CIRRHOSIS PREDICTION

PROJECT 10

By:
Priyanka Chandramohan

1. Introduction

Cirrhosis is a late stage of scarring (fibrosis) of the liver caused by many forms of liver diseases and conditions, such as hepatitis and chronic alcoholism.

Main Goal:

➔ Create an analytical framework to understand
Key factors impacting Cirrhosis
➔ Develop a modeling framework
To predict whether the person contain cirrhosis or not

About Dataset

The dataset is related to different test values related to human body. The dataset describes several popularity and description metrics their effect on its quality. The dataset is of Shape (418,20). The datasets can be used for classification or regression.As the dataset is very long it cannot be displayed in a frame.
It has 418 rows and 20 columns . Here is the list of columns this dataset has,
df.columns

```
Index(['ID', 'N_Days', 'Status', 'Drug', 'Age', 'Sex', 'Ascites',
       'Hepatomegaly', 'Spiders', 'Edema', 'Bilirubin', 'Cholesterol',
       'Albumin', 'Copper', 'Alk_Phos', 'SGOT', 'Tryglicerides',        'Platelets','Prothrombin',
'Stage'],dtype='object')
```

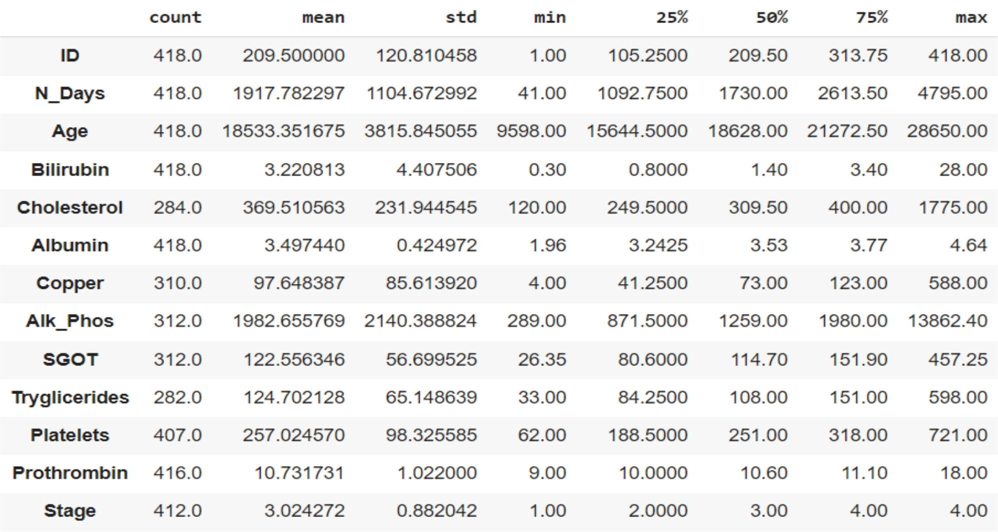
The dataset consists of following columns :

- ID: unique identifier
- N_Days: number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986
- Status: status of the patient C (censored), CL (censored due to liver tx), or D (death)
- Drug: type of drug D-penicillamine or placebo
- Age: age in [days]
- Sex: M (male) or F (female)
- Ascites: presence of ascites N (No) or Y (Yes)
- Hepatomegaly: presence of hepatomegaly N (No) or Y (Yes)
- Spiders: presence of spiders N (No) or Y (Yes)
- Edema: presence of edema N (no edema and no diuretic therapy for edema), S (edema present without diuretics, or edema resolved by diuretics), or Y (edema despite diuretic therapy)
- Bilirubin: serum bilirubin in [mg/dl]
- Cholesterol: serum cholesterol in [mg/dl]
- Albumin: albumin in [gm/dl]
- Copper: urine copper in [ug/day]
- Alk_Phos: alkaline phosphatase in [U/liter]
- SGOT: SGOT in [U/ml]
- Triglycerides: triglicerides in [mg/dl]
- Platelets: platelets per cubic [ml/1000]
- Prothrombin: prothrombin time in seconds [s]
- Stage: histologic stage of disease (1, 2, 3, or 4)

2. EDA and Business Implication

EDA stands for exploratory data analysis where we explore our data and grab insights from it. EDA helps us in getting knowledge in form of various plots and diagrams where we can easily understand the data and its features.

