PES UNIVERSITY, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE18CS203

B.Tech, Sem III
Session : Aug-Dec, 2019

UE18CS203 – INTRODUCTION TO DATA SCIENCE

REPORT
ON
EXPLORATORY ANALYSIS ON
INDIAN LIVER PATIENT RECORDS

SECTION : A

#	SRN	Name	Contact No.	Email ID	Sign
1	PES1201800110	TEJAS SRINIVASAN	9686502395	tejas.srinivasan007@gmail.com	
2	PES1201800328	JOSEPH DOMINIC	6361914117	jd.cherukara@gmail.com	

ABOUT THE DATA SET

Describe the dataset (its purpose, meaning of all the variables), Provide Data Set Size.

The CSV file provided contained 583 rows and 11 columns and contained 4 missing values in Albumin_and_Globulin_Ratio column.

The type of columns and a ***brief description*** of them have been provided below -

1) Age: Age of the patient
2) Gender: Gender of the patient
3) Total Bilirubin: Liver function parameter which is a sum of direct and indirect bilirubin.
4) Direct Bilirubin : Liver Protein.
5) Alkaline Phosphotase: Liver Protein.
6) Alamine Aminotransferase: Liver Protein.
7) Aspartate Aminotransferase: Liver Protein.
8) Total Proteins: Sum of basic liver protein values.
9) Albumin: Liver Protein.
10) Albumin and Globulin Ratio: ratio of albumin and globulin in the blood.
11) Dataset: Has a value of 1 for patients with liver disease and 0 for patients with no liver disease.

ABSTRACT

The purpose of this assignment is to perform exploratory data analysis, plot visualizations on and derive meaningful insights from the dataset on Indian Liver Patient Records.

After performing the required data cleaning (ensuring that there were no missing values or outliers present), the analysis was performed on the dataset. The basic questions given in the dataset such as “Are H - males above 45 are more susceptible than females above 45?” , “Predict whether the patient has liver disease” were performed using the relevant visualizations and regressions. Further the questions on “Which attribute had the highest correlation?” and “Which column can be dropped and why?” can be answered using the correlation matrix and the associated heatmap plotted.

After obtaining some of the basic insights, further visualizations and predictions were made , which are explained in detail before.

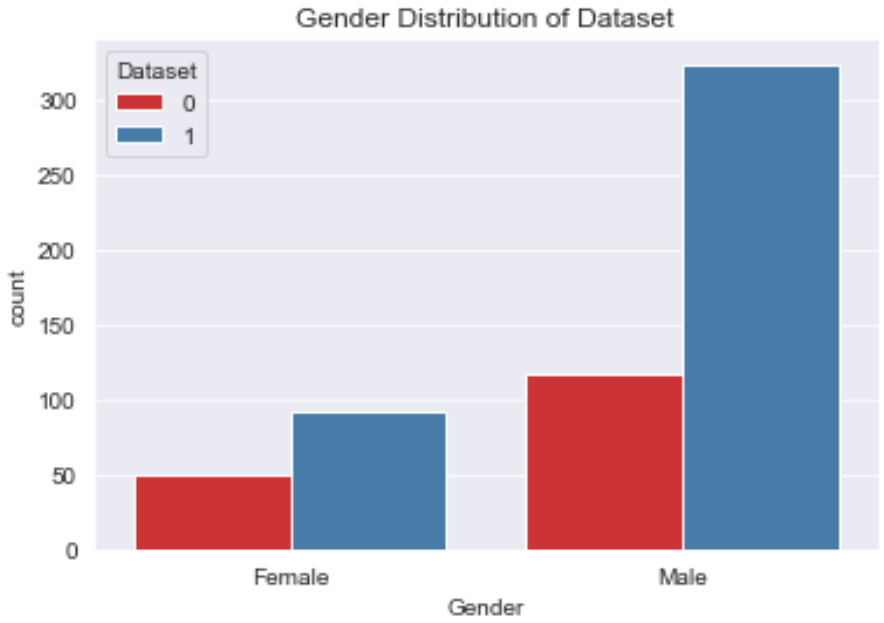
The conclusions obtained have been listed in the appropriate section of the report.

