Analysis of the dataset of Indian Liver Patients

BY

Ashwin

Prasad

Shintu Mathew

Sreelalan S

Swaroop Murali

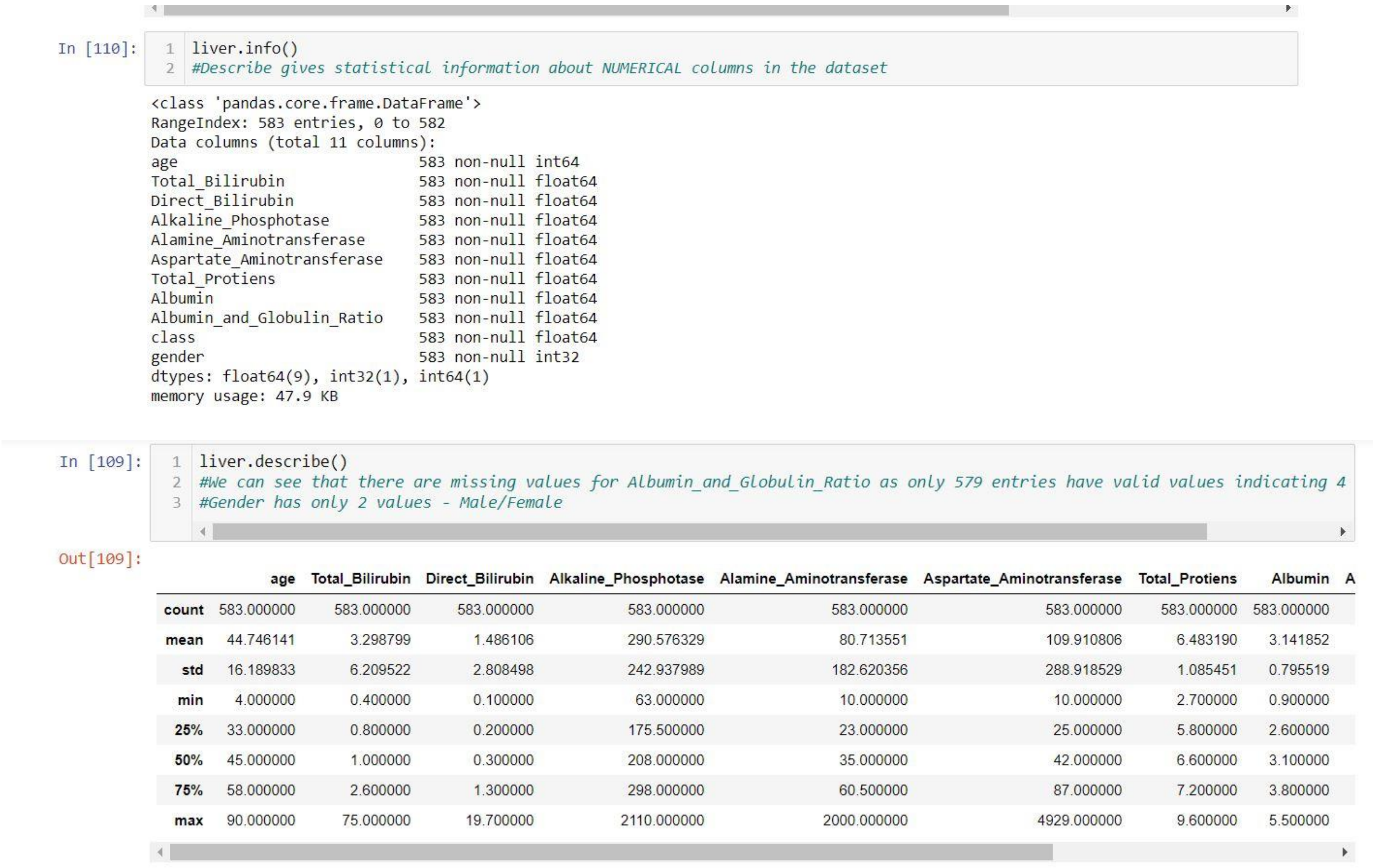
Indian Liver Patients

- Aim of this analysis is to develop a model for early detection of liver disorder from imbalance Liver Function Test
- The initial stage symptoms of the diseases are vague so the medical practitioners often fail to detect the disease. The model tries to improve the accuracy using various classification algorithms.

Sample Data

age	gender	TB	DB	alkphos	sgpt	sgot	TP	ALB	A_G	class
65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
58	Male	1	0.4	182	14	20	6.8	3.4	1	1
72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3	1
26	Female	0.9	0.2	154	16	12	7	3.5	1	1
29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1	1
17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2	2
55	Male	0.7	0.2	290	53	58	6.8	3.4	1	1
57	Male	0.6	0.1	210	51	59	5.9	2.7	0.8	1
72	Male	2.7	1.3	260	31	56	7.4	3	0.6	1

EDA – Visualization of the data



In [110]:

```
liver.info()
#Describe gives statistical information about NUMERICAL columns in the dataset
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 583 entries, 0 to 582
Data columns (total 11 columns):
age                           583 non-null int64
Total_Bilirubin               583 non-null float64
Direct_Bilirubin              583 non-null float64
Alkaline_Phosphotase          583 non-null float64
Alamine_Aminotransferase      583 non-null float64
Aspartate_Aminotransferase    583 non-null float64
Total_Protiens                583 non-null float64
Albumin                       583 non-null float64
Albumin_and_Globulin_Ratio    583 non-null float64
class                         583 non-null float64
gender                        583 non-null int32
dtypes: float64(9), int32(1), int64(1)
memory usage: 47.9 KB
```

In [109]:

