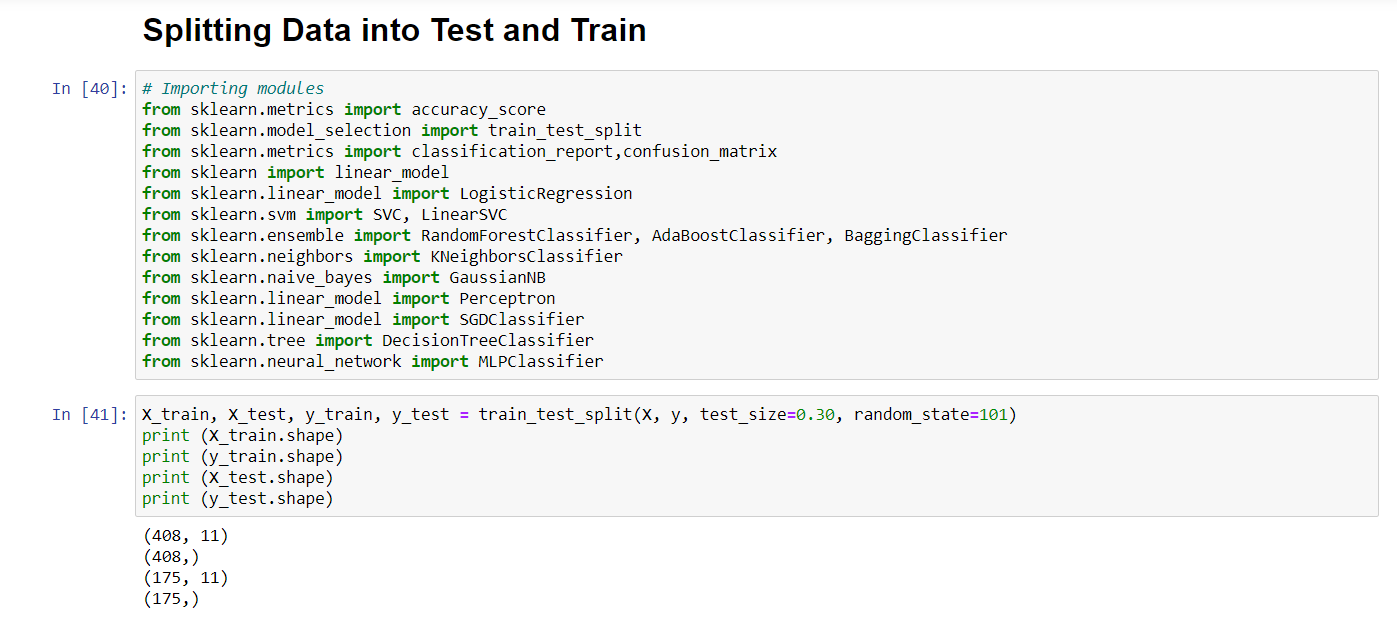






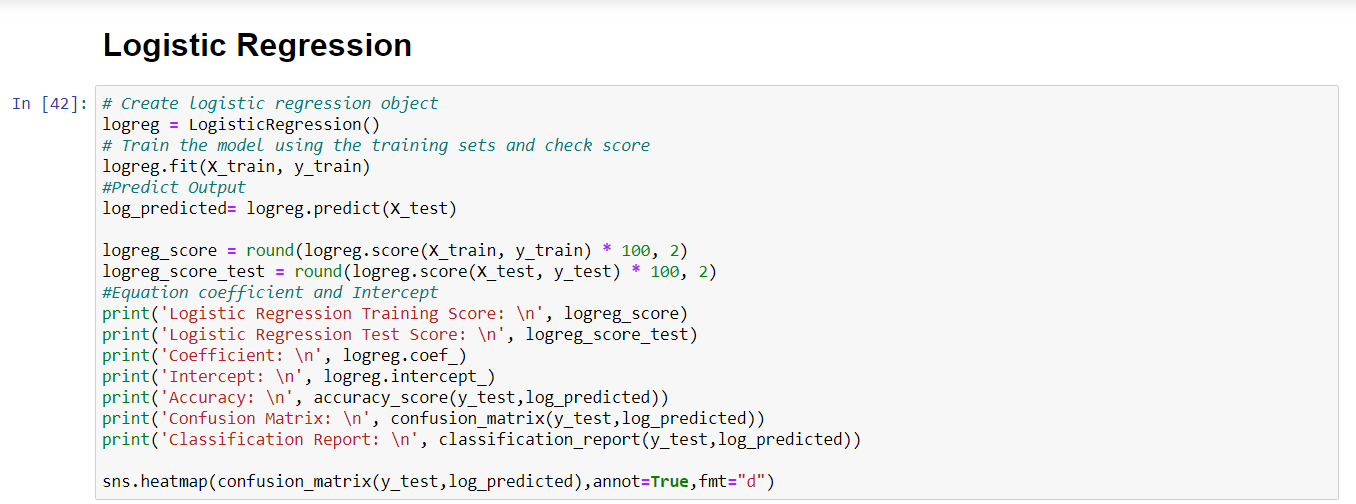
**MODEL BUILDING:**

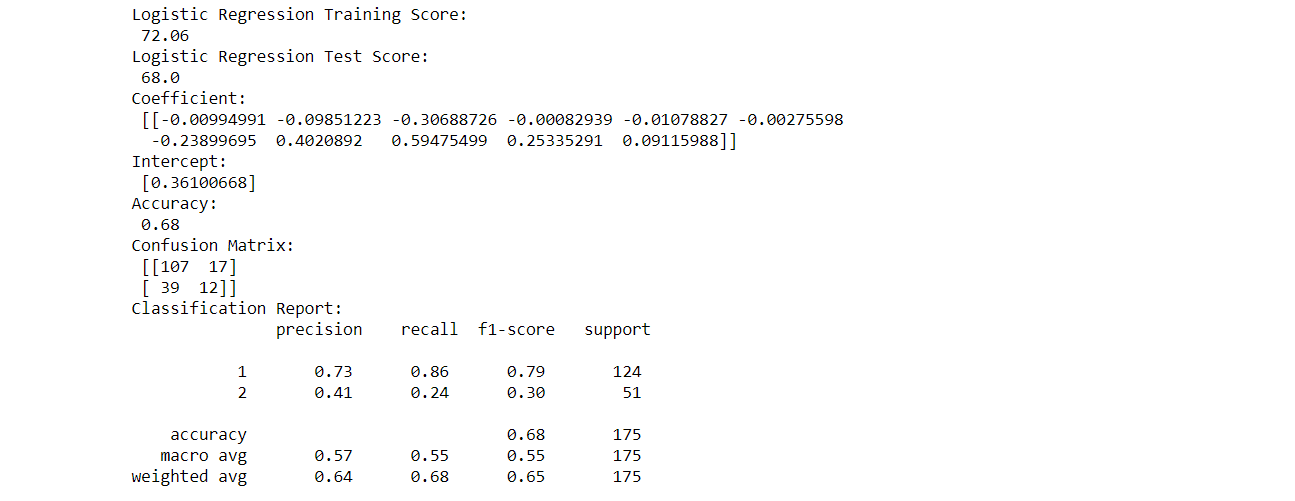
**Training and Testing data:**



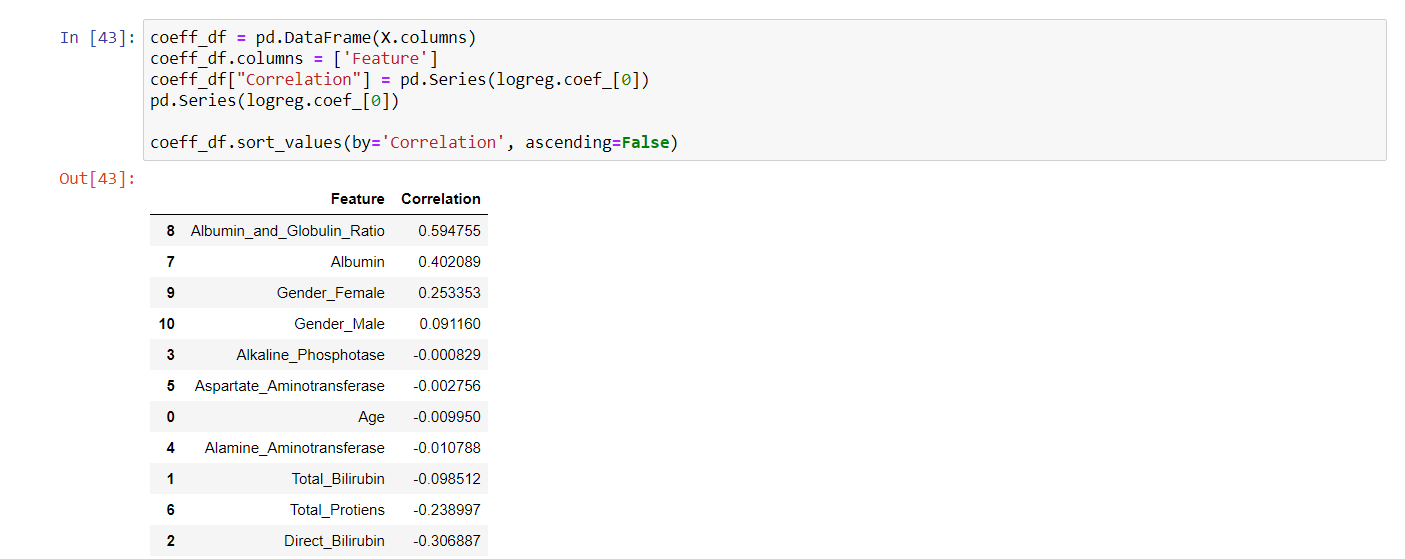
**Logistic Regression:**

Logistic regression is one of the simpler classification models. Because of its parametric nature it can to some extent be interpreted by looking at the parameters making it useful when experimenters want to look at relationships between variables. A parametric model can be described entirely by a vector of parameters = (0, 1... p). An example of a parametric model would be a straight-line y = kx + m where the parameters are k and m. With known parameters the entire model can be recreated. Logistic regression is a parametric model where the parameters are coefficients to the predictor variables written as 0 +1 +X1 + ...PXp Where 0 is called the intercept. For convenience we instead write the above sum of the parameterized predictor variables in vector form as X. The name logistic regression is a bit unfortunate since a regression model is usually used to find a continuous response variable, whereas in classification the response variable is discrete. The term can be motivated by the fact that we in logistic regression found the probability of the response variable belonging to a certain class, and this probability is continuous.