EXPLORATORY ANALYSIS

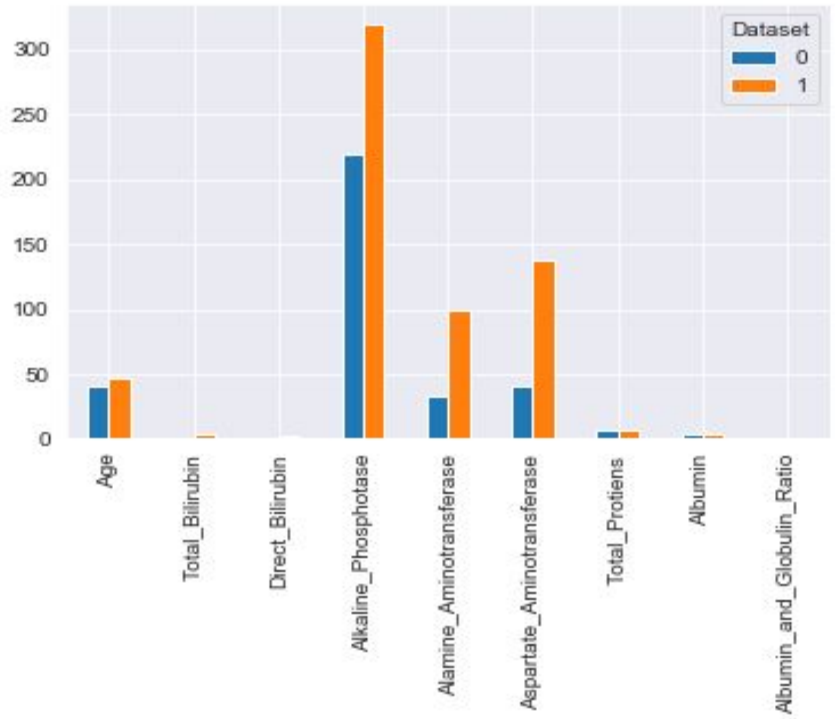
Data Cleaning:

1) Only 4 values were missing from the Albumin_and_Globulin_Ratio column.
2) The column did not have enough NA/missing values to drop so the missing values were replaced instead.
3) These values were filled using interpolation.
4) Since there is no accepted medical definition of an "outlier", the entire range of values in the dataset was considered.
5) The categorical values (2) in the "Dataset" column were changed to 0s for easier understanding and visualization of the column.

Visualisations:

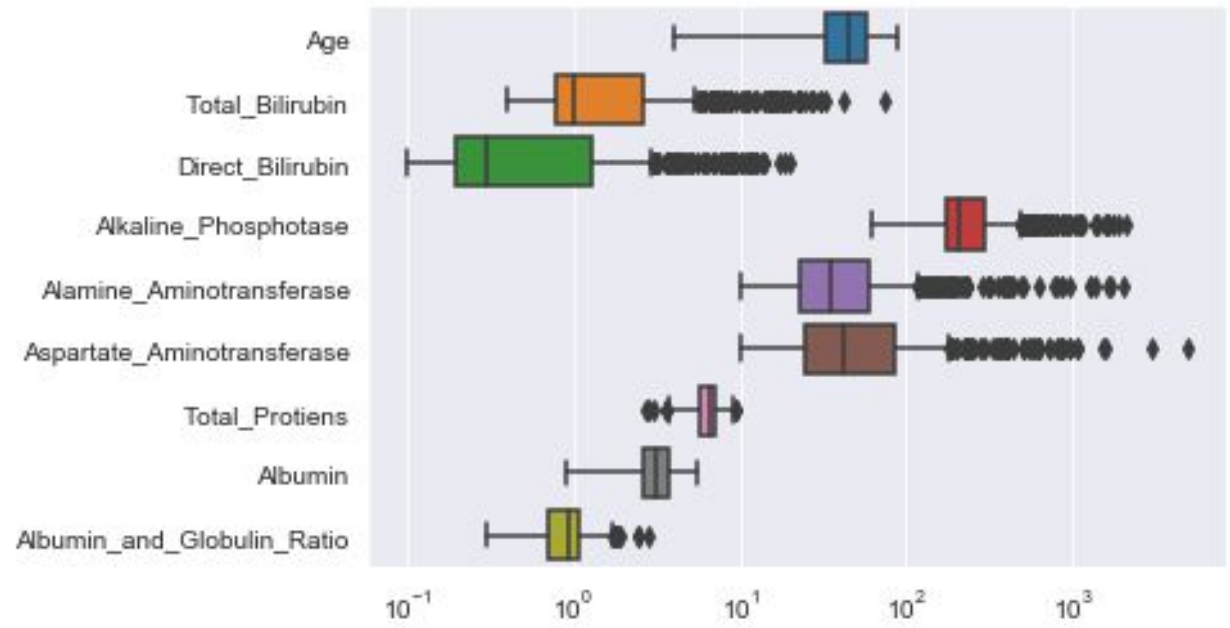


Plotted the distribution of patients of each gender. It is evident that the number of males in the dataset is more than the number of females in the dataset. It can also be inferred that the proportion of liver disease affected males is more than the proportion of liver disease affected females.