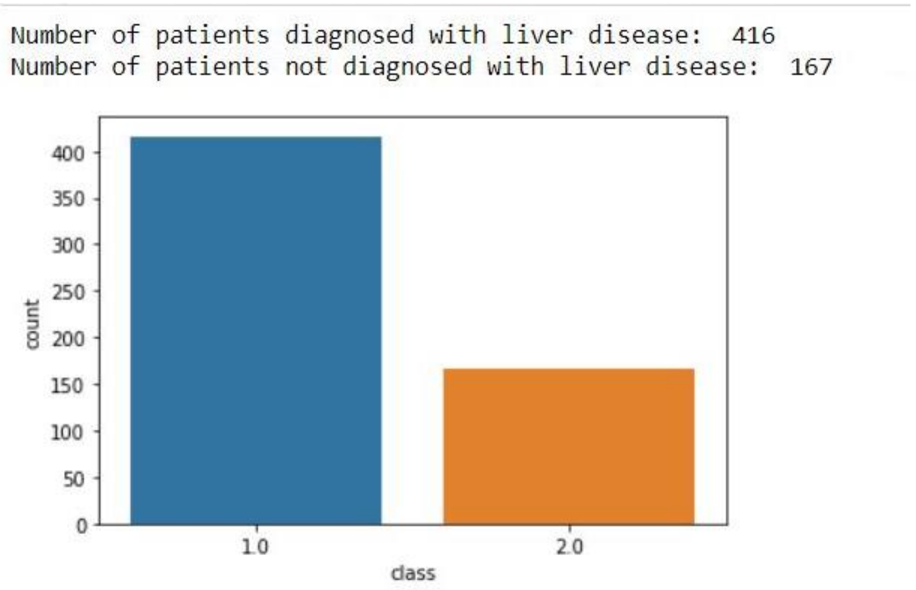
```
liver.describe()
#We can see that there are missing values for Albumin_and_Globulin_Ratio as only 579 entries have valid values indicating 4
#Gender has only 2 values - Male/Female
```

Out[109]:

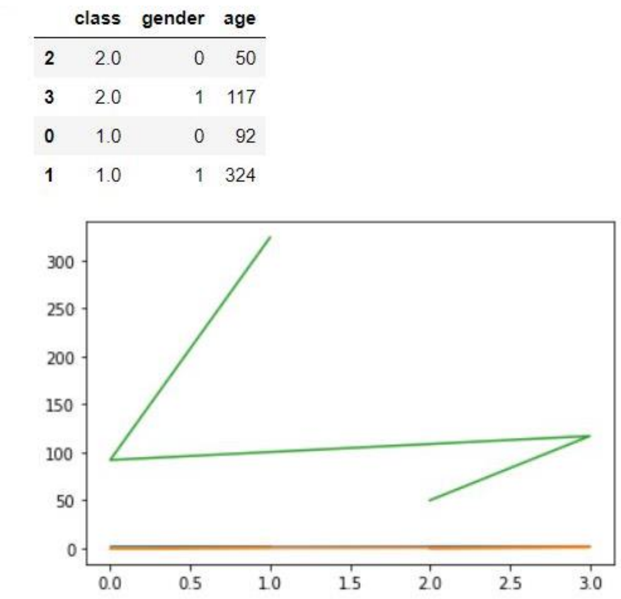
	age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	A
count	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	
mean	44.746141	3.298799	1.486106	290.576329	80.713551	109.910806	6.483190	3.141852	
std	16.189833	6.209522	2.808498	242.937989	182.620356	288.918529	1.085451	0.795519	
min	4.000000	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000	0.900000	
25%	33.000000	0.800000	0.200000	175.500000	23.000000	25.000000	5.800000	2.600000	
50%	45.000000	1.000000	0.300000	208.000000	35.000000	42.000000	6.600000	3.100000	
75%	58.000000	2.600000	1.300000	298.000000	60.500000	87.000000	7.200000	3.800000	
max	90.000000	75.000000	19.700000	2110.000000	2000.000000	4929.000000	9.600000	5.500000	

Explanation of each feature

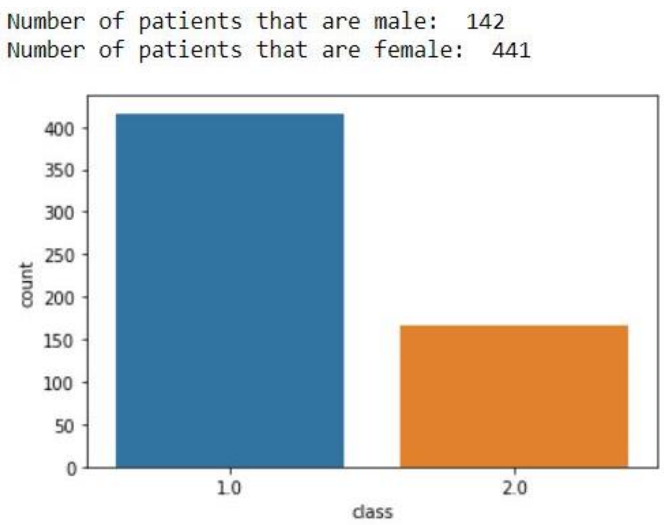
- Bilirubin is a substance made when body breaks down old red blood cells. Bilirubin is also part of bile, which liver makes to help digest the food.
- Alkaline phosphatase, or ALP, Aspartate transaminase, or AST and Alanine transaminase, or ALT are the enzymes which are found in the blood. They are present in higher quantity when the liver is damaged. They are leaked into the blood.
- Total protein, is a biochemical test for measuring the total amount of protein in the blood.
- Albumin is a protein made by the liver. Albumin helps keep fluid in the bloodstream so it doesn't leak into other tissues.
- Globulin. This is a group of proteins. Some of them are made by the liver. Others are made by immune system. They help fight infection and transport nutrients.



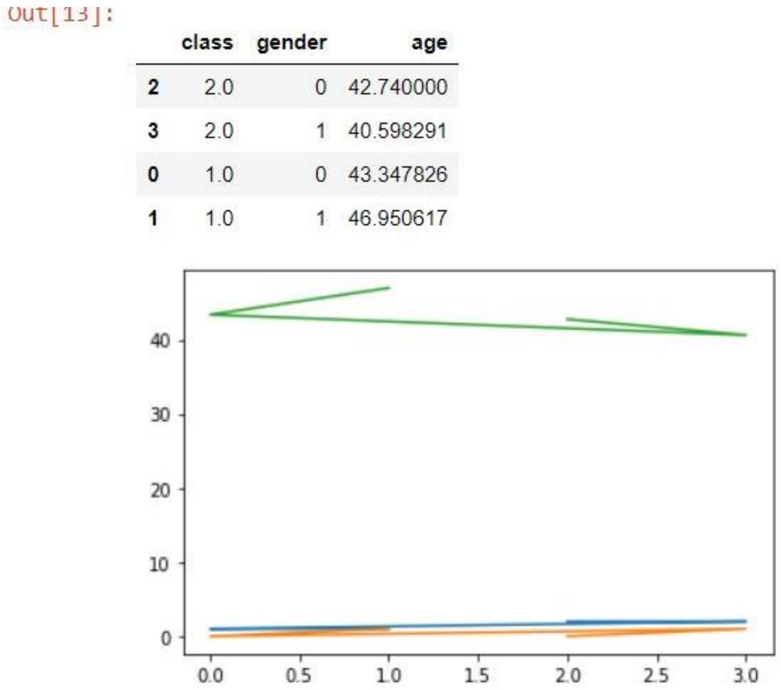
```
Number of patients diagnosed with liver disease:  416
Number of patients not diagnosed with liver disease:  167
```



