CS 60050 MACHINE LEARING

Assignment 1

Question 2. Bayesian Classifier

Provided the data that has list of contains 416 liver patient records and 167 non-liver patient records. The data set was collected from test samples in North East of Andhra Pradesh, India. 'is_patient' is a class label used to divide into groups (liver patient or not). 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".

A Bayesian Classifier model has to be built that classifies if a given person with below attribute details is a patient or not.

Attribute Name	**Definition**
age	Age of the patient
gender	Gender of the patient
tot_bilirubin	Total Bilirubin
direct_bilirubin	Direct Bilirubin
alkphos	Alkaline Phosphotase
sgpt	Alamine Aminotransferase
sgot	Aspartate Aminotransferase
tot_proteins	Total Protiens
albumin	Albumin
ag_ratio	Albumin and Globulin Ratio

Results of the model built:

1.Final number of rows obtained after removing samples that has outlier if its value is greater than 2 x mean + 5 x standard deviation ($\mathbf{2} \times \boldsymbol{\mu} + \mathbf{5} \times \boldsymbol{\sigma}$) is found to be 568.

```
In [314]: print(len(df))
Out[314]: 568

In [337]: print(df)
     age gender tot_bilirubin direct_bilirubin tot_proteins albumin ag_ratio  \
0      4      5             1                1            1       1         1
1      4      1             2                2            2       1         1
2      4      1             2                2            2       1         1
3      4      1             1                1            1       1         1
4      4      1             1                1            1       1         1
..   ...    ...           ...              ...          ...     ...       ...
578    4      1             1                1            2       1         1
579    3      1             1                1            1       1         1
580    3      1             1                1            1       1         1
581    2      1             1                1            1       1         1
582    2      1             1                1            1       1         1

     sgpt sgot alkphos  is_patient
0       3    3        2           1
1       4    3        1           1
2       4    3        2           1
3       3    3        2           1
4       4    2        1           1
..    ...  ...      ...         ...
578     3    1        1           0
579     3    3        2           1
580     3    3        2           1
581     3    3        2           1
582     4    4        3           0

[568 rows x 11 columns]
```

2. Accuracy without Laplace Correction using 5-fold cross validation is found to be as: **0.5661**

```
In [368]: print(sum(accuracy)/len(accuracy))
          0.566150246528308

In [329]: df
```

3. Final accuracy after Laplace correction is found to be as **0.736**

```
In [336]: X_test = test.iloc[:,:-1].values
          Y_test = test.iloc[:,-1].values
          Y_pred = naive_bayes_categorical(train, X=X_test, Y="is_patient")
          print(accuracy_score(Y_test, Y_pred))

          0.7368421052631579
```