Analysis of the dataframe



	count	mean	std	min	25%	50%	75%	max
ID	418.0	209.500000	120.810458	1.00	105.2500	209.50	313.75	418.00
N_Days	418.0	1917.782297	1104.672992	41.00	1092.7500	1730.00	2613.50	4795.00
Age	418.0	18533.351675	3815.845055	9598.00	15644.5000	18628.00	21272.50	28650.00
Bilirubin	418.0	3.220813	4.407506	0.30	0.8000	1.40	3.40	28.00
Cholesterol	284.0	369.510563	231.944545	120.00	249.5000	309.50	400.00	1775.00
Albumin	418.0	3.497440	0.424972	1.96	3.2425	3.53	3.77	4.64
Copper	310.0	97.648387	85.613920	4.00	41.2500	73.00	123.00	588.00
Alk_Phos	312.0	1982.655769	2140.388824	289.00	871.5000	1259.00	1980.00	13862.40
SGOT	312.0	122.556346	56.699525	26.35	80.6000	114.70	151.90	457.25
Tryglicerides	282.0	124.702128	65.148639	33.00	84.2500	108.00	151.00	598.00
Platelets	407.0	257.024570	98.325585	62.00	188.5000	251.00	318.00	721.00
Prothrombin	416.0	10.731731	1.022000	9.00	10.0000	10.60	11.10	18.00
Stage	412.0	3.024272	0.882042	1.00	2.0000	3.00	4.00	4.00

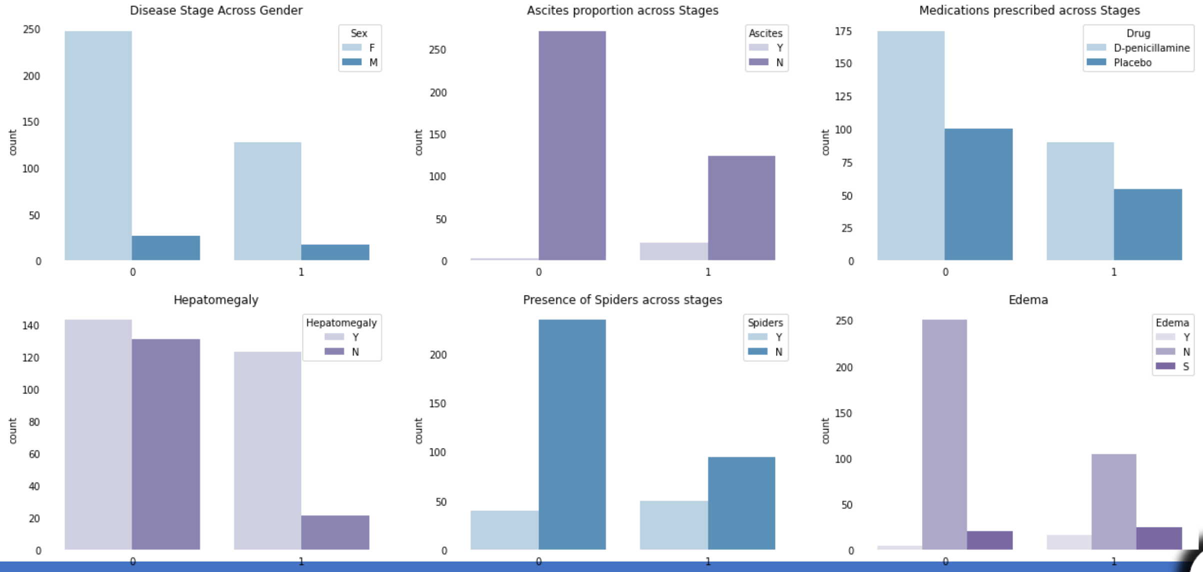
3. Data Cleaning

Handling Missing Values

This is a problem, we could just get rid of all examples with NA values, but in this case our case of small dataset we cannot afford that. We will impute the missing entries with some statistical calculations. **We have two different types of data:**

Numerical data (Age, Cholesterol, Platelets.. etc) Categorical Data (Drug, Sex, Spiders..etc) We will have to use different imputation for each type. For the numerical type we can use mean or median. In this case we will go with median to avoid skewing in the presence of outliers For Categorical type we will impute the most frequent class.