	class	gender	age
2	2.0	0	50
3	2.0	1	117
0	1.0	0	92
1	1.0	1	324

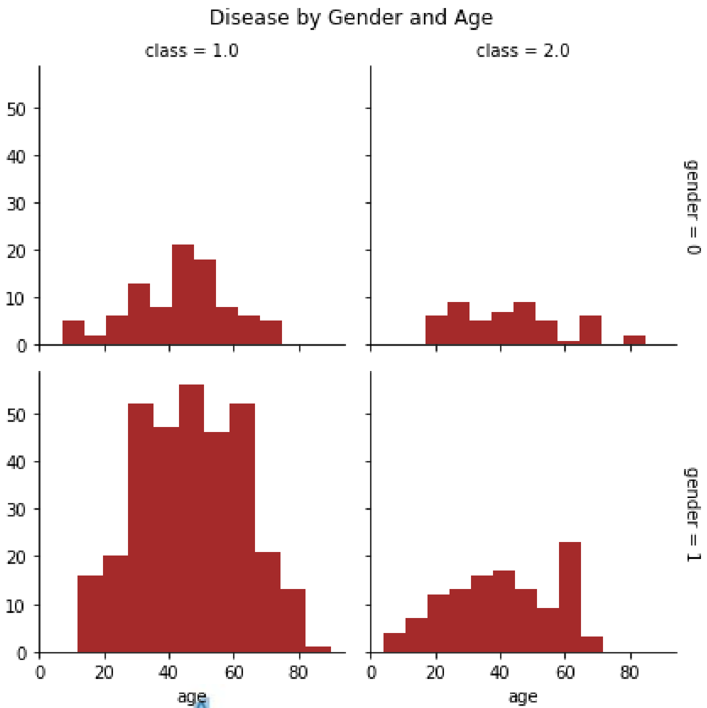


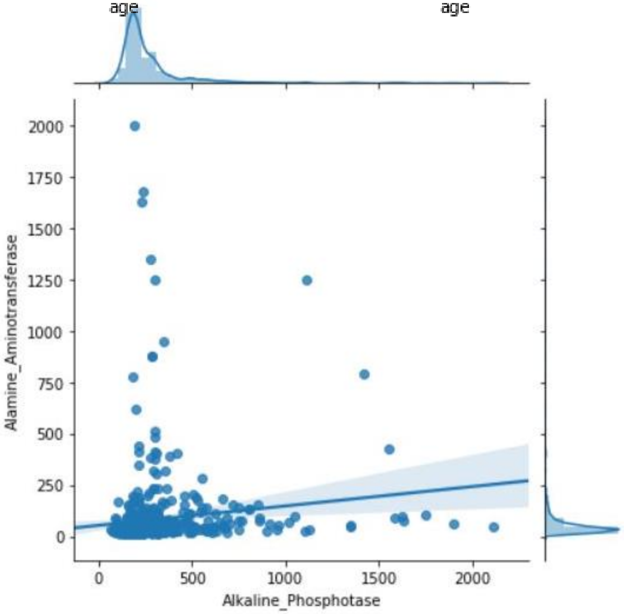
```
Number of patients that are male:  142
Number of patients that are female:  441
```