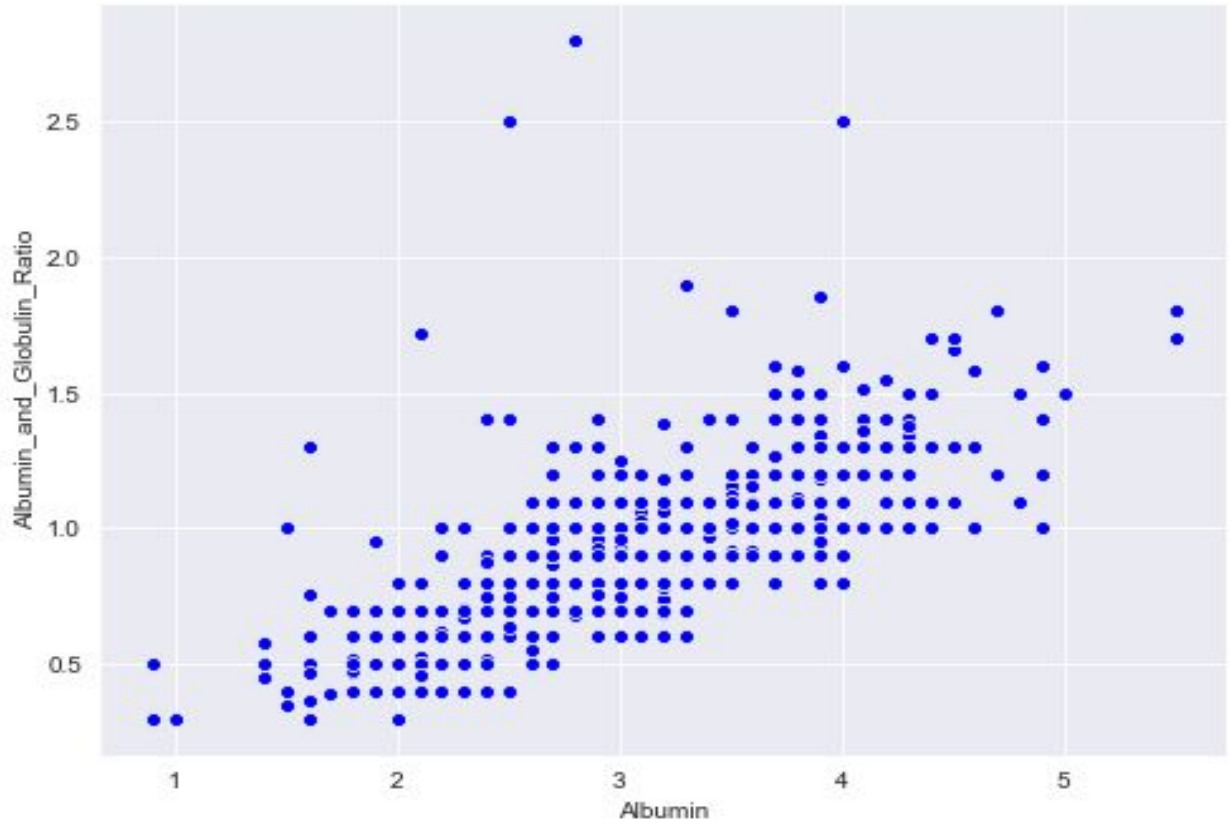


Here the mean values for every column were plotted for healthy and unhealthy people respectively. It is evident that unhealthy people have a higher average value for every column.