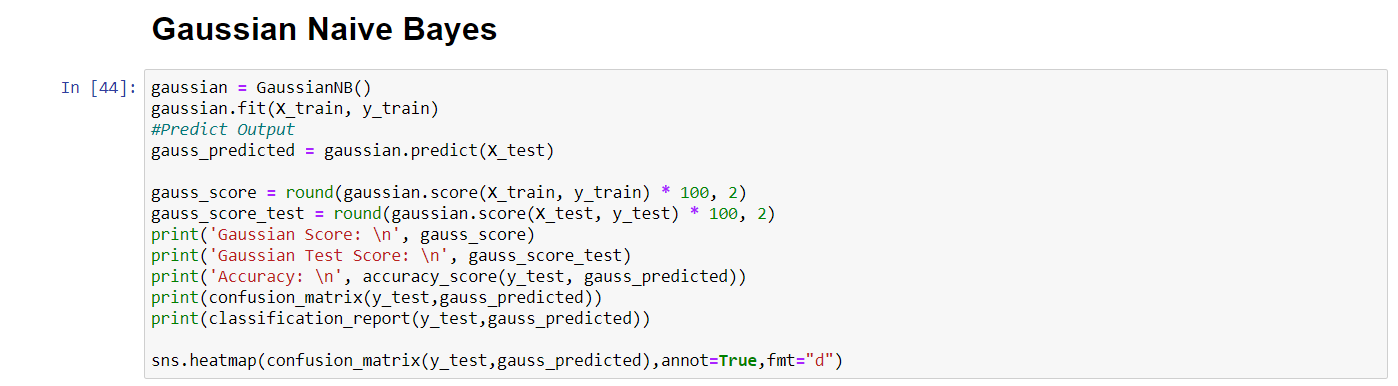


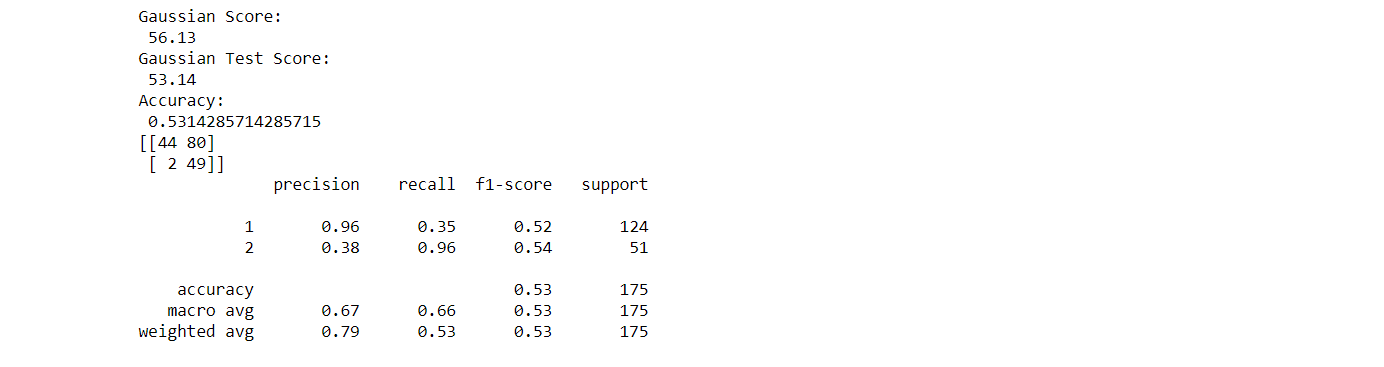


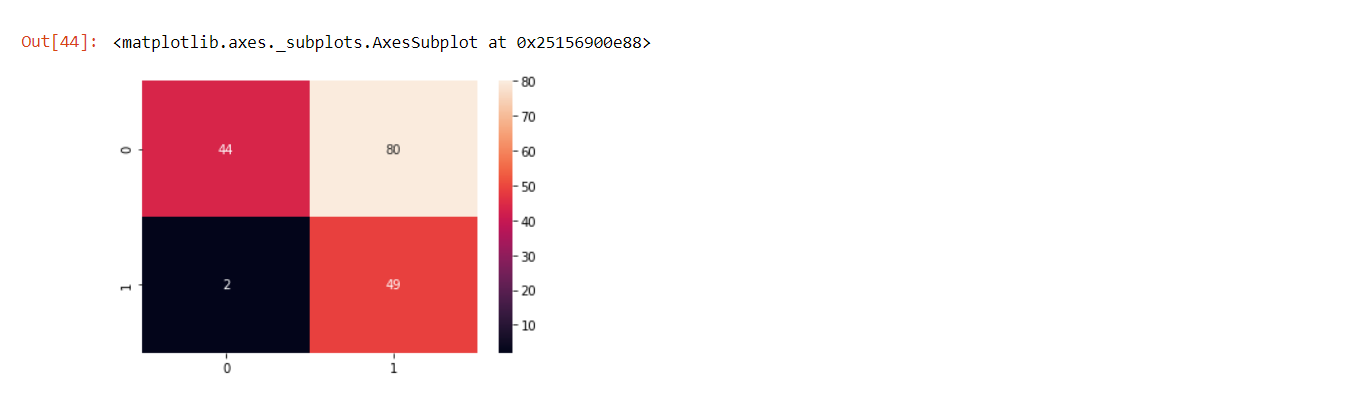




**Gaussian Naive Bayes:**

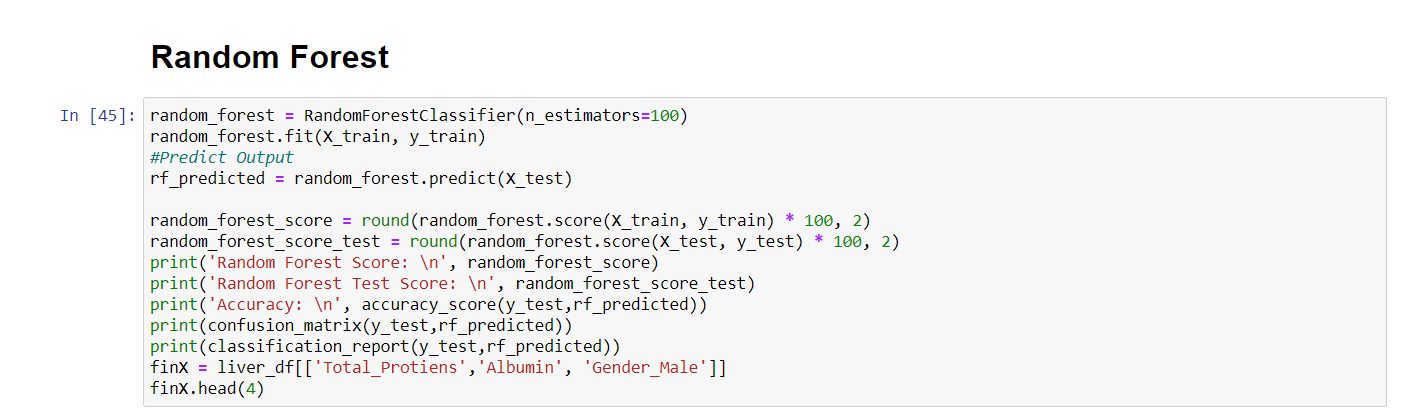


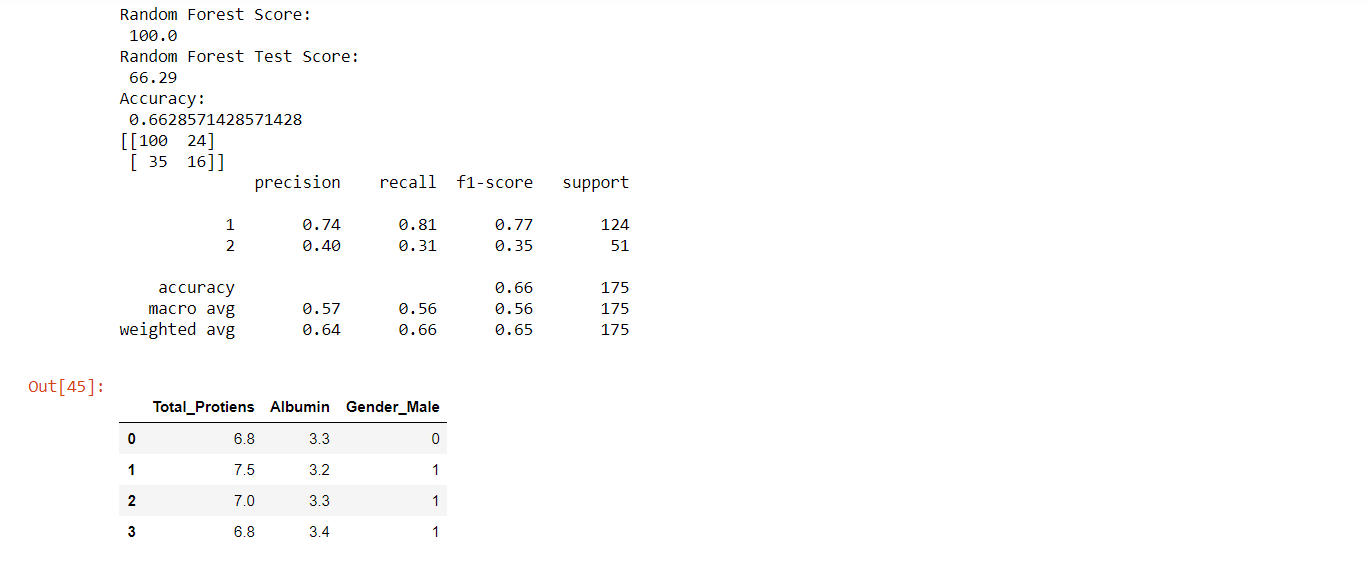




**Random Forest:**

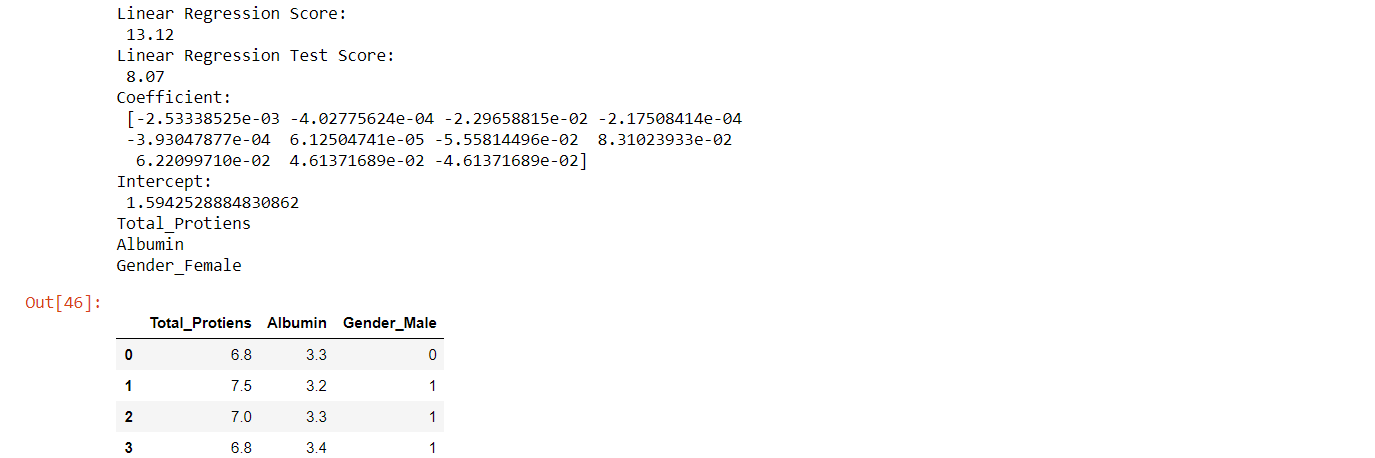
Random forests is a machine learning regression method for classification that drive by constructing liver data into a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. It is unexcelled in accuracy among current algorithms. It output classification efficiently on large liver dataset. It can handle thousands of input attributers without variable deletion. It gives estimates of what variables are important in the classification. Random Forests grows many classification trees. To classify a new liver object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and says the tree "votes" for that class. The forest chooses the classification having the most votes).