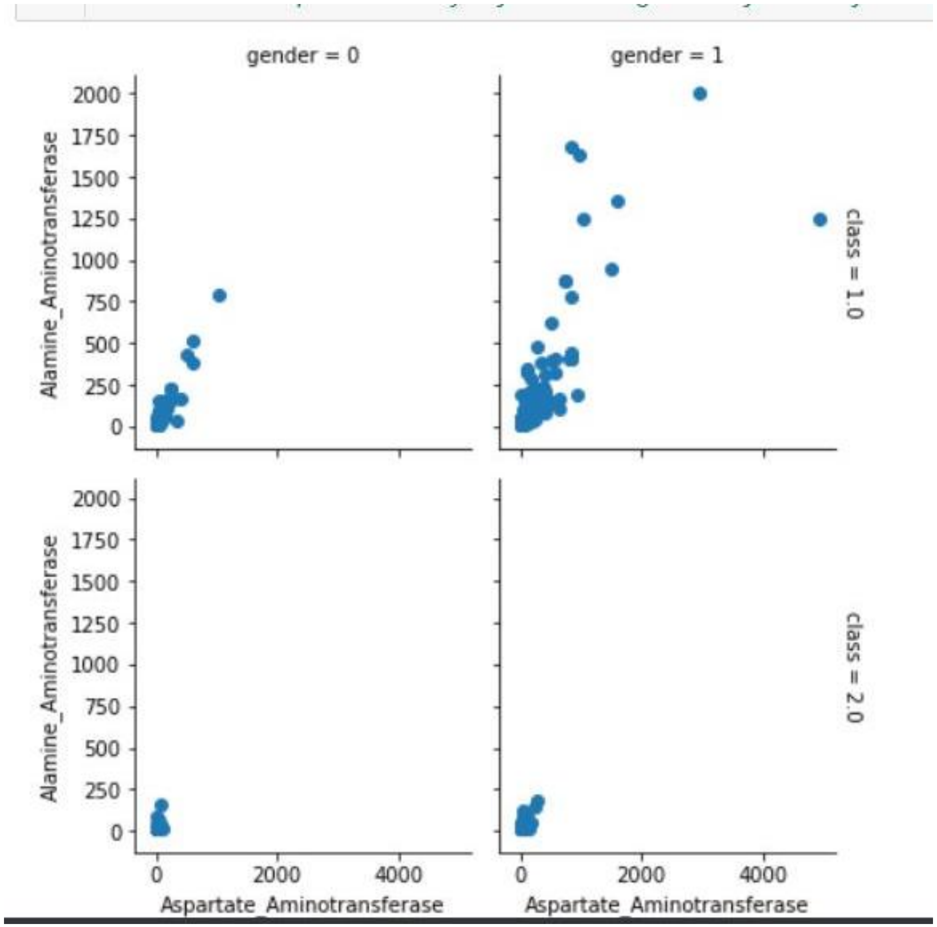


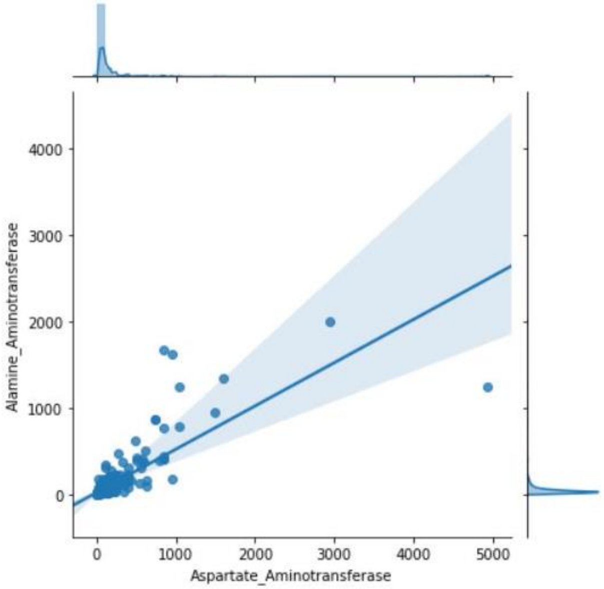
Out[13]:

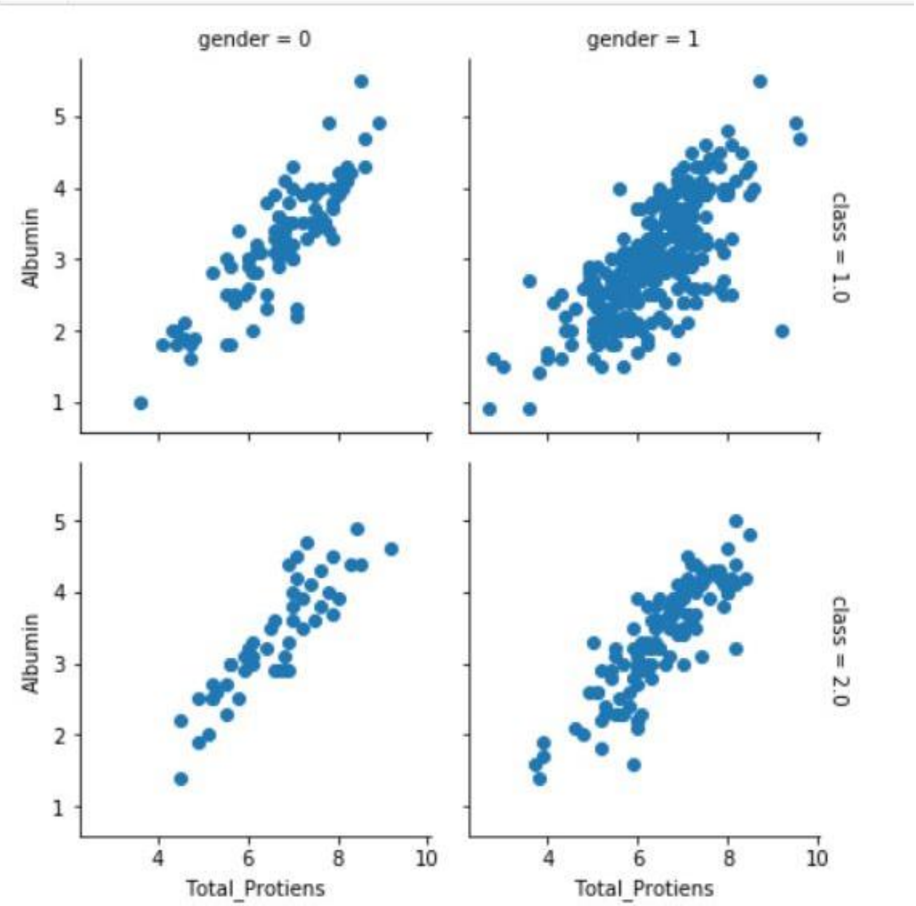
	class	gender	age
2	2.0	0	42.740000
3	2.0	1	40.598291
0	1.0	0	43.347826
1	1.0	1	46.950617

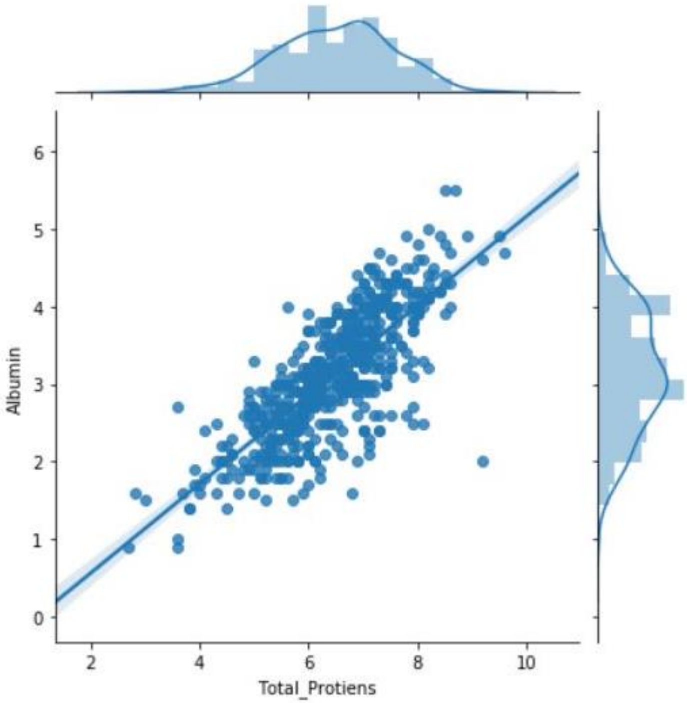

Disease by Gender and Age
class = 1.0
class = 2.0
gender = 0
gender = 1
0
10
20
30
40
50
0
20
40
60
80
age
age

