LiviD (Hepatic Disease Detection and consequent classification.)

Group – 16

- **Participants list**

1. Atul Choudhari
2. Rishik Sharma
3. Umeshkrishna U.

- **Problem statement**

The project aims at classifying those that have contracted liver disease and those that are healthy, based on patient records collected from north east of Andhra Pradesh, India.

- **Overall summary of the solution**

1) Data collection: - data set downloaded from UCI web site - http://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)
2) Data cleaning: - Missing data was first randomly added then treated using appropriate logic and duplicate values were removed.
3) Feature Selection and Modification: - Identified important features. Converted Categorical data, transformed data into appropriate data so that it can be used in classification.
4) Built Different Machine Learning models like
   a. Logistic regression.
   b. Support vector classifier and SVM.
   c. Tree-Based Methods and Ensemble techniques: - Decision tree regression, Bagging, Boosting and Random forest.
   d. Non-parametric methods – KNN.
5) Optimized Hyperparameter of above models using randomized search and grid search wherever necessary.
6) For each model calculated Classification metrics like
   a. Accuracy
   b. Precision
   c. Recall
   d. ROC AUC
7) Finally, compared different methods to select the best model

- **Detailed description and analysis**

1. **Descriptive Analysis**
   I. **Data – 10 Features, 583 Rows**

Age- (age)
The patients age is noted as the enzyme levels change as there is an increase in age.

Gender- (gender)
Gender is considered an important factor as the hormonal and physiological differences in the male and female body produce slightly different responses and it is seen that males have a higher incidence of liver disorders.

Total Bilirubin- (tot bilirubin)
Bilirubin is a bile pigment, and is the excretory end product of heme-degradation. The normal concentration of serum bilirubin is in the range of 0.2-1.0 mg/dl.

Direct Bilirubin- (direct bilirubin)
Direct bilirubin is used as an indicator of obstructive jaundice and the direct bilirubin (unconjugated) serum concentration ranges from 0.2-0.6 mg/dl.

Total Proteins- (tot proteins)
This test checks for the levels of protein, specifically albumin and globulin in the blood. If total protein is abnormal, further testing must be performed to identify which specific protein is abnormally low or high so that a specific diagnosis can be made.

Albumin- (albumin)
Albumin is solely synthesized by the liver. It has a half-life of about 20-25 days; therefore, it is a good marker to assess chronic (and not acute) liver damage. Low serum albumin is commonly observed in patients with severe liver damage. A normal range of albumin is 39 - 51 g/L of blood.

Ag Ratio- (ag_ratio)
This is a blood test to measure the levels of protein in your body. Your liver makes most of the proteins that are found in your blood. This test provides information about the amount of albumin you have compared with globulin. This comparison is called the A/G ratio. Certain diseases tend to lower your level of albumin and raise your level of one or more types of globulins A normal range for albumin is 39 - 51 g/L of blood total globulins is 23 to 35 g/L.

SGPT - (sgpt)
Serum glutamate pyruvate transaminase (SGPT) also called alanine transaminase (ALT), Serum ALT level is increased in liver damage and SGPT is more sensitive and reliable for the assessment of liver functioning.
Normal serum level of SGPT is 5-40 lU/l .

SGOT- (sgot)
Serum glutamate oxaloacetate transaminase (SGOT); also known as aspartate transaminase (AST), this enzyme is not as reliable to asses liver functionality as its rise is not specific to liver dysfunction.
Normal range for serum concentration for SGOT is 5-45 lU/l).

Alkaline Phosphate- (alkphos)
A rise in serum ALP (normal 3-13 KA units/dl), usually associated with elevated serum bilirubin is an indicator of biliary obstruction (obstructive/post-hepatic jaundice). ALP is also elevated in cirrhosis of liver and hepatic tumours.
Liver is not the sole source of alkaline phosphatase. Therefore, its measurement has to be carefully viewed (along with others) before arriving at any conclusion.

Target- (target)
The target is a binary class of patients suffering from liver disease denoted as “2” and patients who donot suffer from liver disease denoted as “1”.

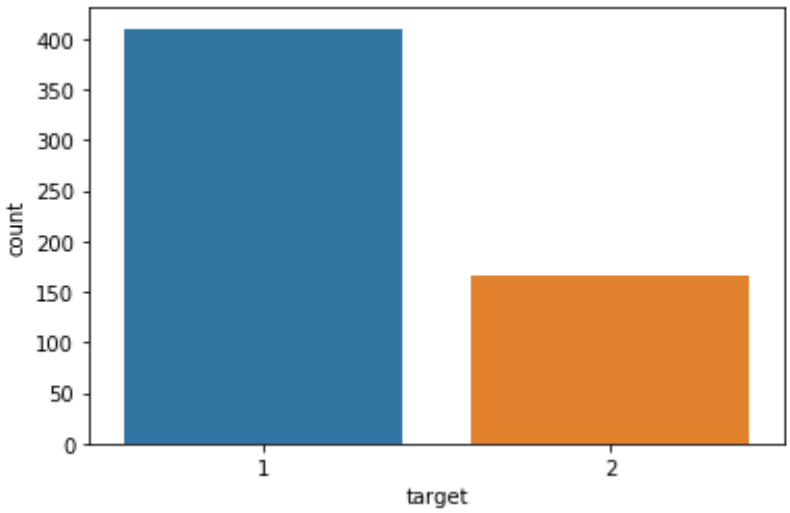
II. Descriptive Statics –

	age	gender	tot_bilirubin	direct_bilirubin	tot_proteins	albumin	ag_ratio	sgpt	sgot	alkphos	target
count	576.000000	576	576.000000	576.000000	576.000000	576.000000	576.000000	576.000000	576.000000	576.00000	576.000000
unique	NaN	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Male	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	450	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	45.071429	NaN	3.303819	1.488542	291.407986	80.815972	110.072917	6.485764	3.141840	0.94599	1.288194
std	15.386858	NaN	6.239048	2.820744	244.174425	183.575998	290.569040	1.088168	0.796363	0.31813	0.453316
min	4.000000	NaN	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000	0.900000	0.30000	1.000000
25%	34.000000	NaN	0.800000	0.200000	176.000000	23.000000	25.000000	5.800000	2.600000	0.70000	1.000000
50%	45.071429	NaN	1.000000	0.300000	208.000000	35.000000	41.000000	6.600000	3.100000	0.93000	1.000000
75%	55.000000	NaN	2.600000	1.300000	298.000000	60.250000	87.000000	7.200000	3.800000	1.10000	2.000000
max	90.000000	NaN	75.000000	19.700000	2110.000000	2000.000000	4929.000000	9.600000	5.500000	2.80000	2.000000

