**LIVER PATIENT DATA ANALYSIS**

**1.1 INTRODUCTION**

Computer Science today infiltrates all other branches to make life easier by using its computational power and storage power. A very powerful tool used these days is Python, which is a high-level programming language with an Object Oriented Programming approach. Python also supports structured programming and is very useful because of its capacity of visualisation. A tool must implement a strategy and such a strategy is Artificial Intelligence.  “Artificial intelligence" is used to describe machines/computers that mimic "cognitive" functions that humans associate with other [human minds](https://en.wikipedia.org/wiki/Human_mind), such as "learning" and "problem solving". In this project, we have used “Machine Learning” which is a subset of Artificial Intelligence.

**1.2 OBJECTIVES OF RESEARCH**

The objective is analyse the parameters, which affect liver health and build a model that takes input variables to classify a patient. To predict whether a patients’ liver is healthy or unhealthy based on certain parameters or inputs.

**1.3 PROBLEM STATEMENT**

The aim is to find the causes of liver diseases and build prediction model that predicts the outcome and gives suggestions for the benefit of the patients’ health. The solution to this is analysis and visualisation of data to derive inference and finally build a model using machine learning and build a UI for use by a practitioner.

**2. REVIEW OF LITERATURE**

The dataset reveals the various factors that affect liver health. Supervised learning is used for the dataset. There are 9 parameters that collectively affect the liver health. There are 583 rows of data of various patients. The various factors are age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alamine aminotransferase, aspartate aminotransferase, total proteins, albumin and lastly albumin and globulin ratio. The dataset has displays the output in the form of 1 and 2, where 1 means an unhealthy liver condition and 2 means a healthy liver condition. The dataset is pre-processed, normalised and then various classification techniques are applied for the prediction model that is built.

**3. DATA COLLECTION**

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). This data set contains 441 male patient records and 142 female patient records.

Any patient whose age exceeded 89 is listed as being of age "90".

Columns:

* Age of the patient
* Gender of the patient
* Total Bilirubin
* Direct Bilirubin
* Alkaline Phosphotase
* Alamine Aminotransferase
* Aspartate Aminotransferase
* Total Protiens
* Albumin
* Albumin and Globulin Ratio
* Dataset: field used to split the data into two sets (patient with liver disease, or no disease)

**4. METHODOLOGY**

The various techniques used are mentioned in this section. It includes pre-processing, visualising, normalising, applying classification techniques to find best accuracy.

**4.1 EXPLORATORY DATA ANALYSIS**

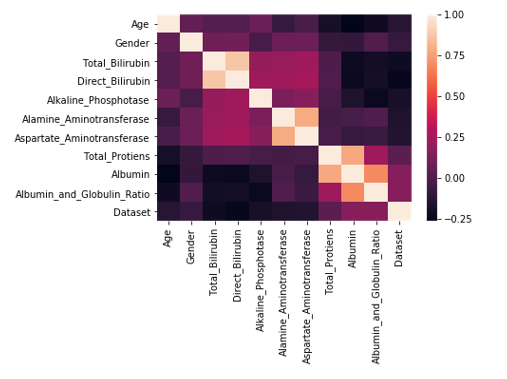
The data contains 583 rows where the age varies from 4 years to 90 years.

The dataset reveals the total bilirubin content which is 0.4 in the minimum case and 75.0 in the maximum case. The median and various other quartiles are also found out for all the columns.

The first quartile defines 25% of data, second quartile defines the 50% or median and the third quartile defines the 75% of the dataset.

The null values are encountered by replacing them with the mean of the entire column.

The correlation between various columns is found out and a heat map is plotted for the same. From the correlation table and the heat map we see that the column “Gender” has the weakest correlation between the other inputs and with the output.