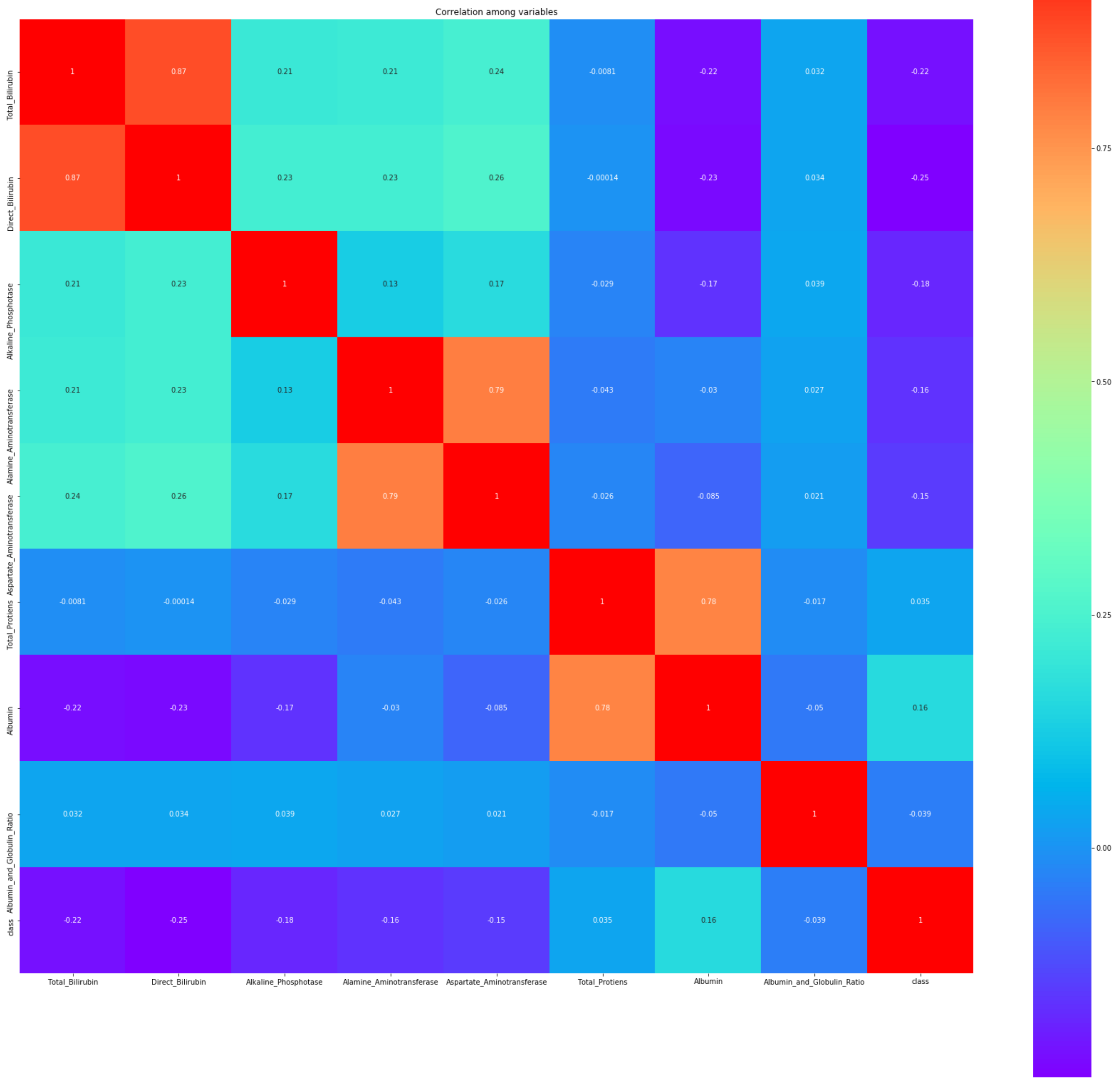

Alamine_Aminotransferase
0
250
500
750
1000
1250
1500
1750
2000
0
500
1000
1500
2000
Alkaline_Phosphotase


gender = 0
gender = 1
class = 1.0
class = 2.0
Alamine_Aminotransferase
0
250
500
750
1000
1250
1500
1750
2000
0
2000
4000
Aspartate_Aminotransferase

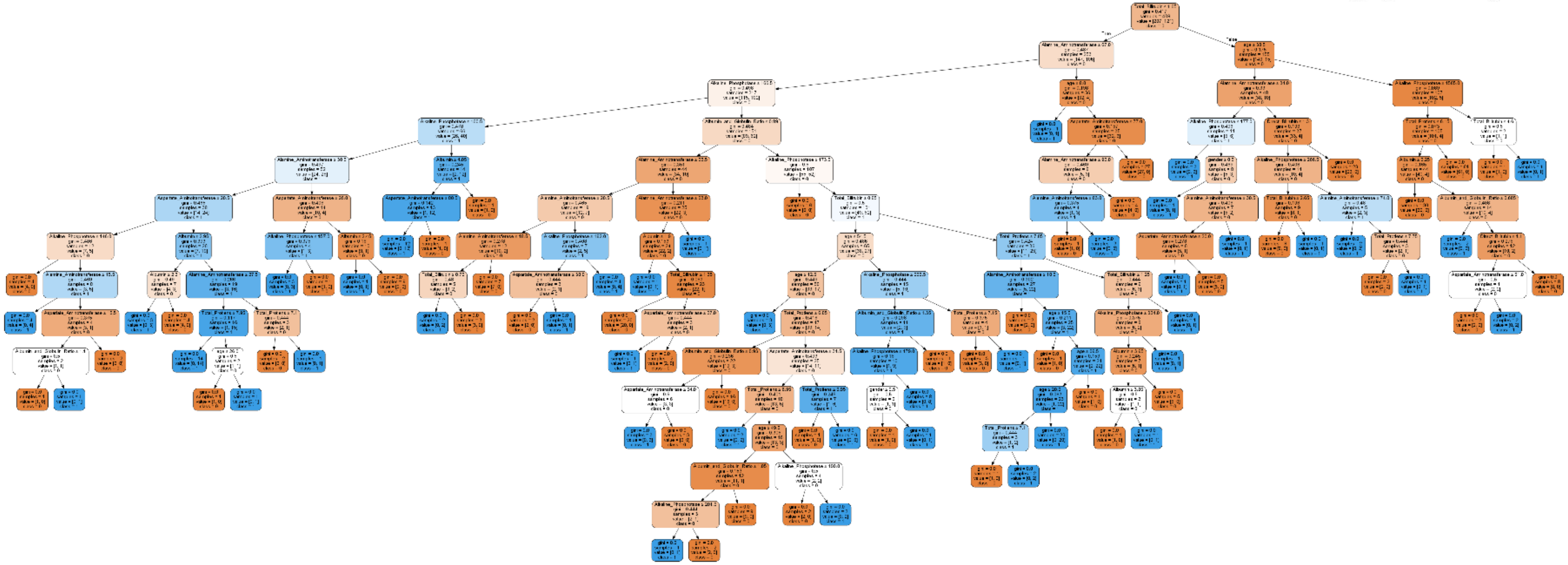

4000
3000
2000
1000
0
Alamine_Aminotransferase
0
1000
2000
3000
4000
5000
Aspartate_Aminotransferase