III. EDA

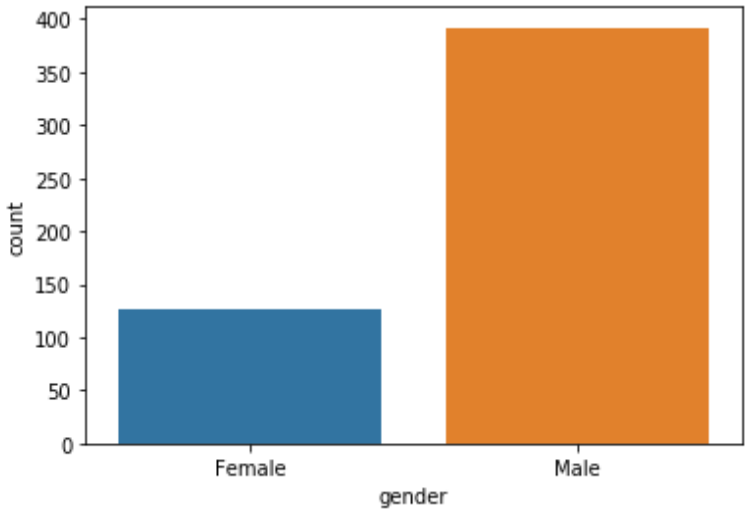
A. Univariate analysis

a. Target variable



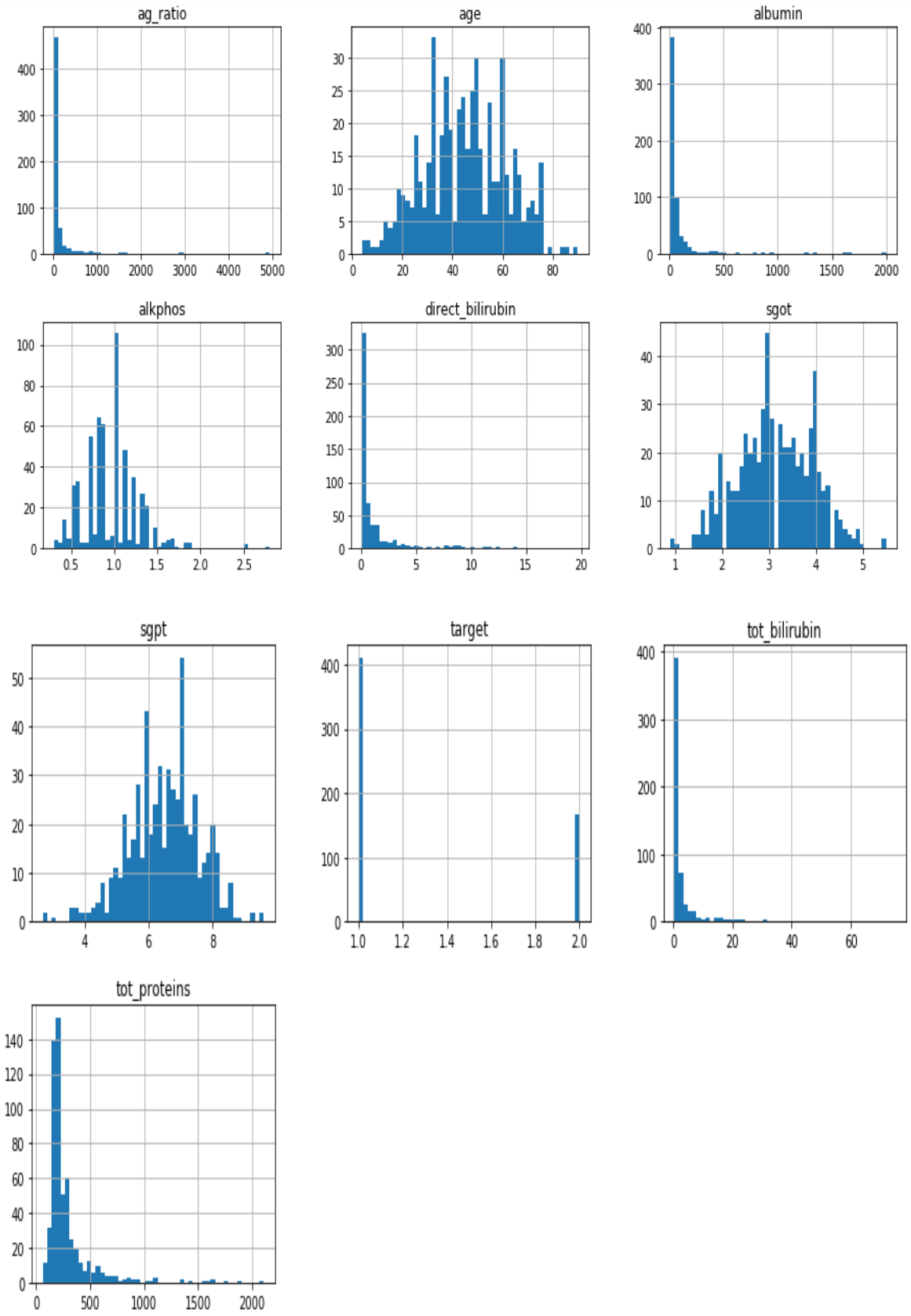
The target variable shows class imbalance with higher cases of positive variable.

b. Gender Ratio



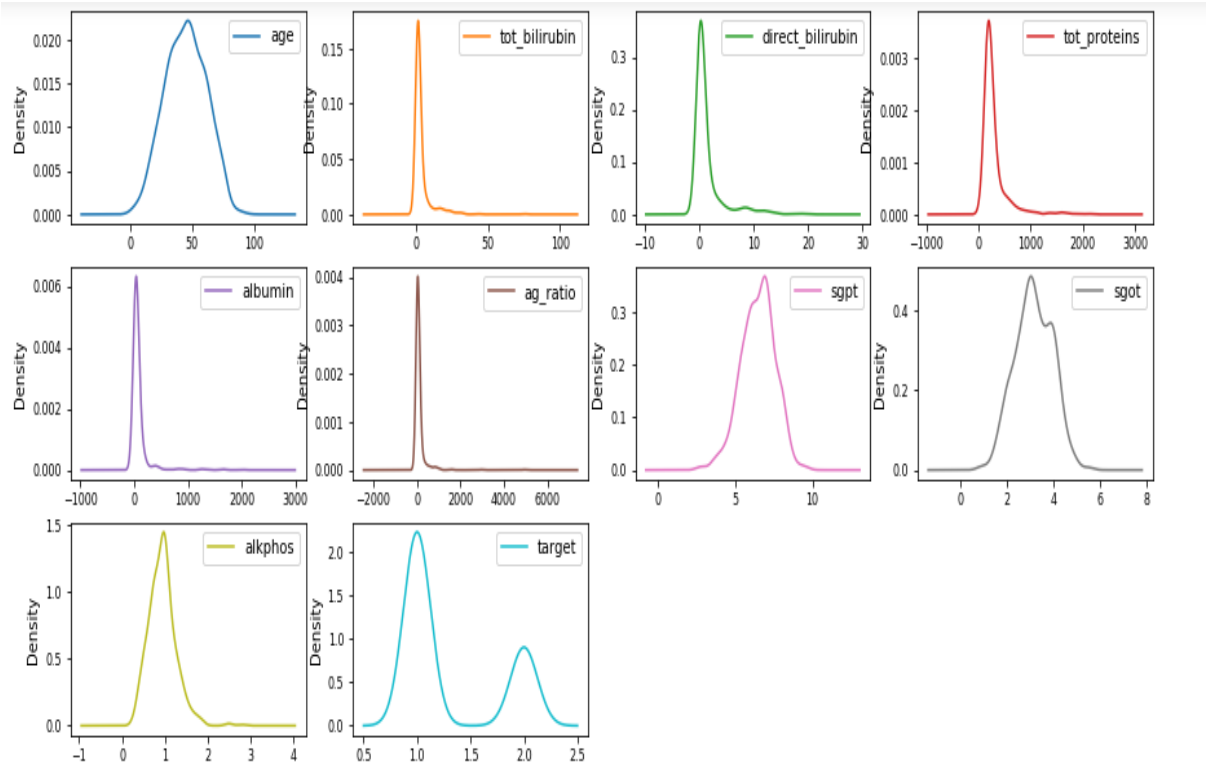
We can see a higher prevalence of males in our sample population.