Heat map of the correlation between features

1 – strong positive correlation

1 <x< 0.5 – good positive correlation

0.5 <x< 0 – weak positive correlation

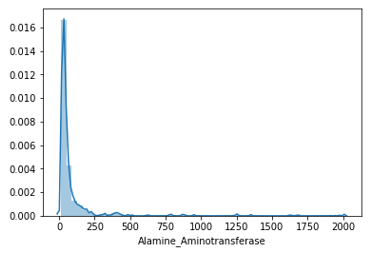
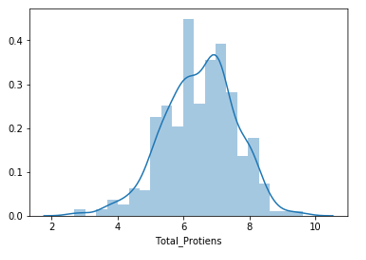
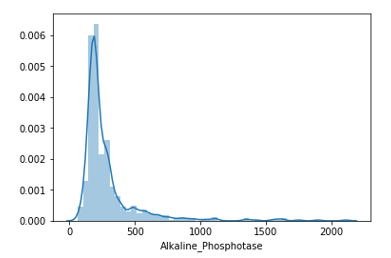
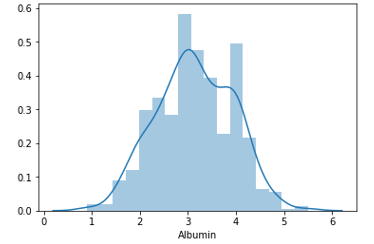
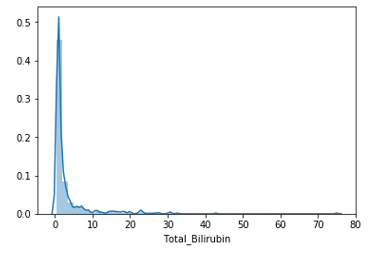
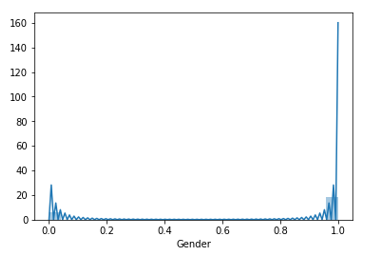
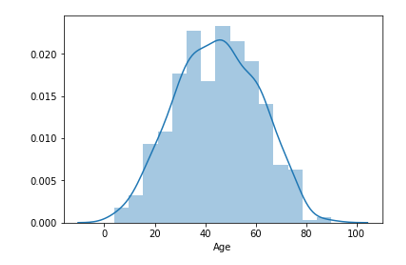
0< x < -0.5 – weak negative correlation

-0.5 < x < -1 - good negative correlation

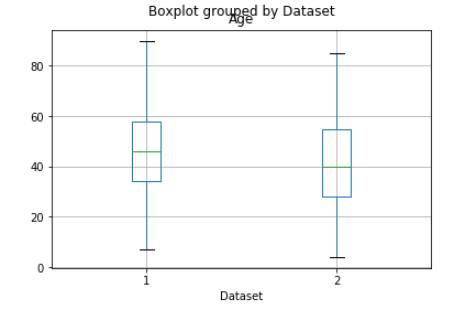
-1 – strong negative correlation

Distplot

The distplot depicts if a feature is normally distributed over a range of values. The distplot for the different features is as follows. From the distplot we conclude if the curve is parabolic or curvilinear for the features.



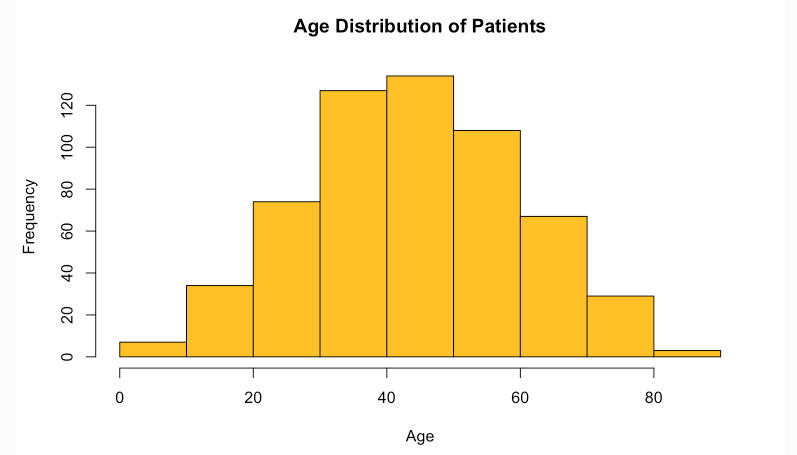
We also plot the box plot for age to the median and see if there are any ouliers. We find no outliers. The box plot also gives the 3 quartiles. The box plot is as follows-