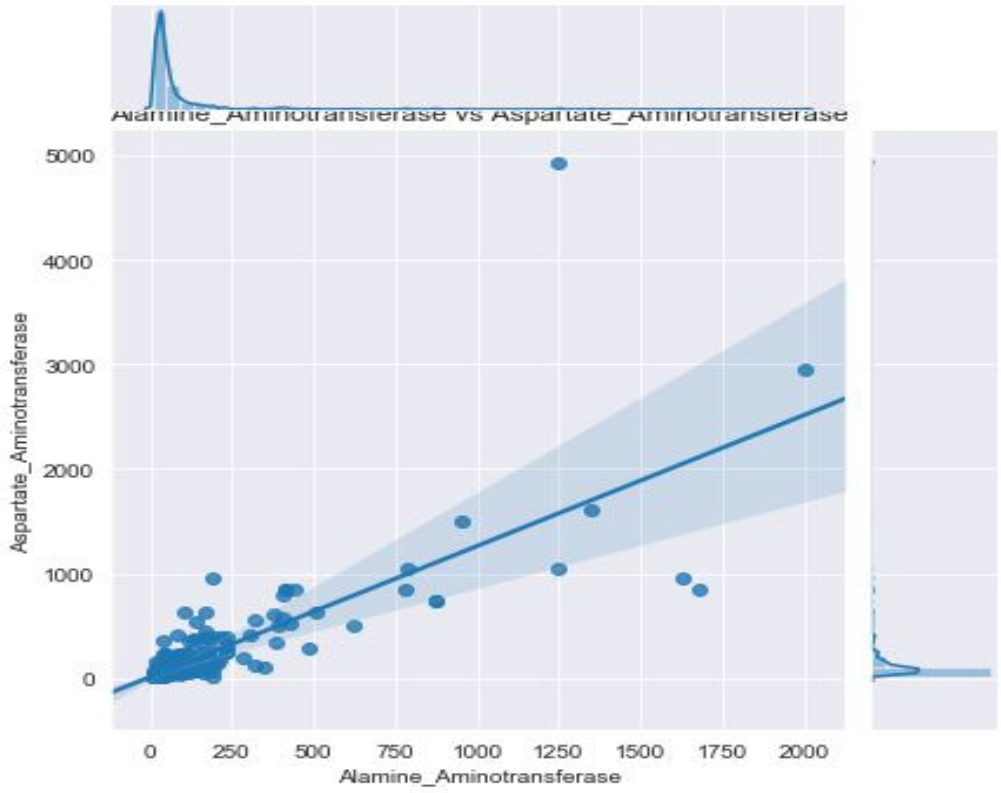
.



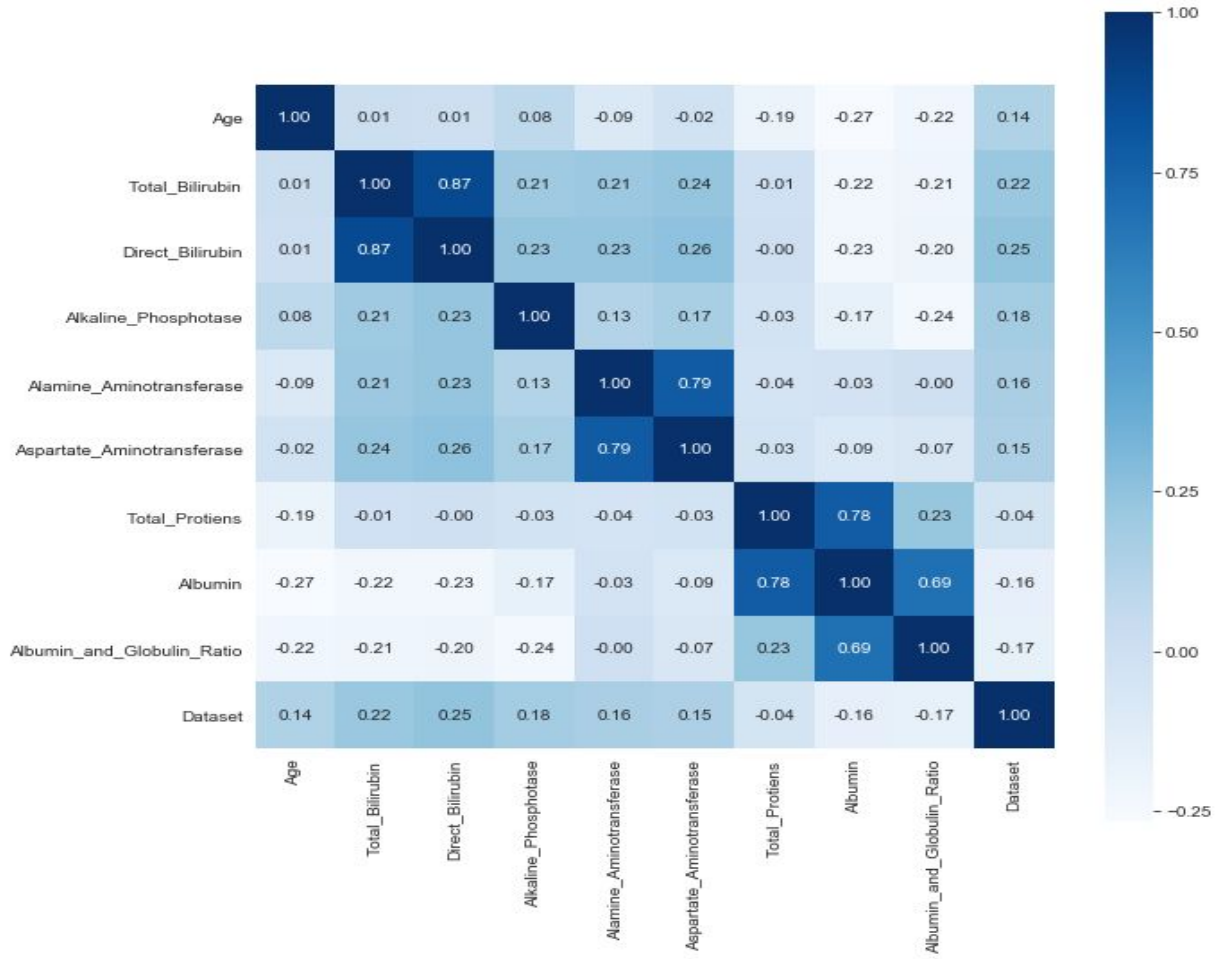
Plotted the box plots for every column on the logarithmic scale to be able to see the outliers.