c. Individual histograms



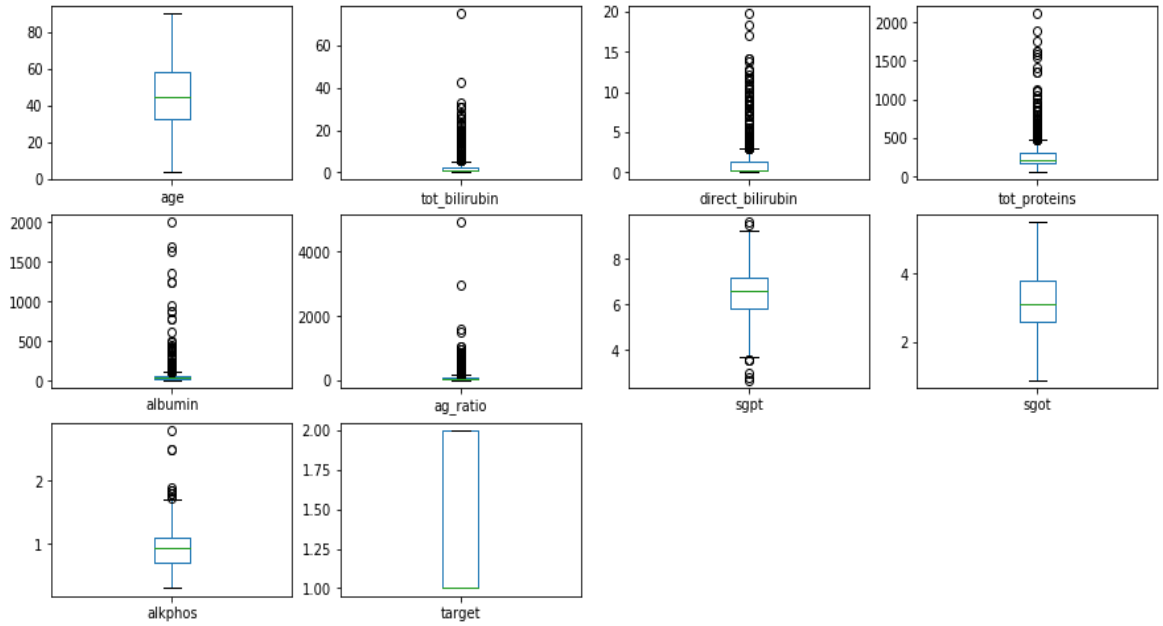
We can see clearly that except for age, alkphos, sgot and sgpt all features are positively skewed which indicates need of a non-linear model and scaling.

d. **Distribution plot**



Here we can see that for most of the features are positively skewed and age, sgpt, sgot are symmetric. And most of the values are highly concentrated in specific ranges.

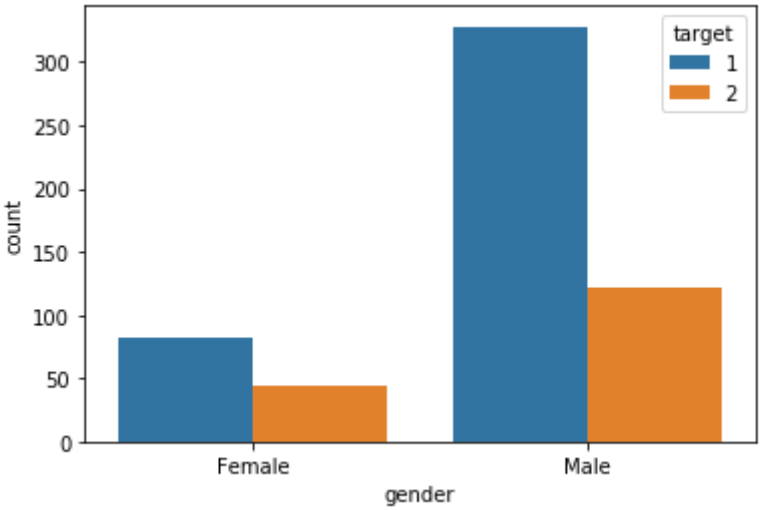
e. Boxplot summary



We can see a high presence of outliers in tot_bilirubin, direct_bilirubin, tot_protiens, albumin, ag_ratio and alkphos. This inturn indicates presence of abnormal serum concentration for these measures.

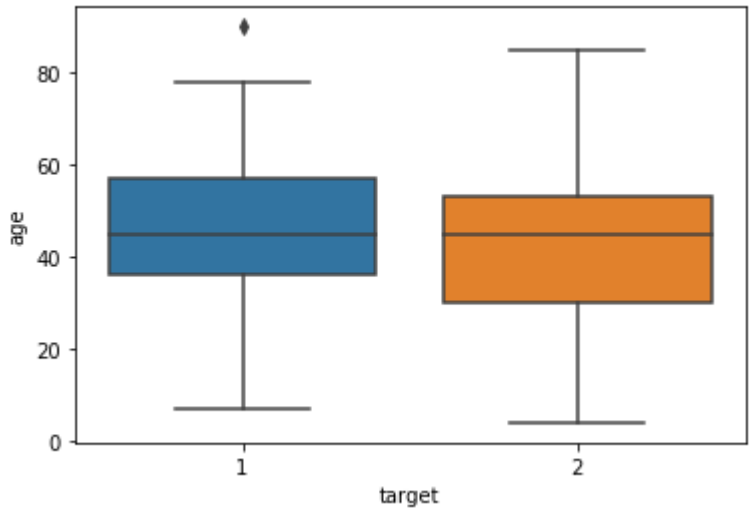
B. Bivariate analysis

a. Gender vs Target count



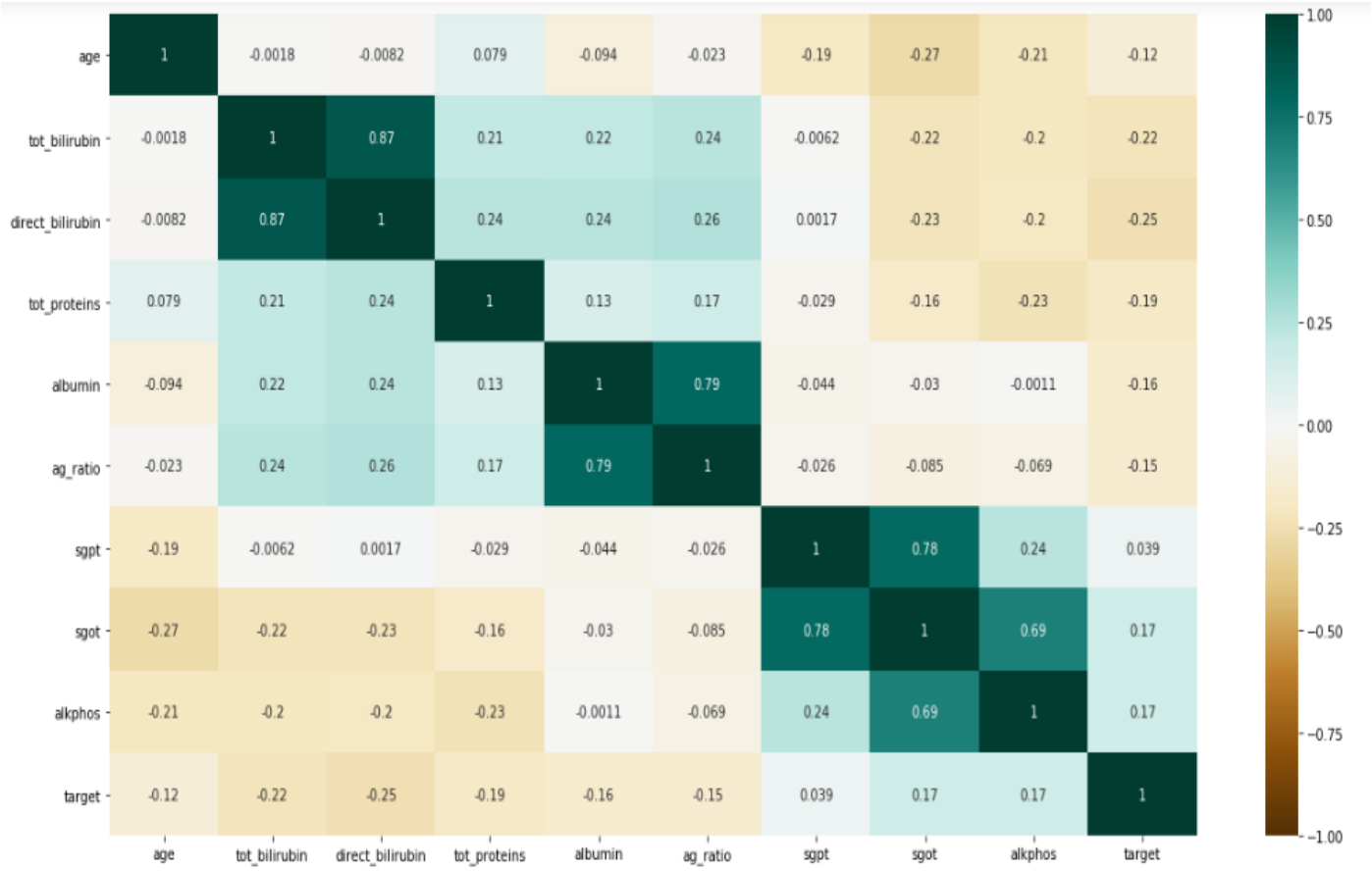
We can see that the proportion of males suffering from liver disorders is higher.