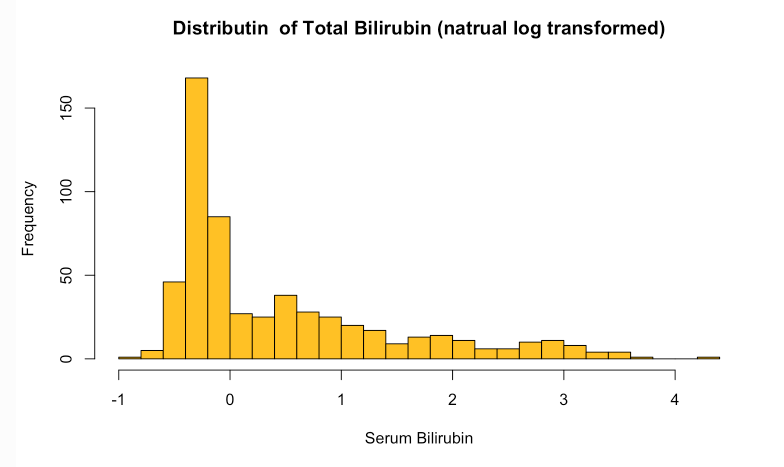
Box plot of age

Column wise frequency distribution

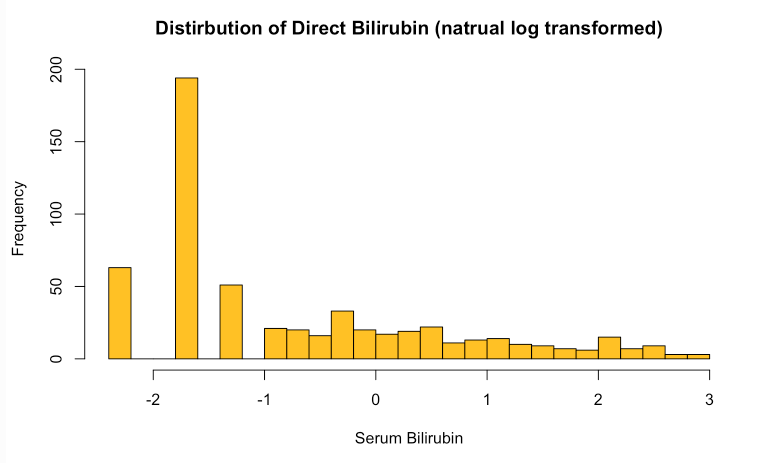
🡺Age



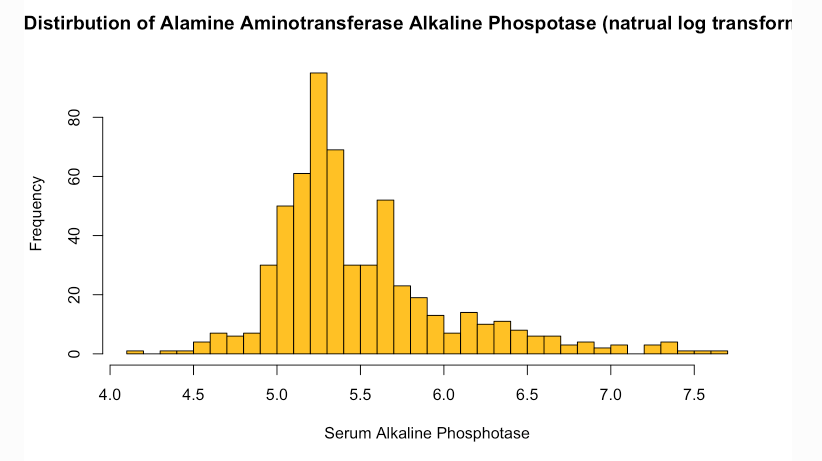
🡺Total Bilirubin



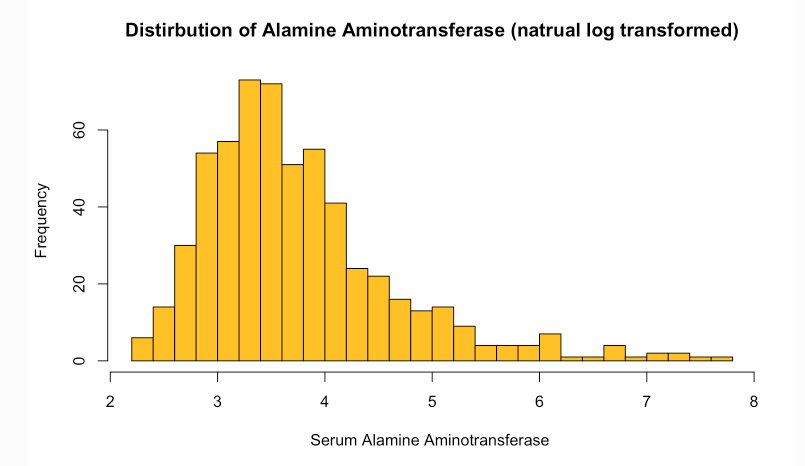
🡺Direct Bilirubin



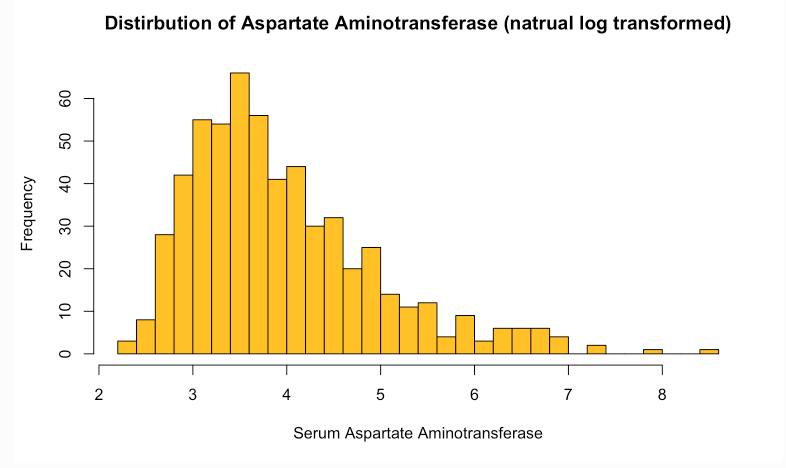
🡺Alkaline Phosphotase



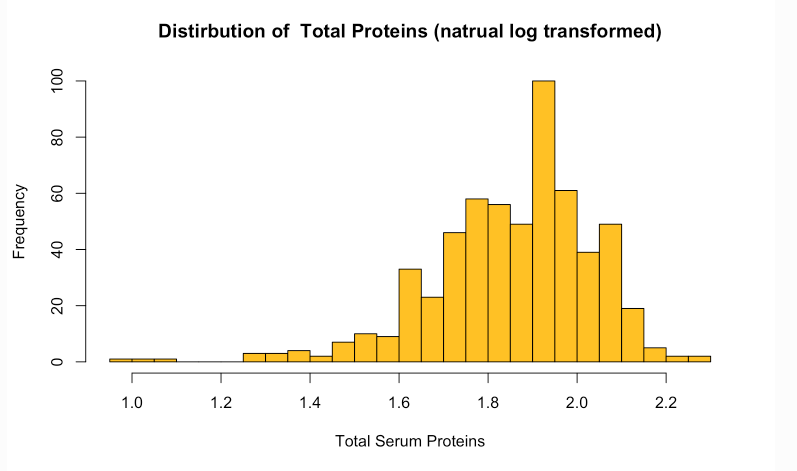
🡺Alamine Aminotransferase



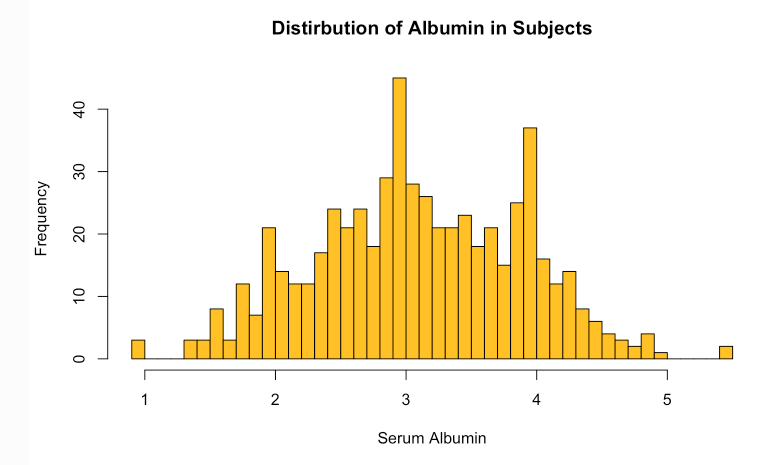