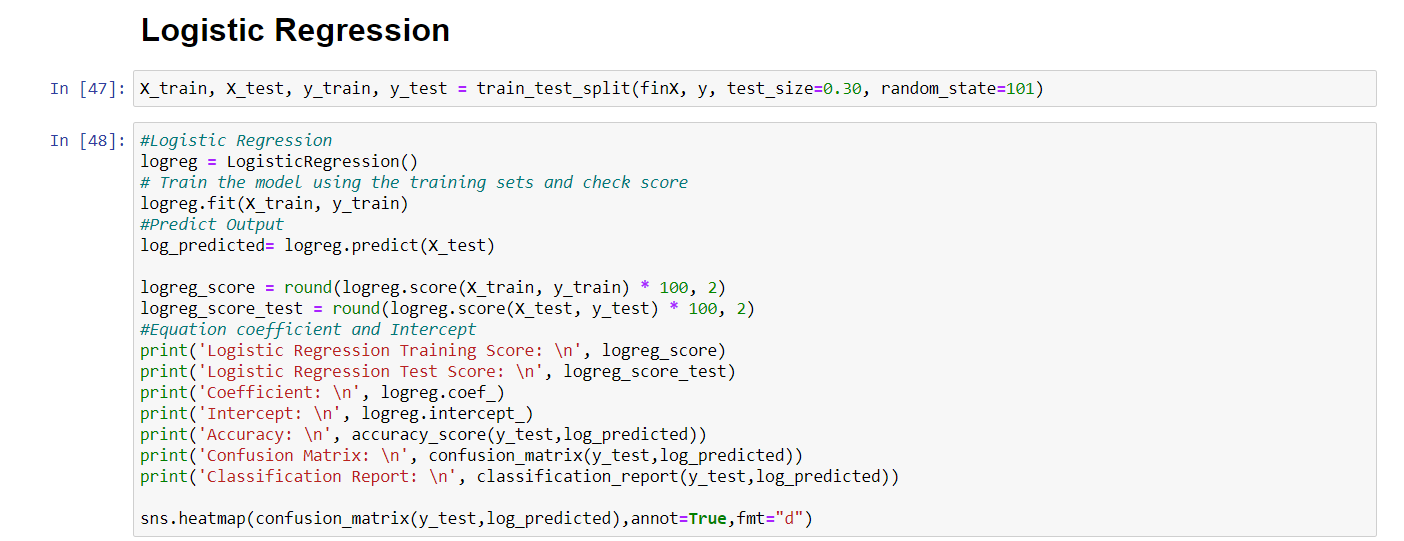


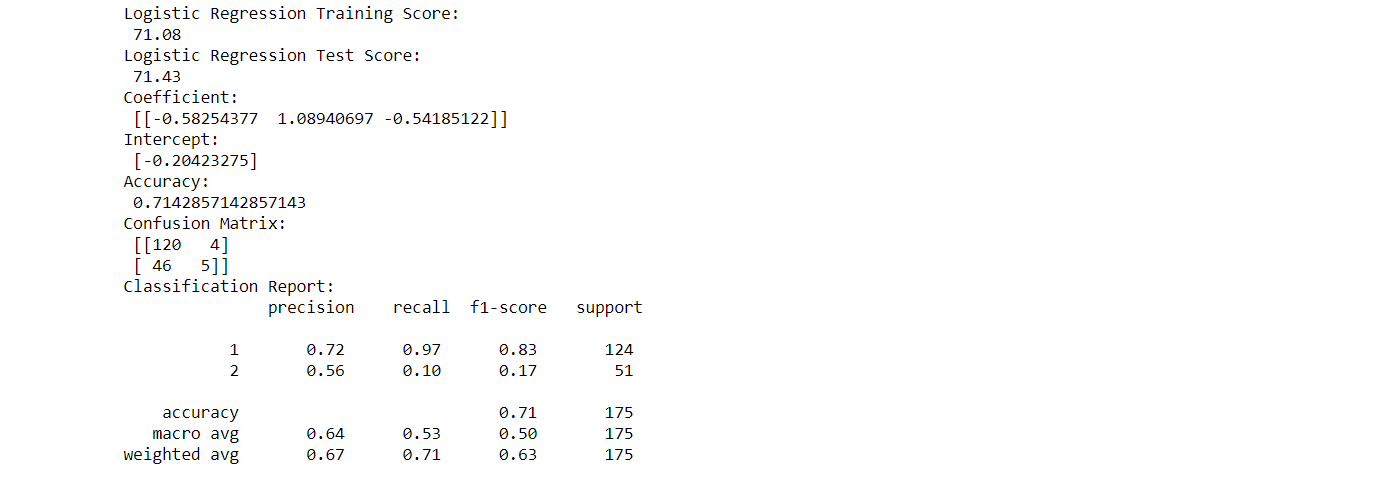


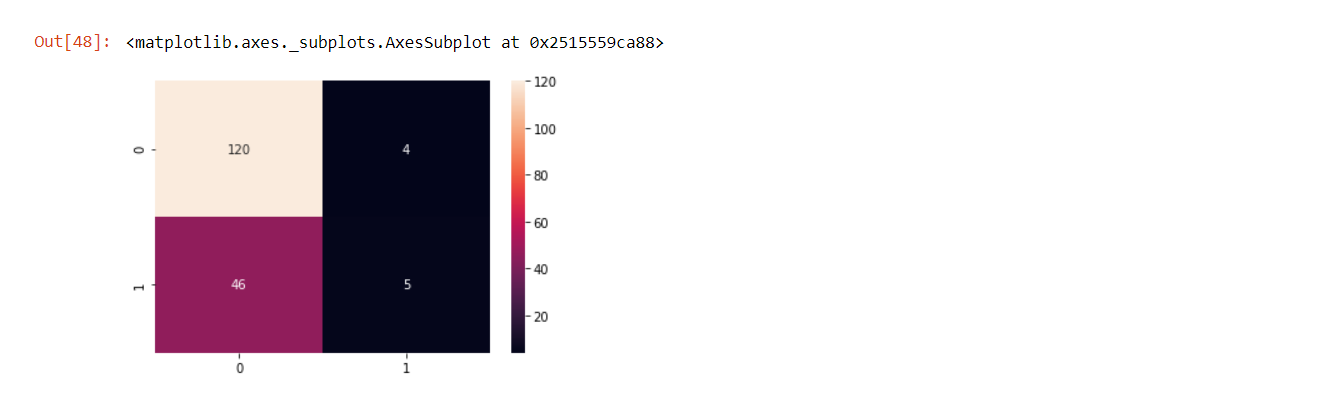