b. Age vs Target Boxplot



We can see that the disorders are normally distributed when we consider age.

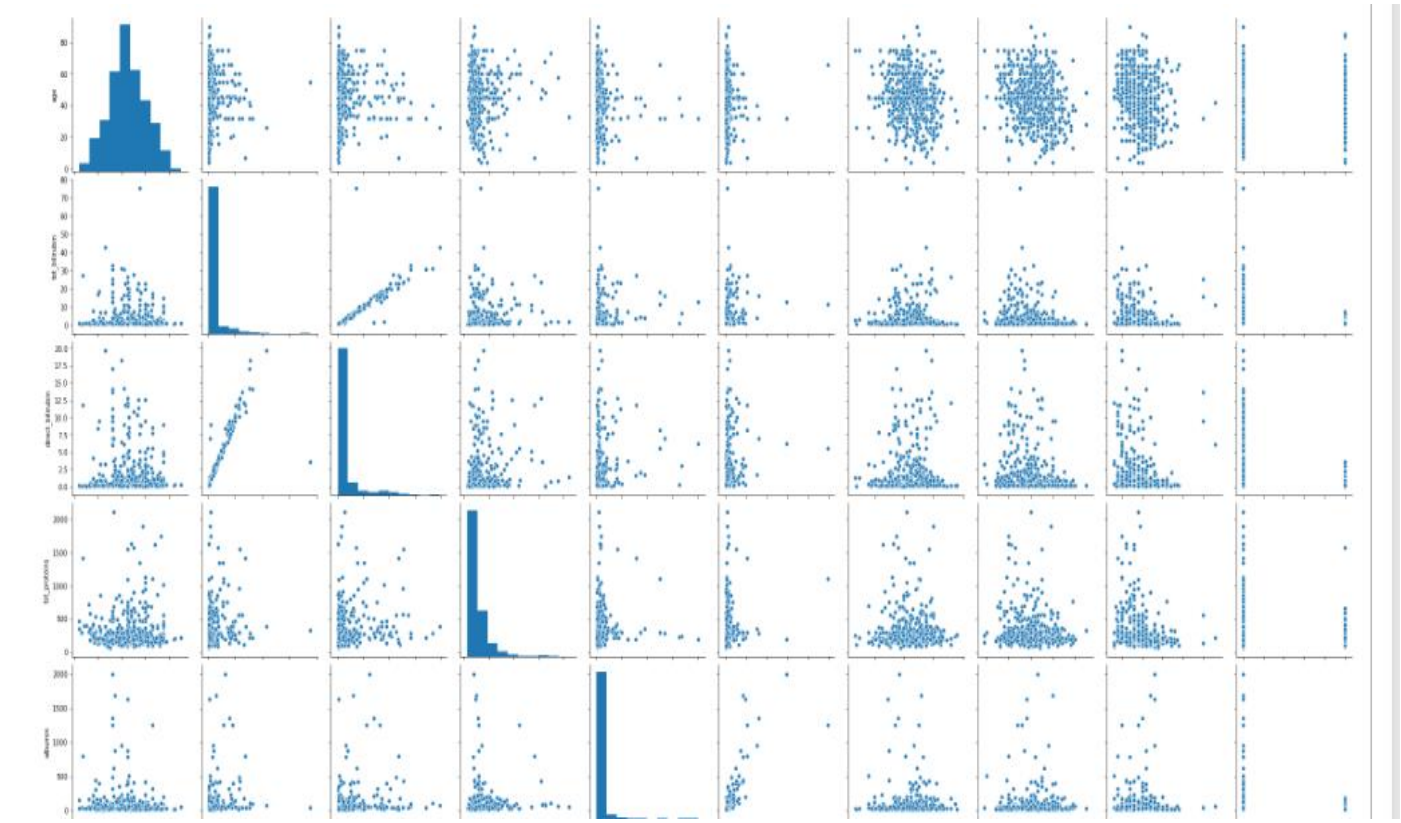
c. Correlation plot and matrix

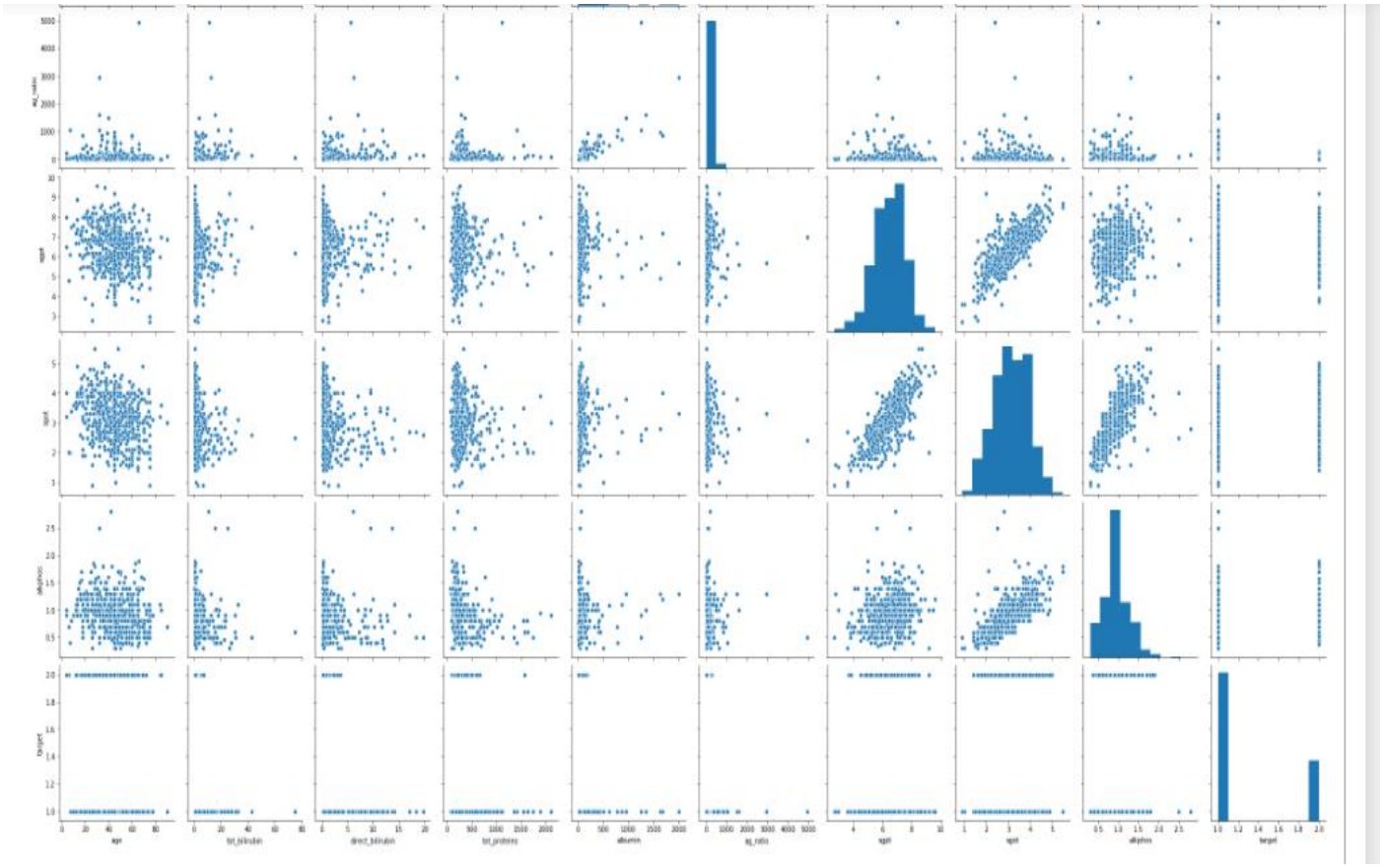


	age	tot_bilirubin	direct_bilirubin	tot_proteins	albumin	ag_ratio	sgpt	sgot	alkphos	target
age	1.000000	-0.001726	-0.007981	0.076870	-0.084316	-0.022514	-0.179613	-0.255211	-0.203586	-0.118077
tot_bilirubin	-0.001726	1.000000	0.874282	0.206796	0.215284	0.238473	-0.006224	-0.219871	-0.203868	-0.223173
direct_bilirubin	-0.007981	0.874282	1.000000	0.235228	0.235364	0.258401	0.001658	-0.225994	-0.197195	-0.249166
tot_proteins	0.076870	0.206796	0.235228	1.000000	0.125037	0.166730	-0.028898	-0.164907	-0.232734	-0.185711
albumin	-0.084316	0.215284	0.235364	0.125037	1.000000	0.791890	-0.043536	-0.029823	-0.000984	-0.163582
ag_ratio	-0.022514	0.238473	0.258401	0.166730	0.791890	1.000000	-0.025857	-0.085135	-0.069063	-0.152750
sgpt	-0.179613	-0.006224	0.001658	-0.028898	-0.043536	-0.025857	1.000000	0.784703	0.234622	0.038652
sgot	-0.255211	-0.219871	-0.225994	-0.164907	-0.029823	-0.085135	0.784703	1.000000	0.684901	0.165502
alkphos	-0.203586	-0.203868	-0.197195	-0.232734	-0.000984	-0.069063	0.234622	0.684901	1.000000	0.165283
target	-0.118077	-0.223173	-0.249166	-0.185711	-0.163582	-0.152750	0.038652	0.165502	0.165283	1.000000

We can see high positive correlation between some of the features which indicates multi collinearity.

d. Pairplot





As seen in the correlation plot we can see the same features exhibiting linearity through this pair plot.

IV. Data Preparation

- Duplicate values were removed from the data

```
df_duplicate = df[df.duplicated(keep = False)] # keep = False gives you all rows with duplicate entries
df = df[~df.duplicated(subset = None, keep = 'first')]