Plotted a scatter plot of albumin globulin ratio vs albumin.



From the graph it is evident that Alanine_aminotransferase is right skewed and Aspartate aminotransferase is left skewed. The two parameters are linearly related.



The above graph is the correlation matrix for each of the columns in the dataset. The Direct bilirubin has the highest correlation with the dataset column. Age has the least correlation with the dataset column.

Predicti -

X_train and y_train were assigned as the feature and the target variables.

1) **Using Linear Regression** - The linear model was fitted and a round of the linear score was found. The test score was calculated using the appropriate linear_score function and the coefficient (Beta 0) and the intercept (for the best fit line was found).

2) **Using Logistic Regression** - The logistic regression function was plotted using the appropriate sklearn function and the test score, intercept and coefficient were found as before. Furthermore metrics such as F1 score, precision and recall scores were calculated and plotted in a table as shown in the notebook.

A confusion matrix was also constructed based on this and can be found in the ipython notebook.

Clearly, the binary logistic regression approach is better than the linear regression approach since there are only two categorical variables ('Dataset' is 0 or 1).

3) **Using Random Forest Classifier** - A random forest classifier is basically a model that consists of multiple decision trees, and it take random subsets of random samples of training data , which leads to lowering the overall variance of the tree, while not compensating on the increase of bias.

   This method was found to give better accuracies, and was hence chosen.

   The random_forest test score and accuracies were found out and on printing the confusion matrix is was found that the number of True positives had drastically increased from that of logistic regression.

A comparison of the different approaches taken for prediction -

Out[311]:

	Model	Score	Test Score
2	Random Forest	100.00	68.57
1	Logistic Regression	73.77	66.86
0	Linear Regression	13.14	8.00

Hypothesis Test:

The Hypothesis we tested was: Men above 45 are more susceptible to liver disease than Women above 45.
H0 : Proportion of affected men above 45 - Proportion of affected women above 45 <=0
H1 : Proportion of affected men above 45 - Proportion of affected women above 45 > 0
Used the Chi – Square Test.
Rejected H0. Implying Men above 45 are more susceptible than women above 45.

CONCLUSION:

Using the various visualizations, predictions and hypotheses testing methods, the following conclusions can be drawn -

1) The different variants of the proteins of the same type (Eg. Aspartate and Alanine Transferase and Total and Direct Bilirubin are linearly correlated - as shown by the jointplots).
2) The skewness of the data towards men (in plots such as Gender vs Total_Bilirubin and Gender vs Albumin) is due to a higher number of men in the dataset, as shown in the barplot.
3) The 'Albumin_and_Globulin_Ratio' column has the highest correlation (about 0.64) with the 'Dataset' column. The 'Age' column has the lowest correlation with 'Dataset' , (0.013) and can be dropped if need be.
4) Prediction of Liver Disease has been been performed using Linear Regression, Logistic Regression and Random Forest and it was found that Random Forest gave the best accuracy since it takes a model subset of the features instead of all of them.
5) From the hypothesis test, it has been concluded that Men above 45 are more susceptible to liver disease than Women above 45.

Further conclusions drawn can be found in the IPython Notebook.