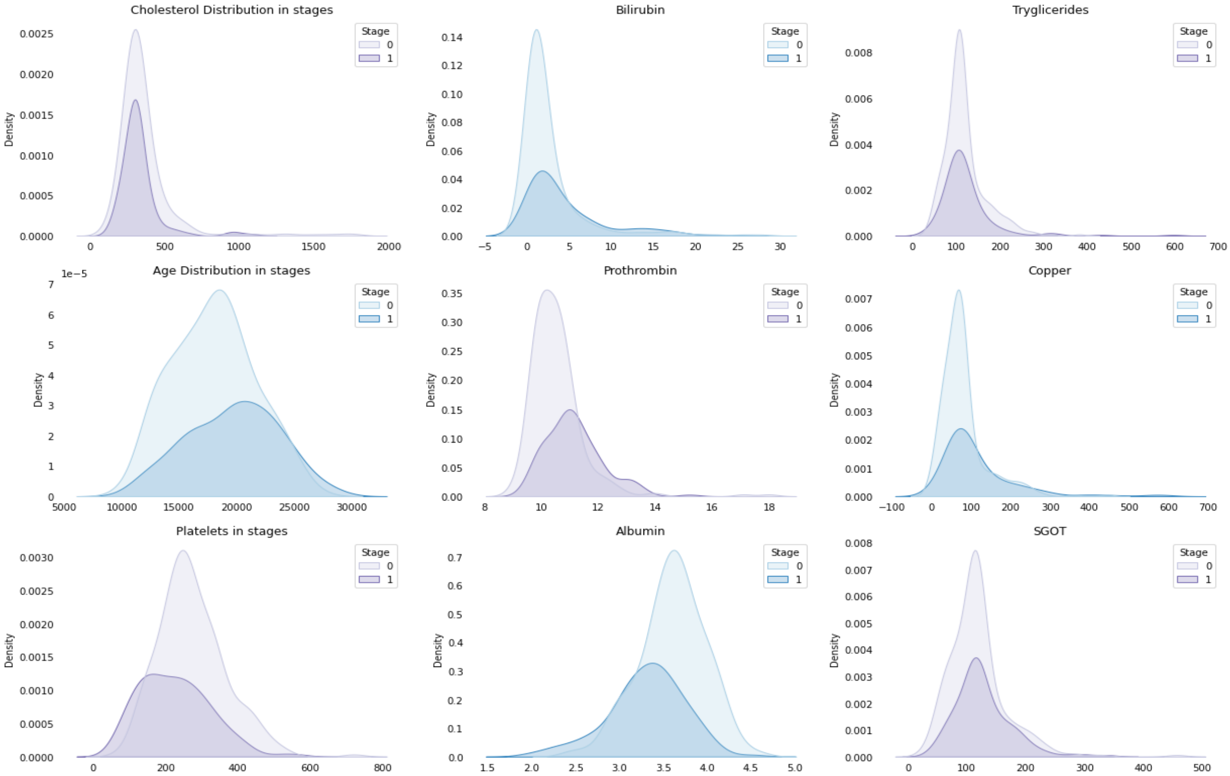
Plots for Analysis



There are some interesting insights if we observe closely. Take the case at Ascites, we observe that the rist of disease is higher with increase in Ascites. also presence of spiders has a positive relation with disease risk.

In **Disease Stage Across Gender** plot we can observe female are high. Coming to the medications prescribed across stage uses two types of medicines like D-penicillamine ad Placebo and that to D-penicillamine is mostly prescribed for female compared to male. Hepatomegaly and Spiders are common sign of liver disease and coming to our data we can clearly observe that Hepatomegaly and Spider are quite opposite when one is increases another is decreased