df.shape
```

```
(576, 11)
```

- The missing data was imputed using mean and mode as logic the EDA helped in taking these decisions.

```
df["age"].fillna(np.nanmean(df["age"]),inplace=True)
```

```
df["gender"].fillna((df["gender"].mode()[0]),inplace=True)
```

```
df["alkphos"].fillna(np.nanmedian(df["alkphos"]),inplace=True)
```

```
df.isnull().sum()
```

```
age                0
gender             0
tot_bilirubin      0
direct_bilirubin   0
tot_proteins       0
albumin            0
ag_ratio           0
sgpt               0
sgot               0
alkphos            0
target             0
dtype: int64
```

- The categorical feature gender was encoded using pandas get dummies function.
- The features were then scaled using a standard scaler.
- Then the data was separated into features set ‘X’ and output set ‘Y’ and split into training and testing sets of parts 7:3.

2. Analysis of Various Classification techniques

1) Logistic Regression

a. Confusion Matrix

	Actual Values		
Predicted Values		Positive (1)	Negative(0)
	Positive (1)	69 (TP)	55 (FP)
	Negative (0)	22 (FN)	100 (TN)

```
[[ 69  55]
 [ 22 100]]
```

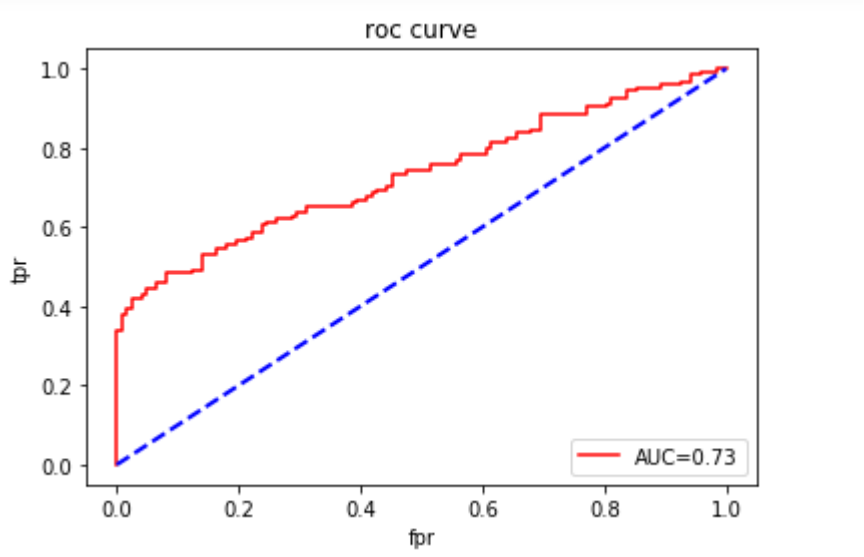
We can see that the true positives are low but true negatives are high but there is still substantial misclassification.

b. Training vs Test Metrics

	Accuracy	Precision	Recall	ROC-AUC score
Train	0.75	0.92	0.54	-
Test	0.69	0.76	0.56	0.73

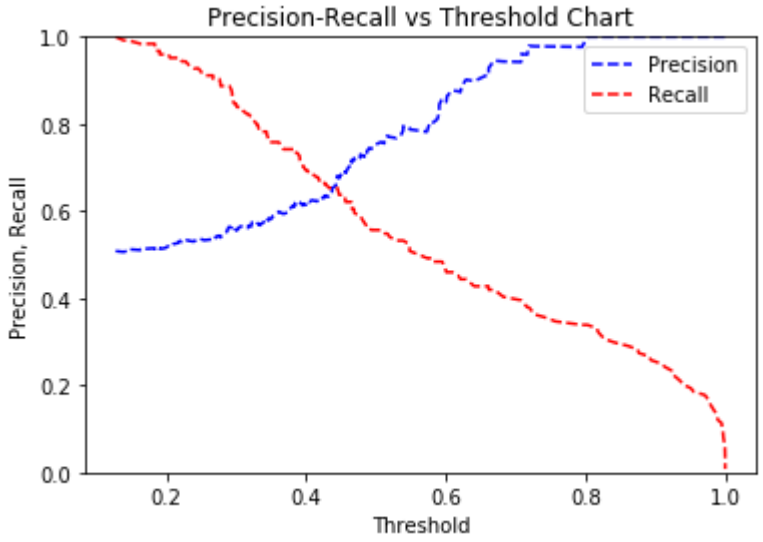
We see high precision but low accuracy and recall.

c. Roc-Auc Curve



The ROC AUC curve gives idea that the model is not up to the mark. The roc score is 0.73.

d. Precision vs Recall Curve



This chart gives idea about how changing the threshold value results in changing of values for precision and recall.