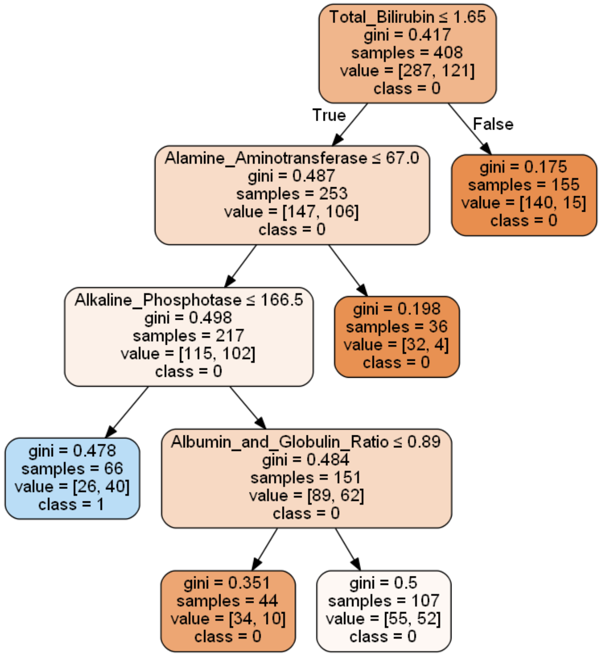

gender = 0
gender = 1
class = 1.0
class = 2.0
5
4
3
2
1
Albumin
4
6
8
10
Total_Protiens


6
5
4
3
2
1
0
Albumin
2
4
6
8
10
Total_Protiens



	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	class
Total_Bilirubin	1	0.87	0.21	0.21	0.24	-0.0081	-0.22	0.032	-0.22
Direct_Bilirubin	0.87	1	0.23	0.23	0.26	-0.00014	-0.23	0.034	-0.25
Alkaline_Phosphotase	0.21	0.23	1	0.13	0.17	-0.029	-0.17	0.039	-0.18
Alamine_Aminotransferase	0.21	0.23	0.13	1	0.79	-0.043	-0.03	0.027	-0.16
Aspartate_Aminotransferase	0.24	0.26	0.17	0.79	1	-0.026	-0.085	0.021	-0.15
Total_Protiens	-0.0081	-0.00014	-0.029	-0.043	-0.026	1	0.78	-0.017	0.035
Albumin	-0.22	-0.23	-0.17	-0.03	-0.085	0.78	1	-0.05	0.16
Albumin_and_Globulin_Ratio	0.032	0.034	0.039	0.027	0.021	-0.017	-0.05	1	-0.039
class	-0.22	-0.25	-0.18	-0.16	-0.15	0.035	0.16	-0.039	1





Total_Bilirubin ≤ 1.65
gini = 0.417
samples = 408
value = [287, 121]
class = 0

True

False

Alamine_Aminotransferase ≤ 67.0
gini = 0.487
samples = 253
value = [147, 106]
class = 0

gini = 0.175
samples = 155
value = [140, 15]
class = 0

Alkaline_Phosphotase ≤ 166.5
gini = 0.498
samples = 217
value = [115, 102]
class = 0

gini = 0.198
samples = 36
value = [32, 4]
class = 0

gini = 0.478
samples = 66
value = [26, 40]
class = 1

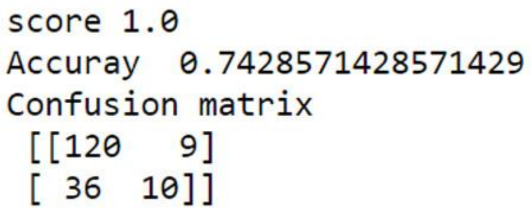
Albumin_and_Globulin_Ratio ≤ 0.89
gini = 0.484
samples = 151
value = [89, 62]
class = 0

gini = 0.351
samples = 44
value = [34, 10]
class = 0

gini = 0.5
samples = 107
value = [55, 52]
class = 0

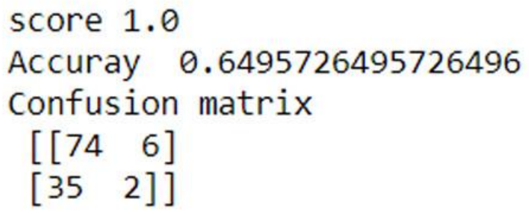
LOGISTIC REGRESSION

WITHOUT SCALING AND PCA



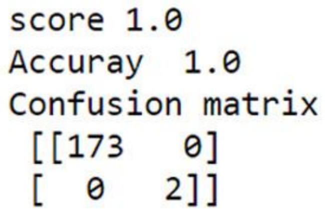
```
score 1.0
Accuray  0.7428571428571429
Confusion matrix
 [[120   9]
 [ 36  10]]
```

WITH SCALING AND PCA

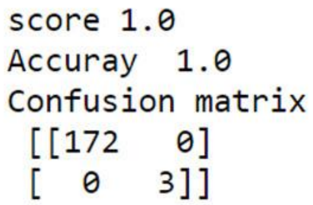


```
score 1.0
Accuray  0.6495726495726496
Confusion matrix
 [[74  6]
 [35  2]]
```

KMEANS & Agglomerative

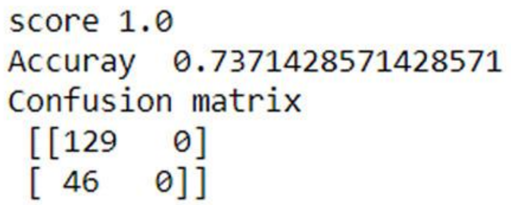


```
score 1.0
Accuray  1.0
Confusion matrix
 [[173   0]
 [  0   2]]
```



```
score 1.0
Accuray  1.0
Confusion matrix
 [[172   0]
 [  0   3]]
```

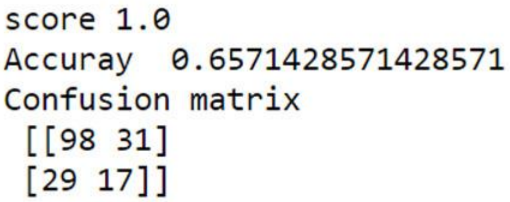
WITH PCA BUT NOT SCALING



```
score 1.0
Accuray  0.7371428571428571
Confusion matrix
 [[129   0]
 [ 46   0]]
```