Distribution Plots

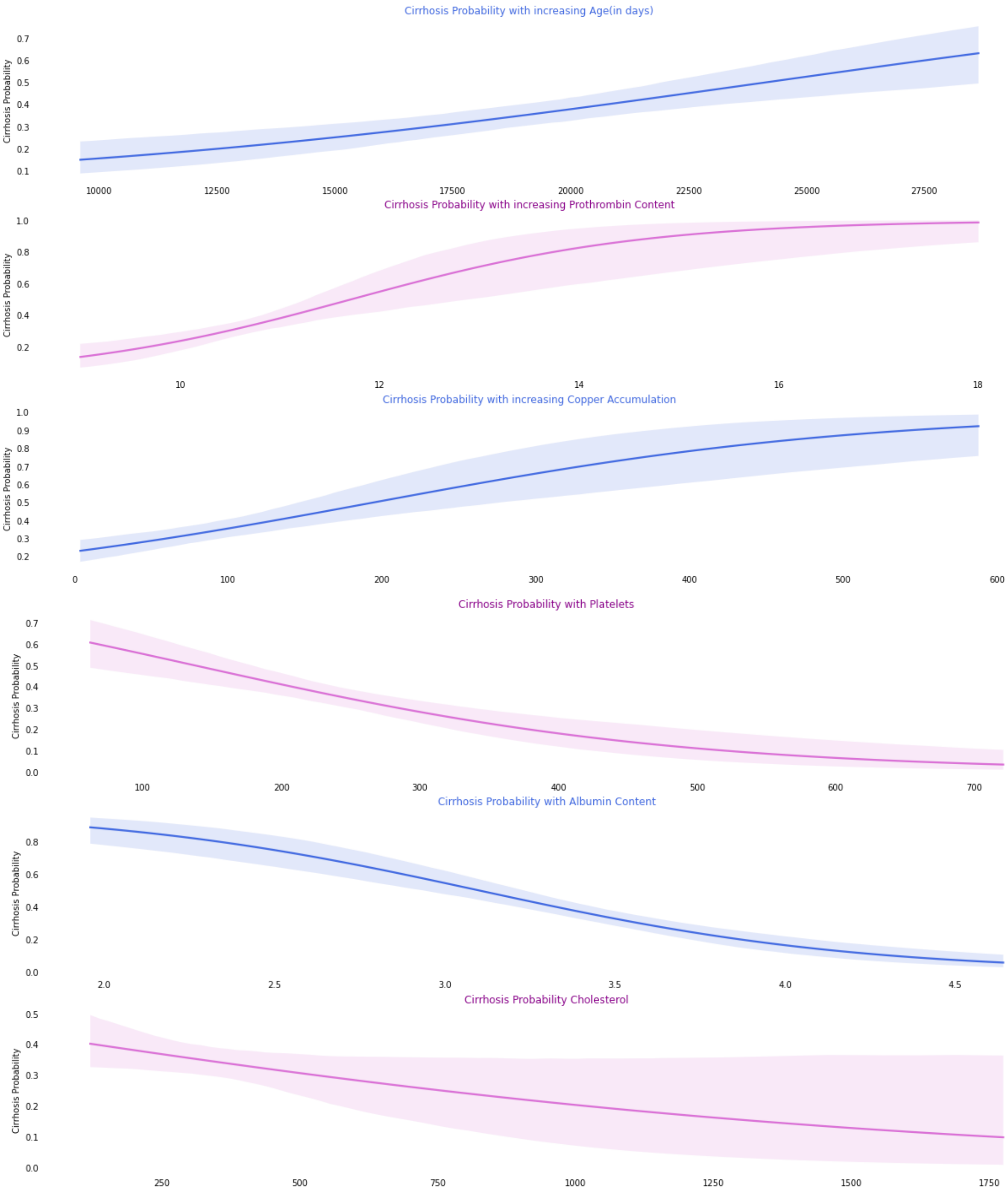


Looking at the feature distribution we can observe that in features such as Age, Prothrombin, Copper the risk of the disease increase with increase in feature value, thus having a positive co-relation on with the disease probability. Lets fit a regression line to check.

Regression

Looks like the data checks with our intuition. These parameters indeed increase the risk of the disease. We can also observe some features such as Platelets, Albumin, Cholesterol where the probability of disease decrease with increase in feature value. Lets tally that with some more regression plots.

Platelets, Albumin checks with our logic the findings about Cholesterol seems interesting! Looks like people with high Cholesterol have lower risk of Cirrhosis, this might not sound correct but our data certainly shows so. This should help our model predict the target. We will e looking at what features contribute the most in later part of the project.



Cirrhosis Probability with increasing Age(in days)

Cirrhosis Probability with increasing Prothrombin Content

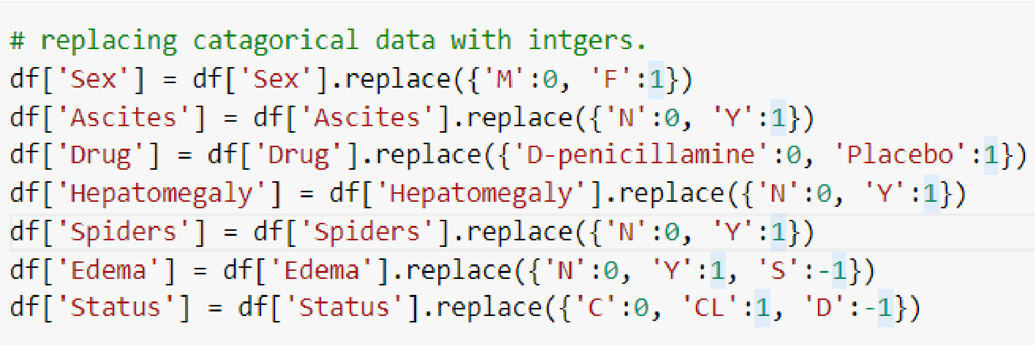
Cirrhosis Probability with increasing Copper Accumulation