2) Support vector Classifier

Confusion Matrix

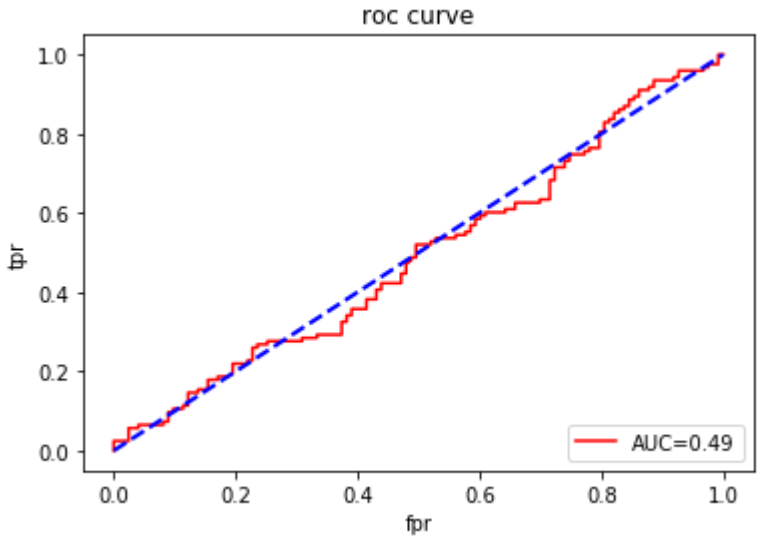
	Actual Values		
Predicted Values		Positive (1)	Negative(0)
	Positive (1)	74 (TP)	49 (FP)
	Negative (0)	10 (FN)	113 (TN)

SVC performs slightly better than the logistic model.

Training vs Test Metrics

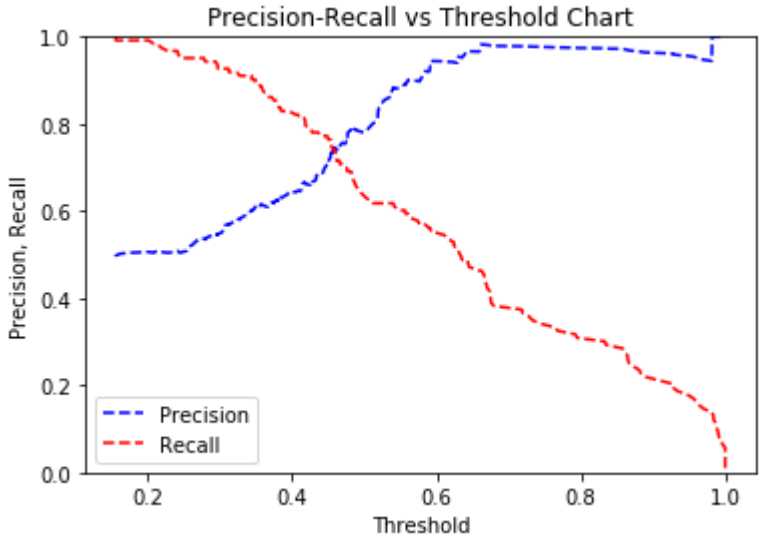
	Accuracy	**Precision**	**Recall**	**ROC-AUC score**
Train	0.73	0.84	0.56	
Test	0.76	0.88	0.60	0.49

a. Roc-Auc Curve



The bad AUC show that the SVC does not perform properly even though it gives slightly higer accuracy and precision.

b. Precision vs Recall Curve



We can see that at around 0.45 precision and recall become equal.

3) Support Vector Machines

a. Confusion Matrix

	Actual Values		
Predicted Values		Positive (1)	Negative(0)
	Positive (1)	81 (TP)	42 (FP)
	Negative (0)	34 (FN)	89 (TN)

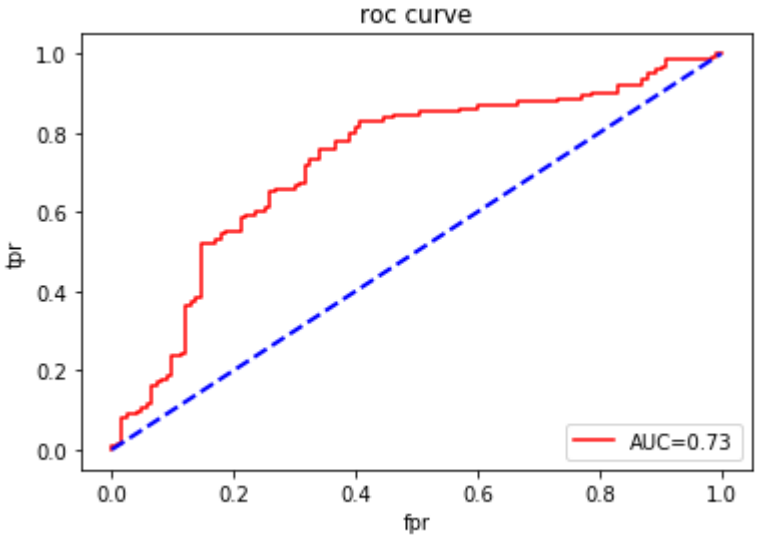
The SVM radial basis function does not work very well there are still many falsely classified patients.

b. Training vs Test Metrics

	Accuracy	Precision	Recall	ROC-AUC score
Train	**0.95**	**0.97**	**0.93**	**-**
Test	**0.69**	**0.70**	**0.66**	**0.73**

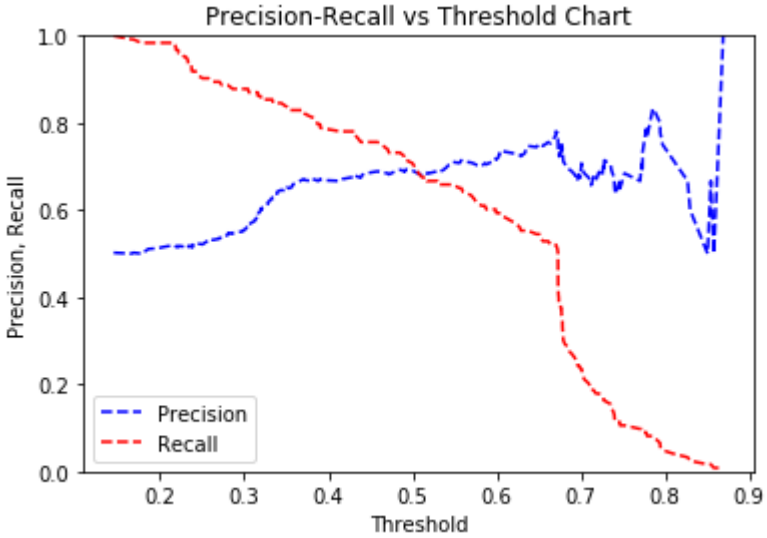
We can clearly see that SVM overfits the given training data.

c. Roc-Auc Curve



The AUC is around 0.73 which does not indicate a very good model.

d. Precision vs Recall Curve



The precision and recall converge around 0.52.