🡺Aspartate Aminotransferase



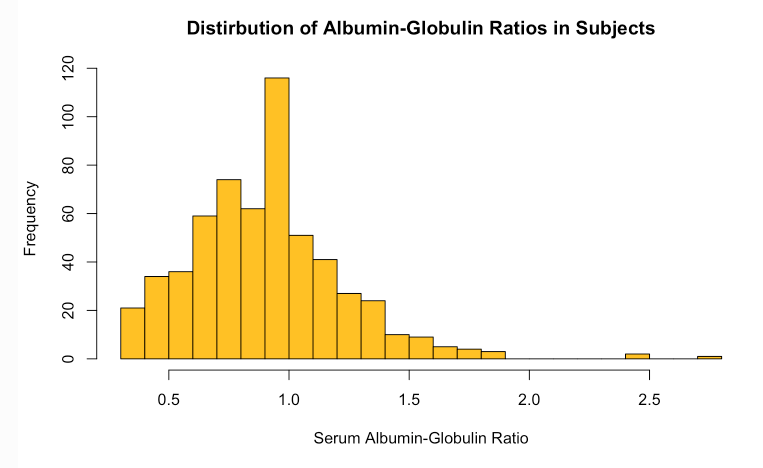
🡺Total Proteins



🡺Albumin



🡺Albumin and Globulin Ratio



**4.2 DATA MODELLING**

For data modelling, kNN Classifier, Logistic Regression, Decision Tree, Random Tree classifier, Support Vector Classifier and Naive Bayes is used. Out of all these, Support vector give the best accuracy.

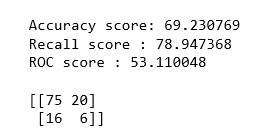
**ROC curve** represents a sensitivity/specificity pair corresponding to a particular decision threshold.

**Recall Score** expresses the ability to find all relevant instances in a dataset

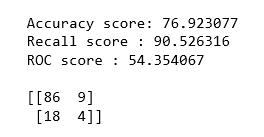
The Confusion Matrix and Accuracy scores are given below for all of the above Techniques. A confusion matrix is made up as follows-

|  |  |
| --- | --- |
| True Positive | False Positive |
| False Negative | True Negative |