Cirrhosis Probability with Platelets

Cirrhosis Probability with Albumin Content

Cirrhosis Probability Cholesterol

Preprocessing



```
# replacing catagorical data with intgers.
df['Sex'] = df['Sex'].replace({'M':0, 'F':1})
df['Ascites'] = df['Ascites'].replace({'N':0, 'Y':1})
df['Drug'] = df['Drug'].replace({'D-penicillamine':0, 'Placebo':1})
df['Hepatomegaly'] = df['Hepatomegaly'].replace({'N':0, 'Y':1})
df['Spiders'] = df['Spiders'].replace({'N':0, 'Y':1})
df['Edema'] = df['Edema'].replace({'N':0, 'Y':1, 'S':-1})
df['Status'] = df['Status'].replace({'C':0, 'CL':1, 'D':-1})
```

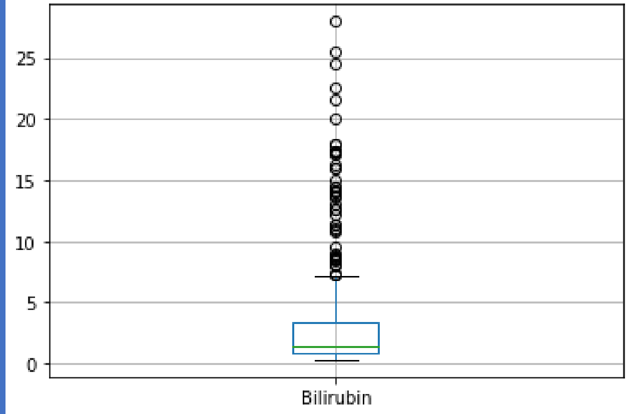
# replacing catagorical data with intgers.
 # Male : 0 , Female :1
 # N : 0, Y : 1
 # D-penicillamine : 0, Placebo : 1
 # N : 0, Y : 1
 # N : 0, Y : 1
 # N : 0, Y : 1, S : -1
 # 'C':0, 'CL':1, 'D':-1

Showing the all the unique elements and their count by Using the value_counts function for all the features

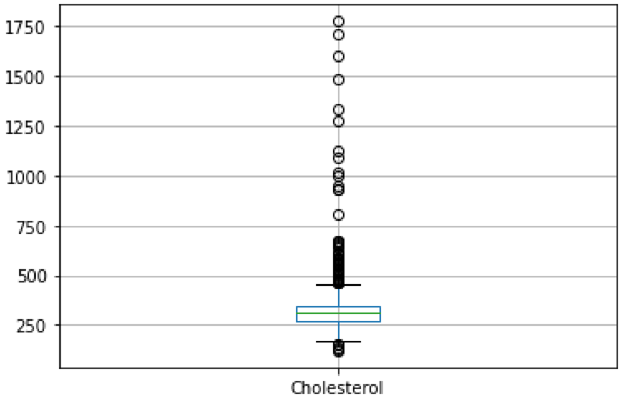
df['Column_name'].value_counts()

Uni-variate Analysis - By BoxPlot

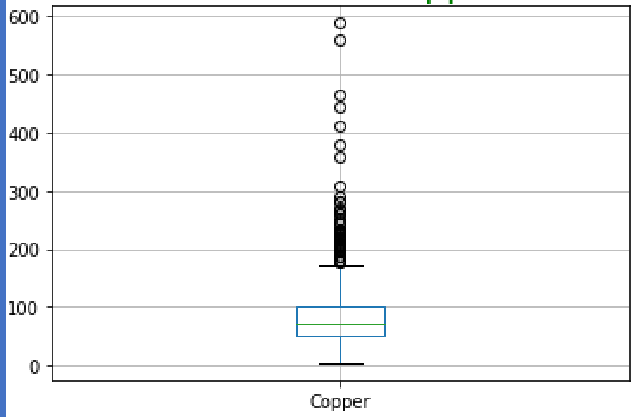
Box Plot Of Bilirubin



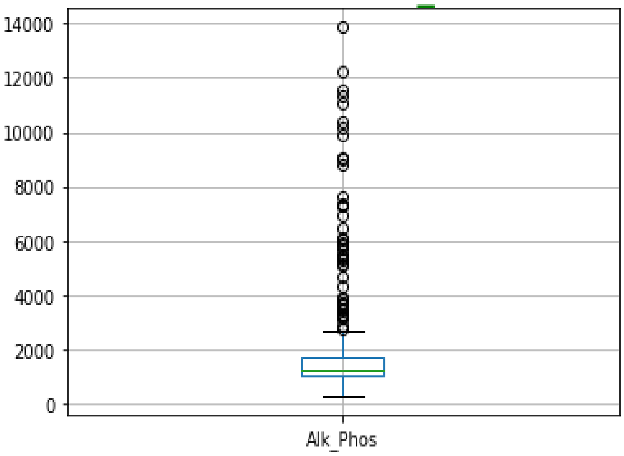
Box Plot Of Cholesterol



Box Plot Of Copper



Box Plot Of Alk_Phos



Observation:

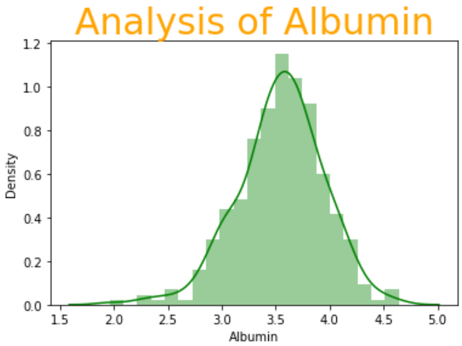
After observing all the box plots there some columns which have outliers they are:

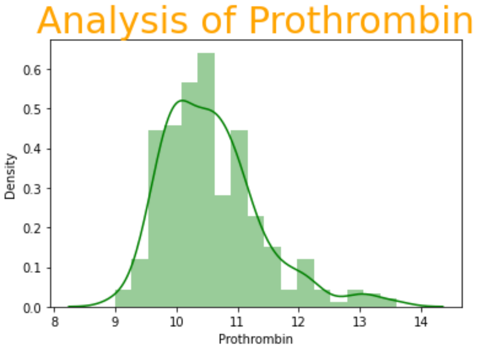
➔ Sex,Ascites,Spiders,Edema,Bilirubin,Cholesterol,Albumin,Copper,Alk_Phos,SGOT,Tryglicerides,Platelets,Prothrombin

Removing outliers for the above columns:

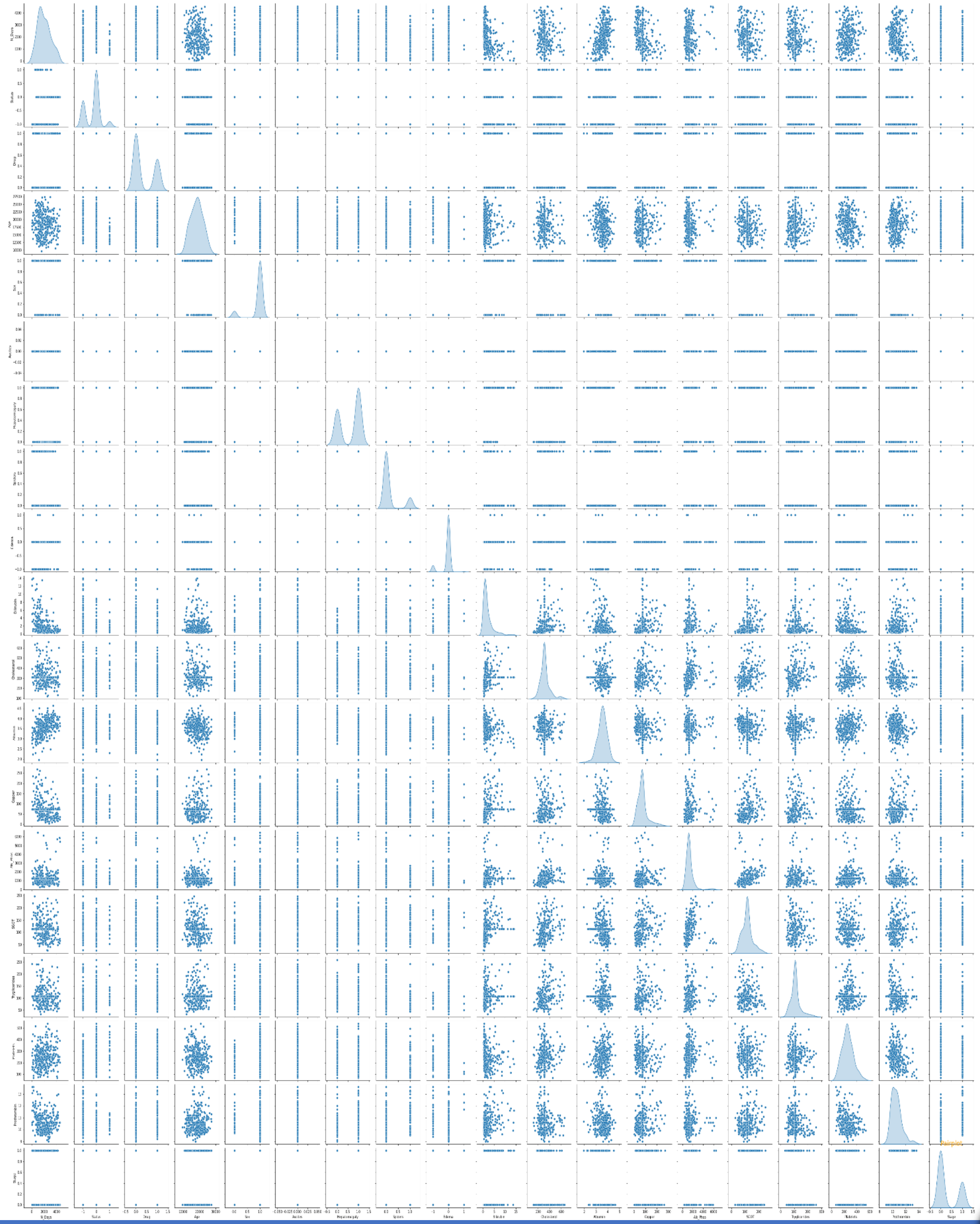
After removing the all the outliers ,I have done the dist plot analysis on the features.