4) Decision Tree Classifier

a. Confusion Matrix

	Actual Values		
Predicted Values		Positive (1)	Negative(0)
	Positive (1)	74 (TP)	49 (FP)
	Negative (0)	29 (FN)	94 (TN)

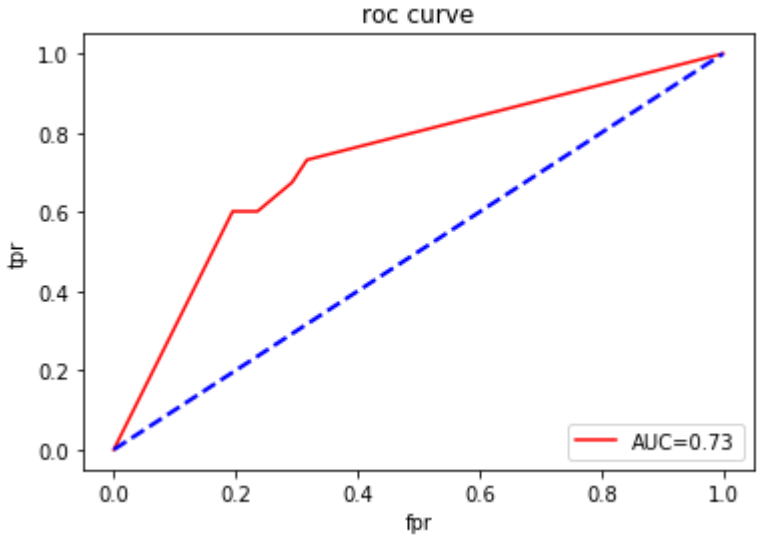
We can see that there are many miss classified data points.

b. Training vs Test Metrics

	Accuracy	**Precision**	**Recall**	**ROC-AUC score**
Train	**0.95**	**0.97**	**0.93**	
Test	**0.68**	**0.72**	**0.60**	**0.73**

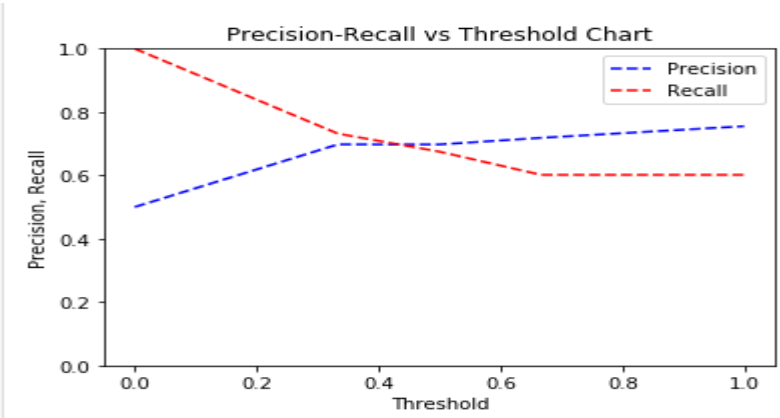
The model over fits the training data set.

c. Roc-Auc Curve



The AUC score is 0.73 similar to previous models.

d. Precision vs Recall Curve



The precision and recall converge around 0.45.

5) Random Forest

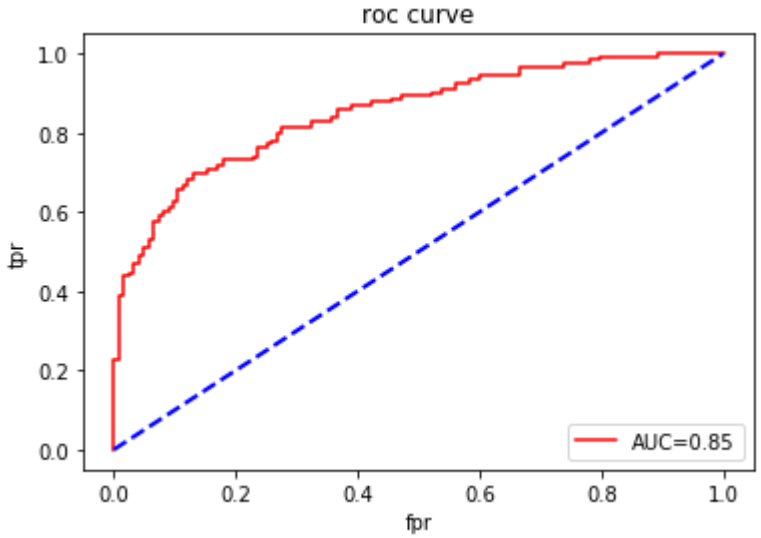
a. Confusion Matrix

	Actual Values		
Predicted Values		Positive (1)	Negative(0)
	Positive (1)	89 (TP)	34 (FP)
	Negative (0)	22 (FN)	101 (TN)

b. Training vs Test Metrics

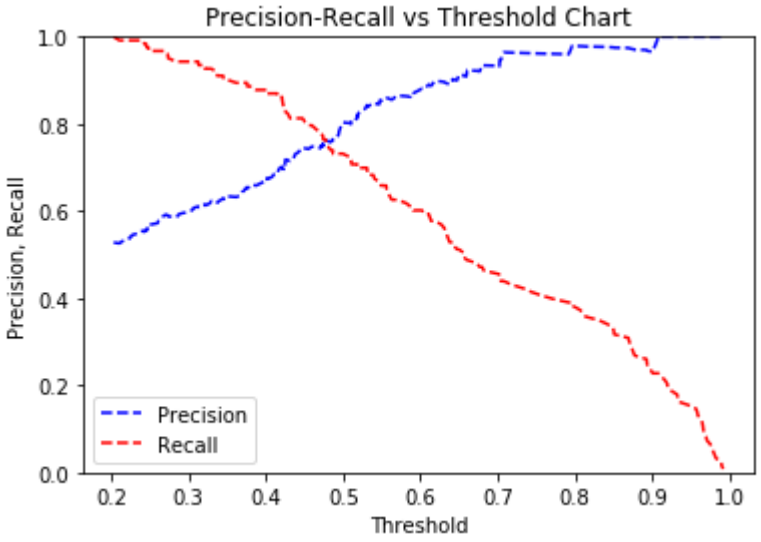
	Accuracy	Precision	Recall	ROC-AUC score
Train	0.93	0.96	0.89	-
Test	0.77	0.80	0.72	0.85

c. Roc-Auc Curve



The random forest model gives an AUC of 0.85 which is a very good score.

d. Precision vs Recall Curve



The precision and recall converge somewhere near 0.48.

6) Adaboost

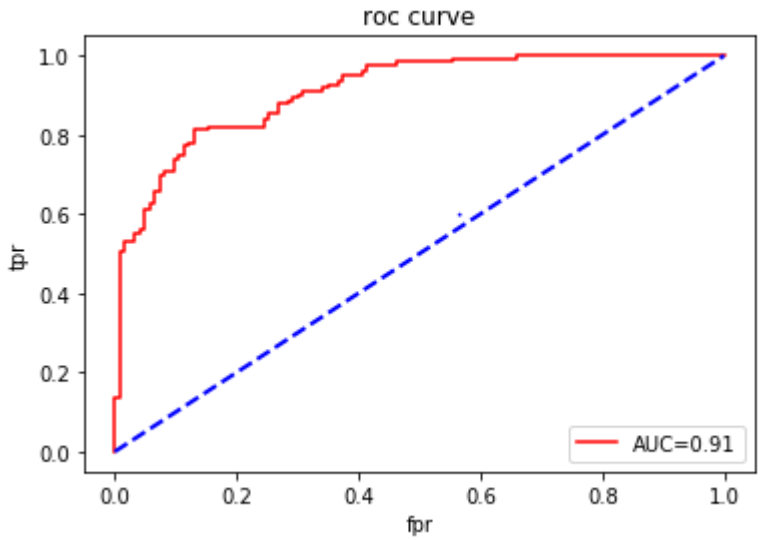
a. Confusion Matrix

	Actual Values		
Predicted Values		Positive (1)	Negative(0)
	Positive (1)	97 (TP)	26 (FP)
	Negative (0)	16 (FN)	107 (TN)