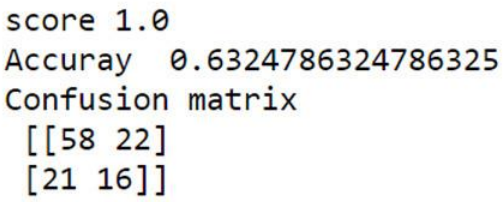
Decision tree (Entropy)

WITHOUT SCALING AND PCA



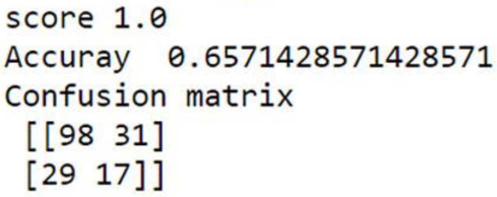
```
score 1.0
Accuray  0.6571428571428571
Confusion matrix
 [[98 31]
 [29 17]]
```

WITH SCALING AND PCA

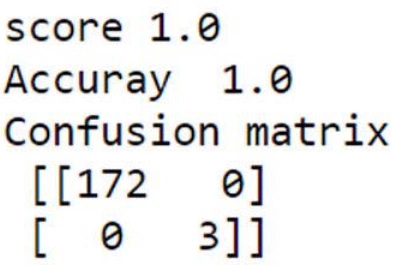


```
score 1.0
Accuray  0.6324786324786325
Confusion matrix
 [[58 22]
 [21 16]]
```

KMEANS & Agglomerative

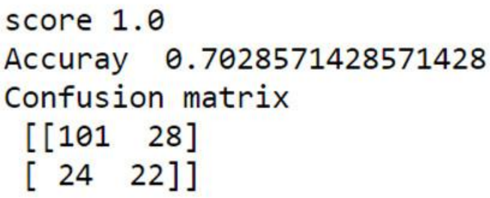


```
score 1.0
Accuray  0.6571428571428571
Confusion matrix
 [[98 31]
 [29 17]]
```



```
score 1.0
Accuray  1.0
Confusion matrix
 [[172   0]
 [  0   3]]
```

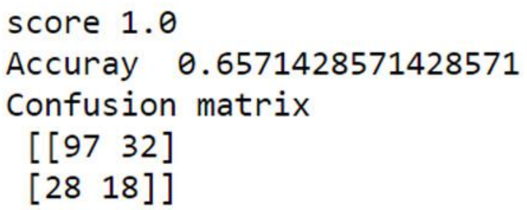
WITH PCA BUT NOT SCALING



```
score 1.0
Accuray  0.7028571428571428
Confusion matrix
 [[101  28]
 [ 24  22]]
```

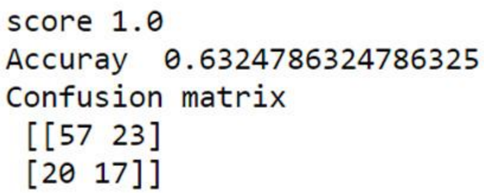
Decision tree (Gini)

WITHOUT SCALING AND PCA



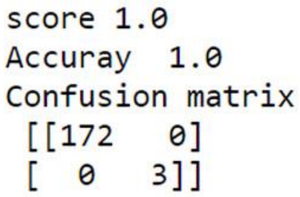
```
score 1.0
Accuray  0.6571428571428571
Confusion matrix
 [[97 32]
 [28 18]]
```

WITH SCALING AND PCA

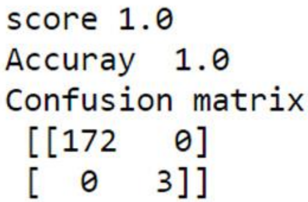


```
score 1.0
Accuray  0.6324786324786325
Confusion matrix
 [[57 23]
 [20 17]]
```

KMEANS & Agglomerative

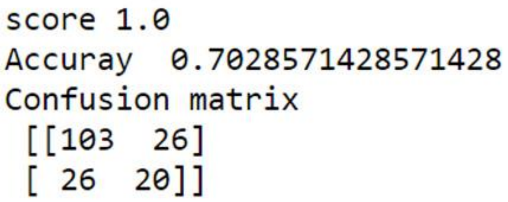


```
score 1.0
Accuray  1.0
Confusion matrix
 [[172   0]
 [  0   3]]
```



```
score 1.0
Accuray  1.0
Confusion matrix
 [[172   0]
 [  0   3]]
```

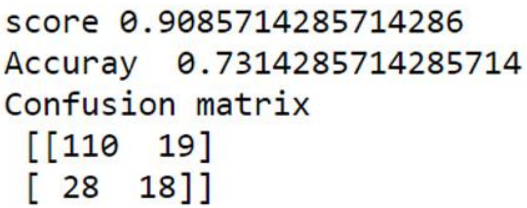
WITH PCA BUT NOT SCALING



```
score 1.0
Accuray  0.7028571428571428
Confusion matrix
 [[103  26]
 [ 26  20]]
```

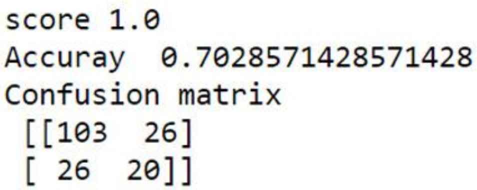
Random Forest

WITHOUT SCALING AND PCA



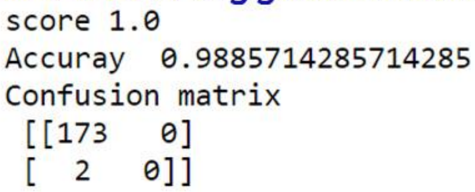
```
score 0.9085714285714286
Accuray  0.7314285714285714
Confusion matrix
 [[110  19]
 [ 28  18]]
```

WITH SCALING AND PCA

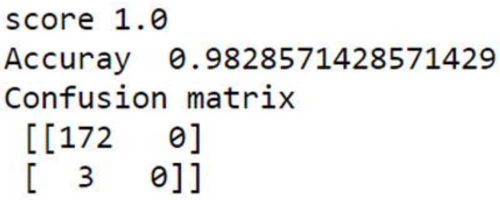


```
score 1.0
Accuray  0.7028571428571428
Confusion matrix
 [[103  26]
 [ 26  20]]
```

KMEANS & Agglomerative

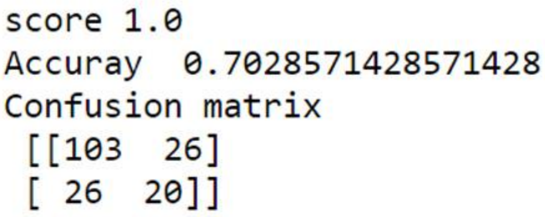


```
score 1.0
Accuray  0.9885714285714285
Confusion matrix
 [[173   0]
 [  2   0]]
```



```
score 1.0
Accuray  0.9828571428571429
Confusion matrix
 [[172   0]
 [  3   0]]
```