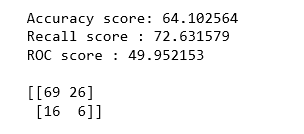
🡺kNN Classifier



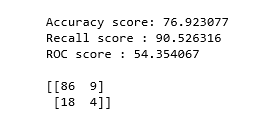
🡺Logistic Regression



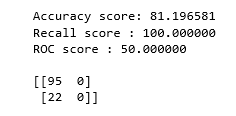
🡺Decision Tree Classifier



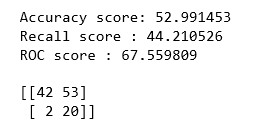
🡺Random Forest Classifier

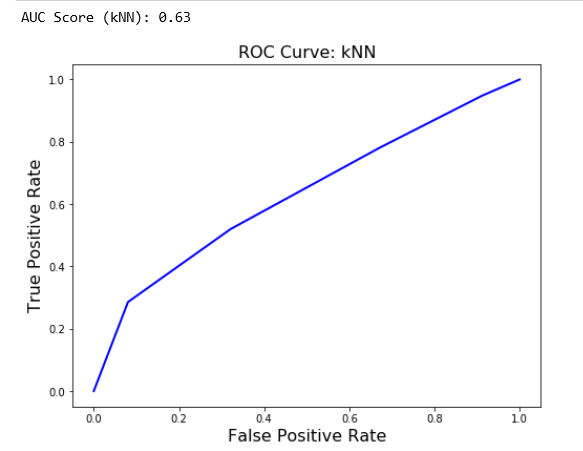


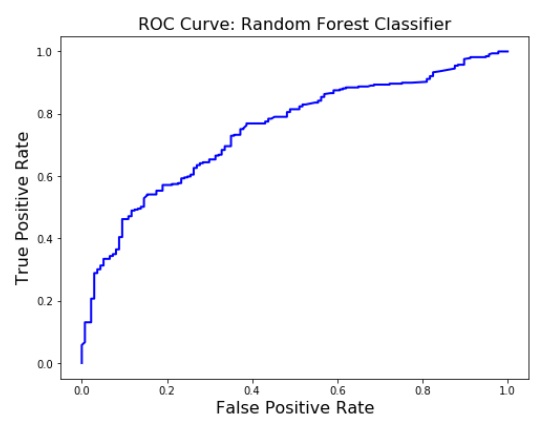
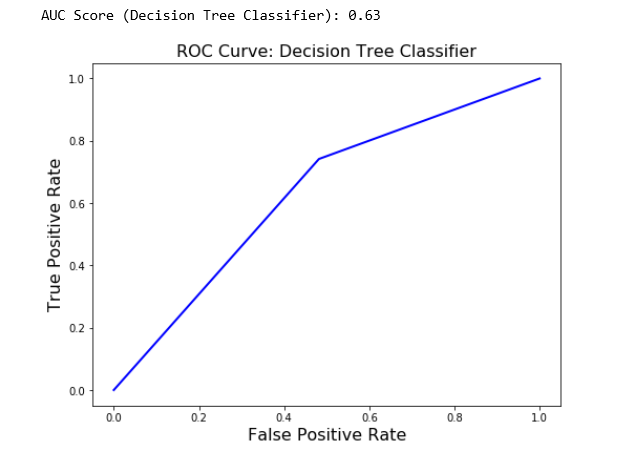
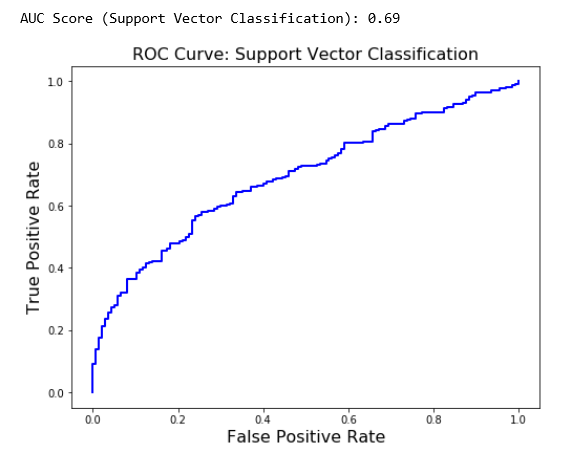
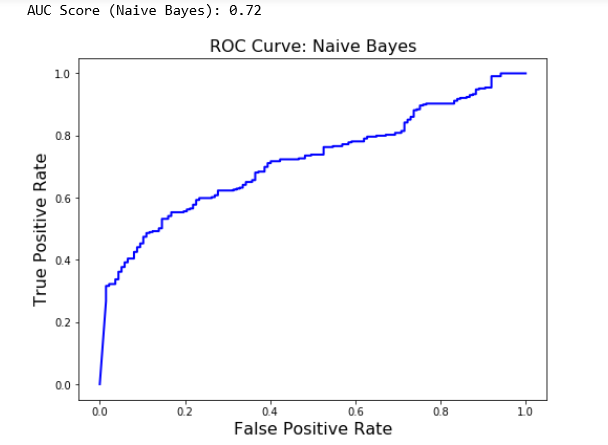
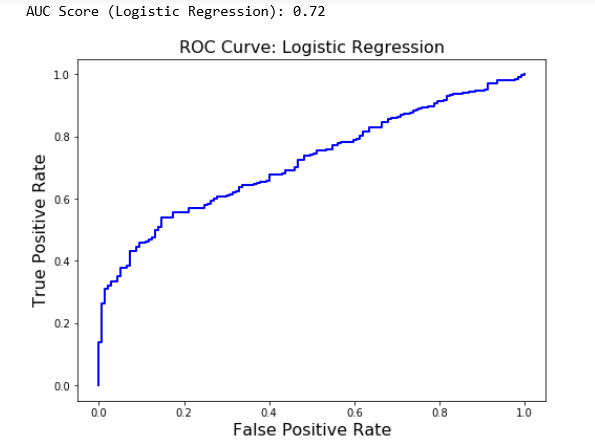
🡺Support Vector Classifier