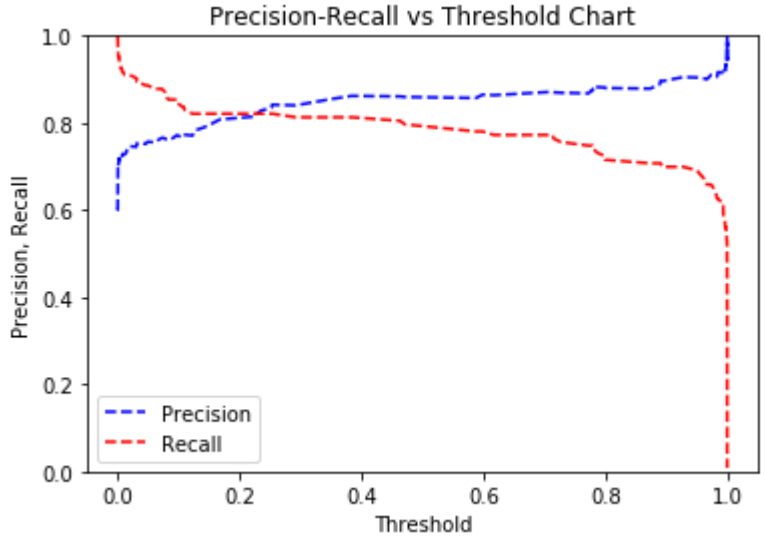
b. Training vs Test Metrics

	Accuracy	Precision	Recall	ROC-AUC score
Train	1.0	1.0	1.0	-
Test	0.83	0.86	0.79	0.91

c. Roc-Auc Curve



d. Precision vs Recall Curve



.

7) KNN

a. Confusion Matrix

	Actual Values		
Predicted Values		Positive (1)	Negative(0)
	Positive (1)	70 (TP)	53 (FP)
	Negative (0)	18 (FN)	105 (TN)

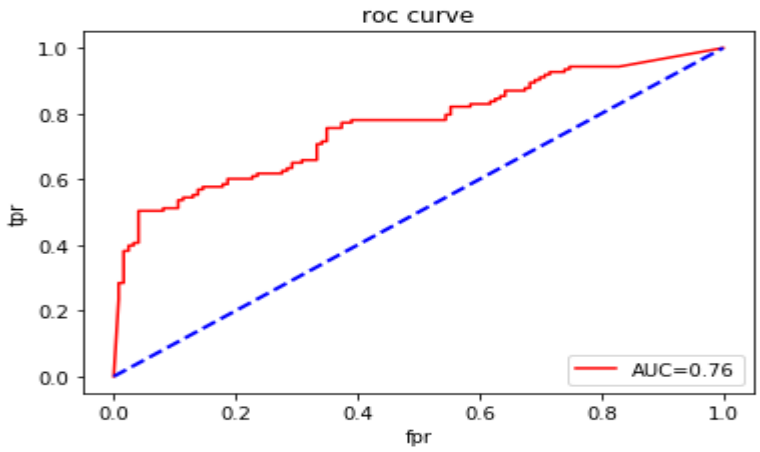
There are many falsely classified datapoints which indicates the model does not work well.

b. Training vs Test Metrics

	Accuracy	**Precision**	**Recall**	**ROC-AUC score**
Train	**1.0**	**1.0**	**1.0**	**-**
Test	**0.71**	**0.80**	**0.57**	**0.76**

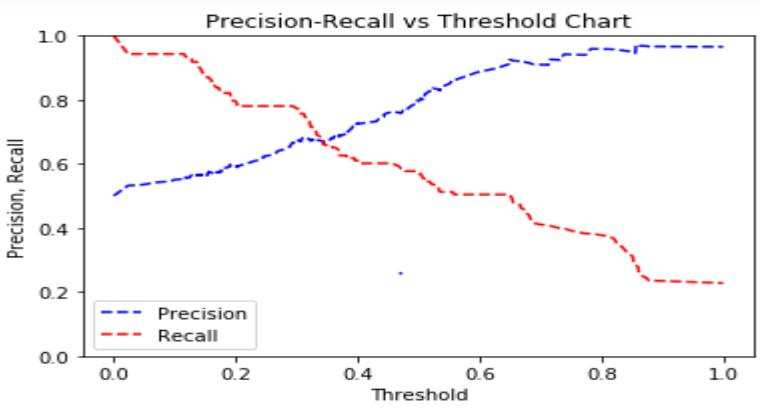
We can cleary see that the model overfits the given training set.

c. Roc-Auc Curve



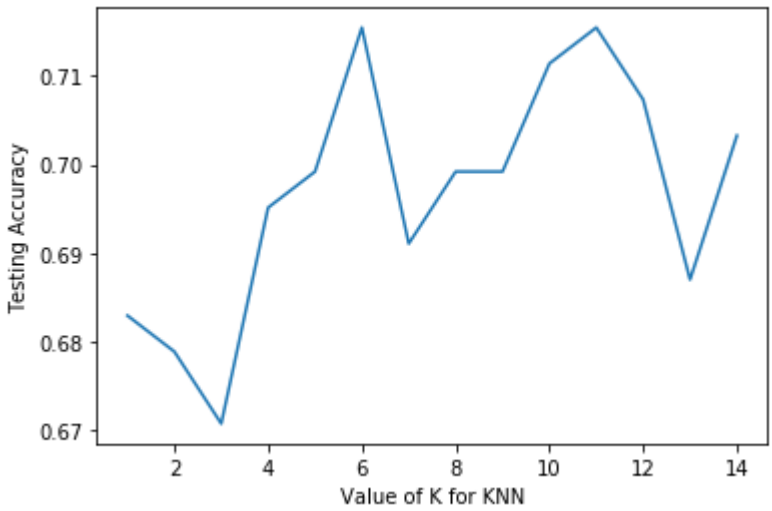
The roc auc score is observed to be 0.76 which indicated its bot the best model.

d. Precision vs Recall Curve



We can see convergence of precision and recall at approx. 0.35 as threshold value.

e. **Best value for K neighbours**



The best values for K is observed to be 6 and 11.

3. Conclusion

1) Comparison Table

Sr.no.	Technique name / metrics	Accuracy	Precision	Recall	ROC AUC Score
1.	Logistic Regression	0.69	0.76	0.56	0.73
2.	Support vector Classifier	0.76	0.88	0.60	0.49
3.	Support Vector Machines	0.69	0.70	0.66	0.73
4.	Decision Tree Classifier	0.68	0.72	0.60	0.73
5.	Random Forest	0.77	0.80	0.72	0.85
6.	Adaptive Boosting	0.83	0.86	0.79	0.91
7.	KNN	0.71	0.80	0.57	0.76

3. Inferences

- **The best model for predictions of liver disease is Adaptive boosting with the highest ROC AUC Score of 0.91. The model does over fit the given data but performs best on testing samples.**
- **The second-best model is observed to be the random forest model which too does have the problem of overfitting but compared to adaboost its slightly less but in-turn we have to compromise on all three metrics with approximately 5-6 % loss in each metric.**

4. References

1) ‘Biochemistry’ Dr U. K. Satyanarayan published by Books and Allied (P) Ltd. Edition – 2007 page no. 453-459
2) ‘Investigating ILPD for most significant features’ Jothi Lakshmi U, K.Jayanthi and M.Sathya International Journal of Mechanical Engineering and Technology (IJMET) Volume 8, Issue 10, October 2017, pp. 741–749, Article ID: IJMET_08_10_080
3) https://pandas.pydata.org/docs/
4) https://scikit-learn.org/stable/
5) https://imbalanced-learn.readthedocs.io/en/stable/api.html
6) https://stackoverflow.com/

==========
==========