Dist Plots





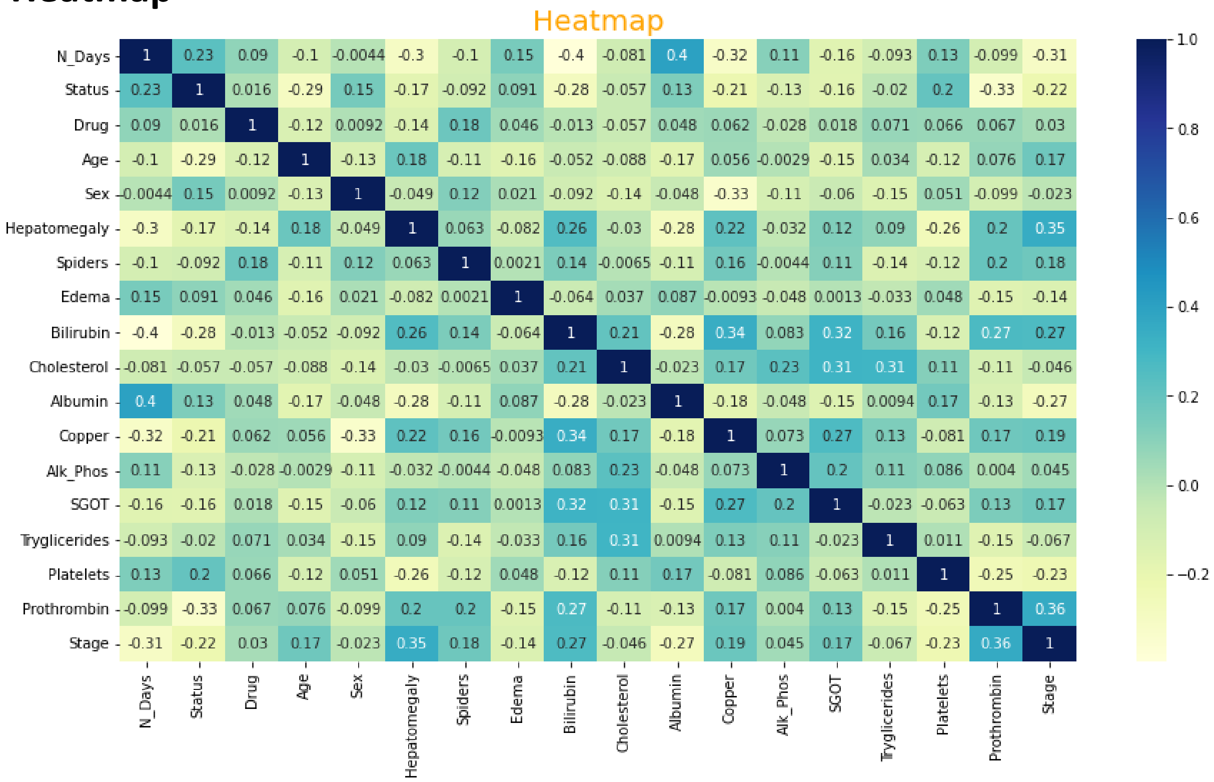
Pair Plot



By pair plot analysis we can see that Ascites feature doesn’t show any variation.We removed “Ascites”

Bivariavte Analysis

Heatmap



After performing EDA we can observe that Ascites feature doesn't show any variation so we can remove that feature as it is an entry level there is no need to take any other additional data.

Making the data flexible to train the machine learning model

4. Model Building

Model Selection and Why?

After cleaning and processing the data then comes the modelling part which includes building Machine Learning models, let’s first understand in brief what Machine Learning is?

Machine Learning is a technique that analyses past data and tries to extract meaningful insights and patterns from them which can be further used to perform predictions in future. For example, classifying whether a tumor is benign or malignant, predicting stock prices, etc. One such application which we’re using right here is predicting Cirrhosis. Before making predictions first we need to build a model and train it using past data.