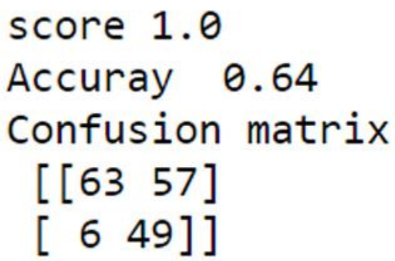
WITH PCA BUT NOT SCALING



```
score 1.0
Accuray  0.7028571428571428
Confusion matrix
 [[103  26]
 [ 26  20]]
```

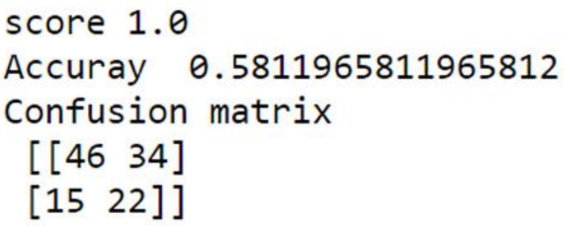
Naive Bayes

WITHOUT SCALING AND PCA



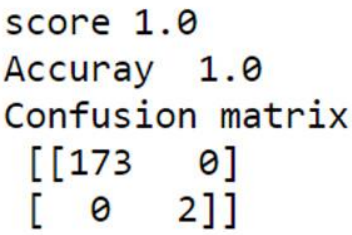
```
score 1.0
Accuray  0.64
Confusion matrix
 [[63 57]
 [ 6 49]]
```

WITH SCALING AND PCA

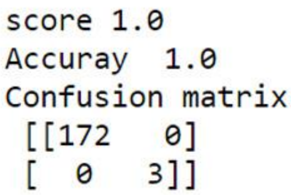


```
score 1.0
Accuray  0.5811965811965812
Confusion matrix
 [[46 34]
 [15 22]]
```

KMEANS & Agglomerative

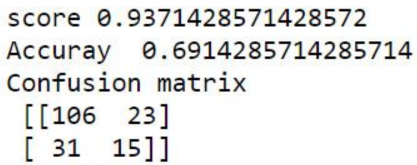


```
score 1.0
Accuray  1.0
Confusion matrix
 [[173   0]
 [  0   2]]
```



```
score 1.0
Accuray  1.0
Confusion matrix
 [[172   0]
 [  0   3]]
```

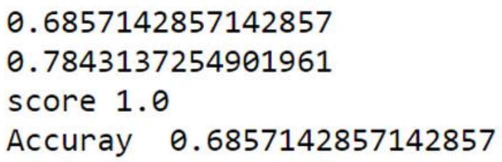
WITH PCA BUT NOT SCALING



```
score 0.9371428571428572
Accuray  0.6914285714285714
Confusion matrix
 [[106  23]
 [ 31  15]]
```

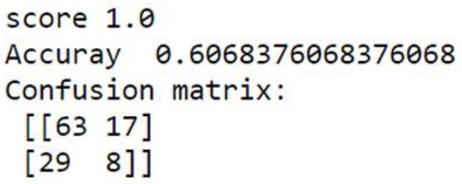
K neighbors

WITHOUT SCALING AND PCA



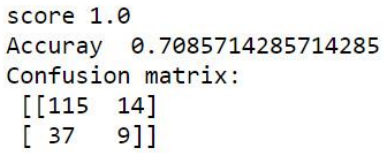
```
0.6857142857142857
0.7843137254901961
score 1.0
Accuray  0.6857142857142857
```

WITH SCALING AND PCA

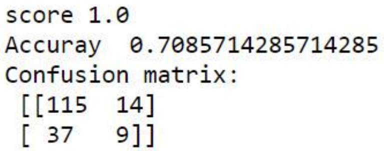


```
score 1.0
Accuray  0.6068376068376068
Confusion matrix:
 [[63 17]
 [29  8]]
```

KMEANS & Agglomerative

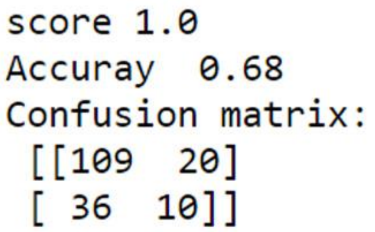


```
score 1.0
Accuray  0.7085714285714285
Confusion matrix:
 [[115  14]
 [ 37   9]]
```



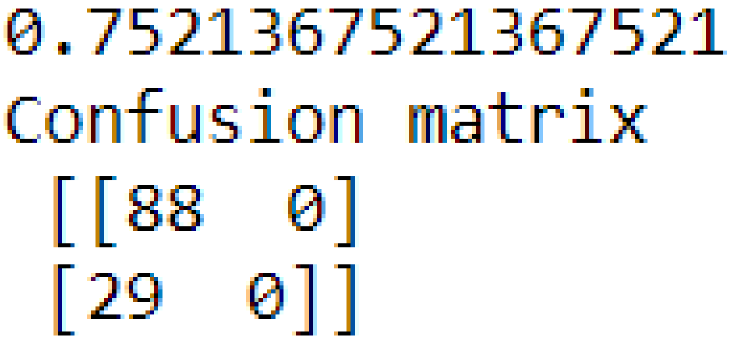
```
score 1.0
Accuray  0.7085714285714285
Confusion matrix:
 [[115  14]
 [ 37   9]]
```

WITH PCA BUT NOT SCALING



```
score 1.0
Accuray  0.68
Confusion matrix:
 [[109  20]
 [ 36  10]]
```

Support Vector Machine



```
0.7521367521367521
Confusion matrix
 [[88  0]
 [29  0]]
```

Inference

- logistic regression accuracy is highest without scaling and PCA - 0.74285
- Decision tree (entropy)accuracy is highest without scaling but with PCA -0.70285
- Decision tree (gini)accuracy is highest without scaling but with PCA -0.70285

- Random forest accuracy is highest without scaling and PCA - 0.731
- Naive Bayes accuracy is highest without scaling but with PCA -0.6914

- K Neighbours accuracy is highest without scaling but with PCA -0.68
- Support Vector machine gives accuracy of 0.752136 but false negative is higher in the model and true negative is 0. This model for this dataset should be ignored.

- Overall, for this dataset, Logistic regression yields better accuracy with 0.74285