**Linear Regression:**







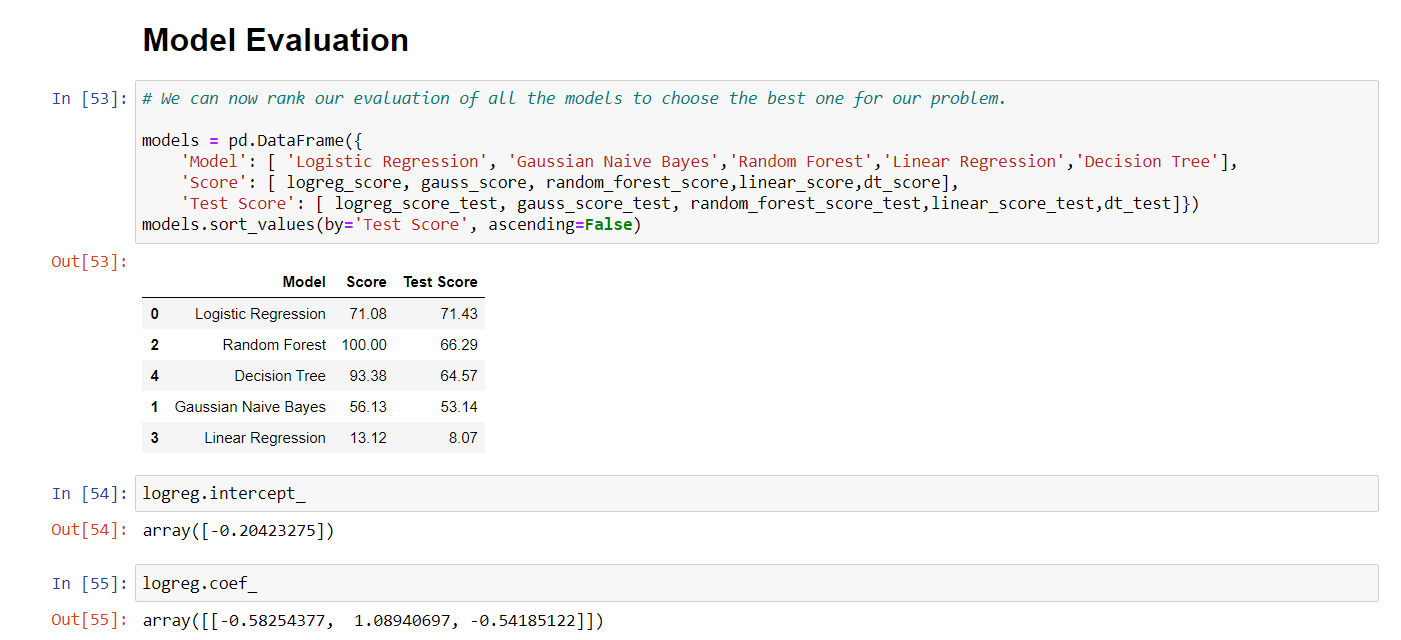




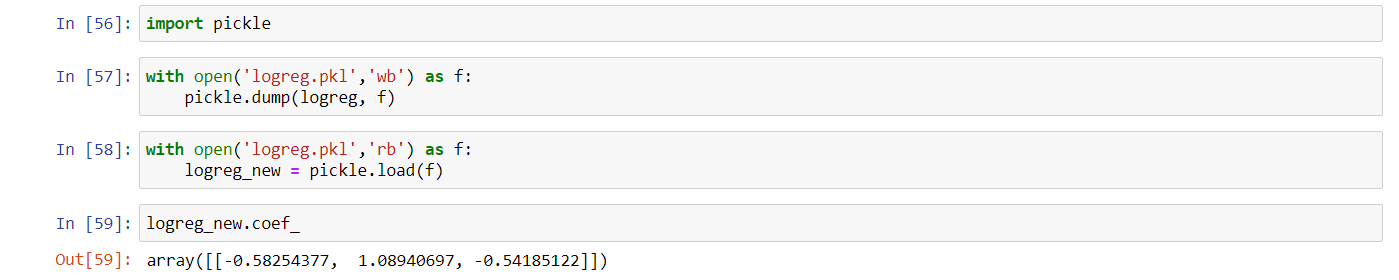
**Decission Tree Classifier**:



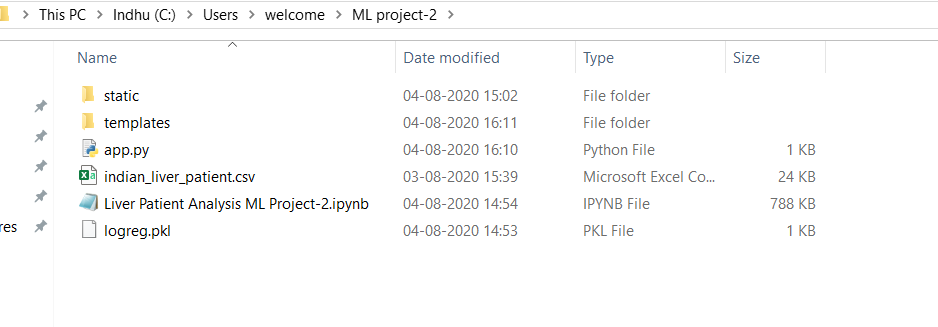
**Model Evaluation:**



**Saving the model using pickle:**

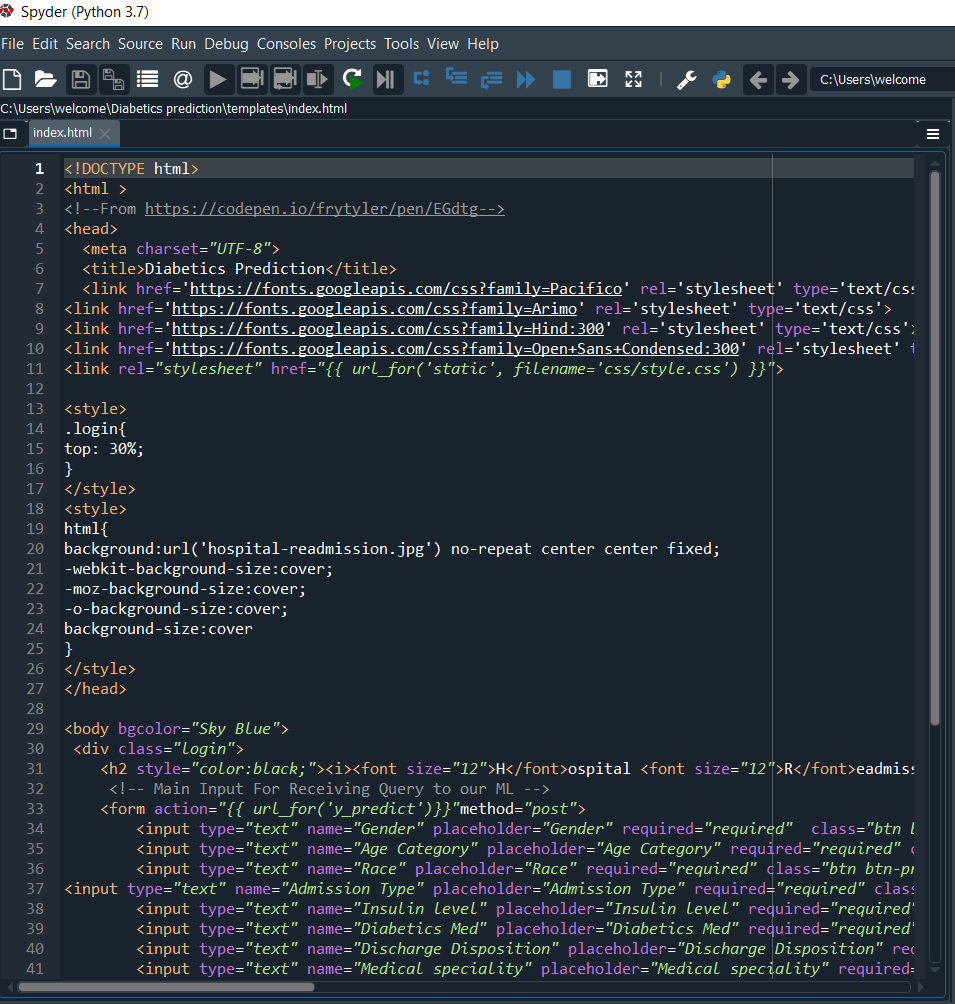


**File structure:**

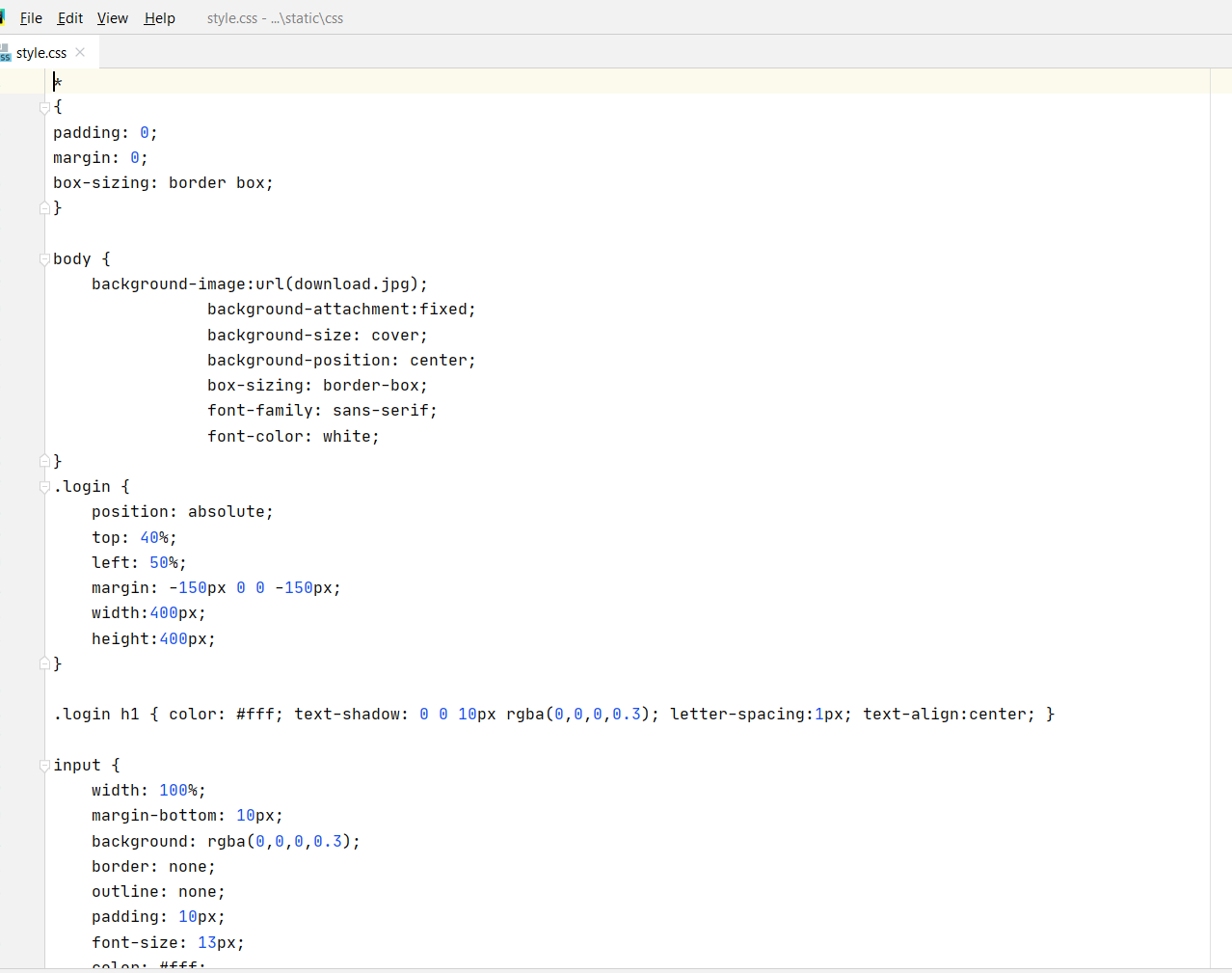


**Building HTML Page:**

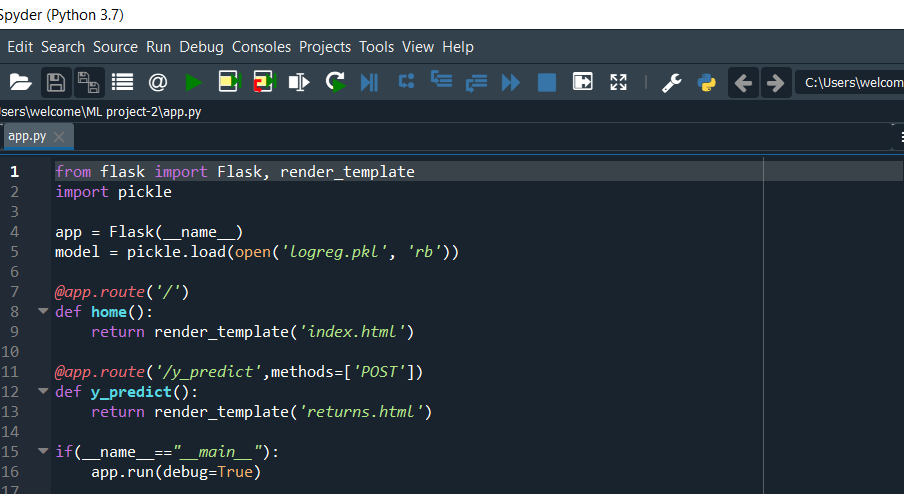
This is the basic HTML page for our project. H1 tag is used to give heading to the project. The user has to enter the details of and the output will be predicted in such a way that the particular patient has a liver disease or not.