First, we need to separate the dataset into two parts: features (Columns) and labels (Stage) which is the required format for any model to be trained on.

Then the data needs to be split into 3 sets

1. Training set - This will be the part of the dataset which the model will be using to train itself, the size should be at least 60-70% of the total data we've.

2. Validation set - This set is used for validating our model's performance for a different set of hyperparameters. After taking out the train set, the remaining set can be split into validation and test set.

3. Testing set - To evaluate how the model is performing on the unseen data on which the model will be doing future predictions on, test set is used. It helps to understand how much error is there between actual and predicted values.

Splitting the data

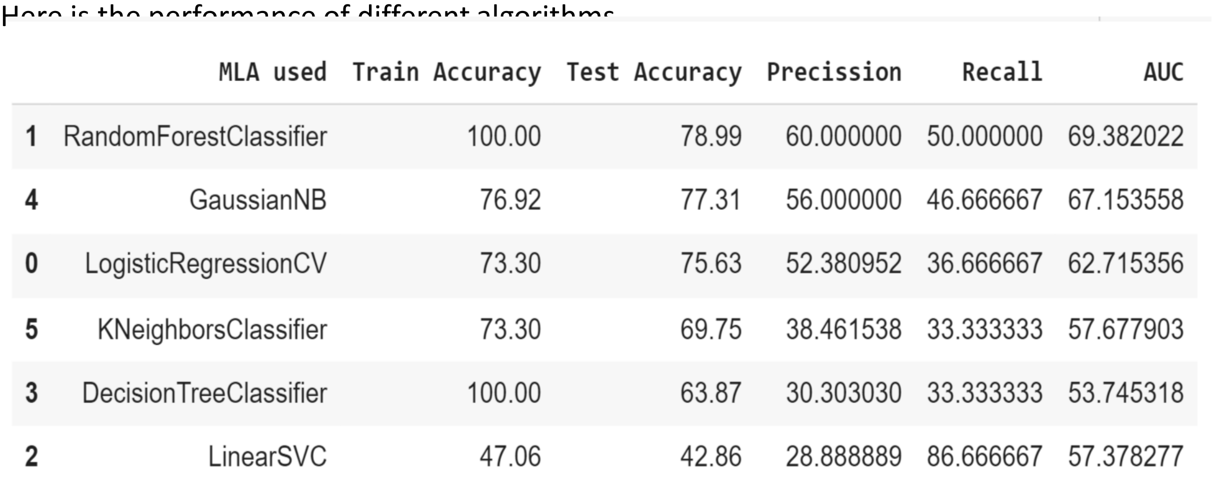
We need to build different regression algorithms and using the testing set we can determine which model to keep for making final predictions.

Here is the list of all the algorithms we've to build and evaluated:

1. DecisionTreeClassifier
2. SVM Classifier
3. Logistic regression
4. Random Forest Classifier
5. KNeighborsClassifier
6. NaiveBayes Classifier

Here different algorithms are used to train the model firstly we used the Classifiers and predicted the accuracy

Here is the performance of different algorithms



	MLA used	Train Accuracy	Test Accuracy	Precission	Recall	AUC
1	RandomForestClassifier	100.00	78.99	60.000000	50.000000	69.382022
4	GaussianNB	76.92	77.31	56.000000	46.666667	67.153558
0	LogisticRegressionCV	73.30	75.63	52.380952	36.666667	62.715356
5	KNeighborsClassifier	73.30	69.75	38.461538	33.333333	57.677903
3	DecisionTreeClassifier	100.00	63.87	30.303030	33.333333	53.745318
2	LinearSVC	47.06	42.86	28.888889	86.666667	57.378277

LogisticRegression :- Here as it is a regressor we can perform classification as the target variables are two.I have got more test accuracy than train accuracy

RandomForestClassifier :-It gives the highest test accuracy and train accuracy.

LinearSVM :- Linearsvm shows the lowest performance that means it is not a good model now .It may be due training and testing set differences

DecisionTreeClassifier :-As it gave 100 % train accuracy it failed at the testing test .so it cannot be considered as the best fit .

Naïve Bayes Classifier :-It is a probabilistic classifier it classifies based on the bayes theorem.

KNearestNeighbour Classifier :-It doesn’t train the model before the testing it only makes prediction when the test set is provided.

5. Final Conclusion

- The ensemble model has performed well compared to that of linear, Decision Tree.
- The best performance is given by the Random Forest model.
- The top key features that effects the cirrhosis are :Age, Prothrombin, Copper
- Features like Ascites and Cholestrol doesn’t show more effect on the disease.
- We can conclude from above that Random Forest Classifier is giving better results compared to that of normal Methods.