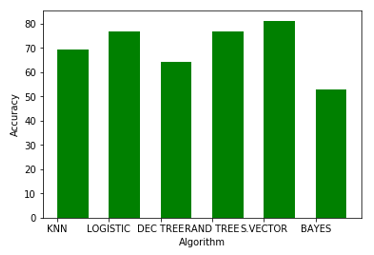
🡺Naive Bayes



The ROC AUC curves are as follows for the different models – 



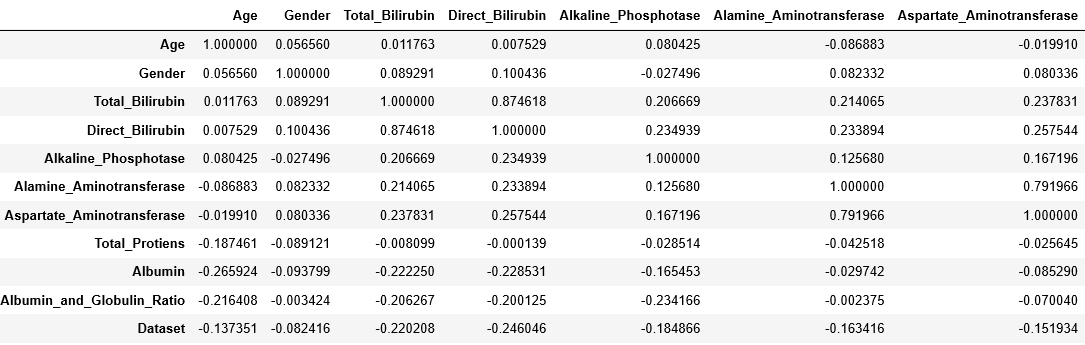
The vivid comparison of accuracy scores of all the models is done through Bar Graph. The Bar Graph is as follows –



**5. FINDINGS AND SUGGESTIONS**

It is observed that the models that perform well for the given dataset are Logistic Regression, Random Tree Classifier and Support Vector Classifier. It is further observed that the Support Vector Classifier performs the best among the rest.

From the correlation table –



It is observed that Gender doesn’t play a major role and hence it can be dropped while doing the analysis. The correlation is very weak.

It is found that children aged below 10 have comparatively lesser unhealthy liver.

The main causes of this liver ailment can be excessive use of oil, egg which increases albumin. The increased consumption of alcoholic drinks also play a major factor and that’s why more middle aged people from 30-60 suffer from liver problems.

**6. CONCULSION**

This type of model might be useful in early detection, while not diagnosis because of the accuracy rate.

The model can be really helpful with racially more diverse training data set. We can also derive the further effect of Gender on the model.

Given the accuracy is 81% on the validation set, with further tuning of this model on more validation data, it might be useful as a tool to suggest further definitive testing of liver disease for patients who this model suggests might be positive.

**7. BIBLIOGRAPHY**

* [www.kaggle.com](http://www.kaggle.com)
* UCI Machine Learning Repository: ILPD (Indian Liver Patient Dataset) Data Set. (n.d.). Retrieved December 8, 2017, from <https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)>
* Liver Function Test Levels

<https://www.emedicinehealth.com/liver_blood_tests/article_em.htm#what_are_the_symptoms_of_abnormal_levels_of_liver_enzymes>