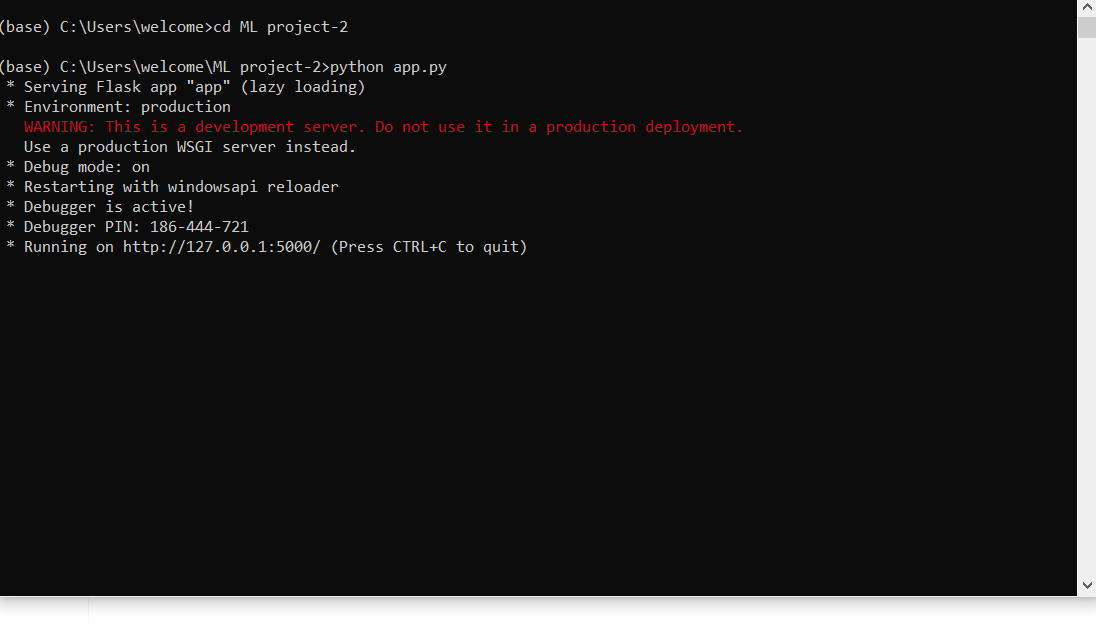
**CSS Code:**



**Python Code:**



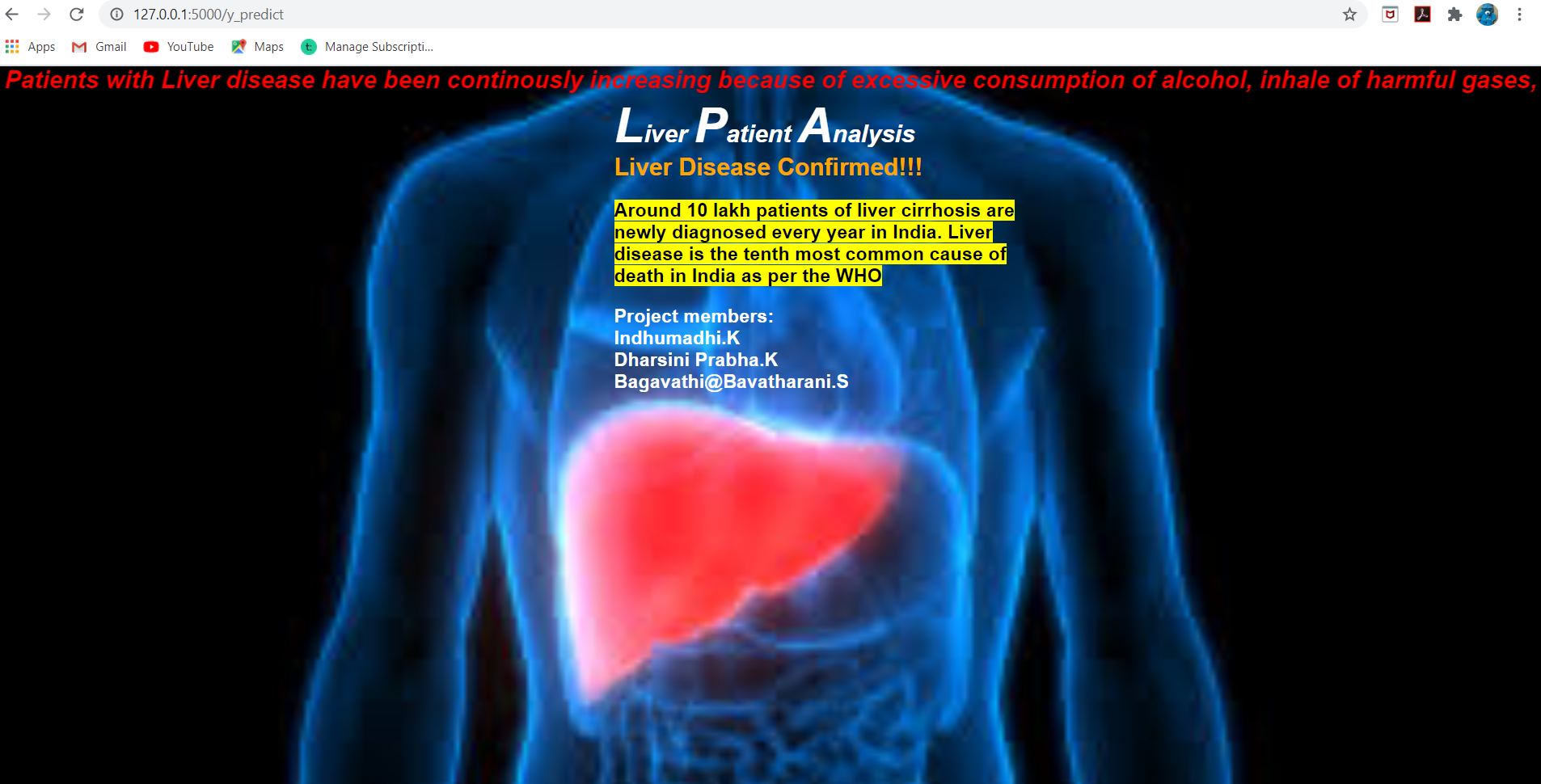
**Prompt:**



**Webpage:**



**Output Screen:**



**Conclusion:**

In this project, we have proposed methods for diagnosing liver disease in patients using machine learning techniques. The five machine learning techniques that were used include Logistic Regression, Gaussian Naive Bayes, Random Forest, Linear Regression and Decision Tree Classifier. The system was implemented using all the models and their performance was evaluated. Performance evaluation was based on certain performance metrics. Logistic Regression was the model that resulted in the highest accuracy with an accuracy of 71.43%.

**References**:

* + www.kaggle.com
  + www.geeksforgeeks.com
  + www.github.com
  + towardsdatascience.com