INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

CS60050 – Machine Learning

Assignment 1
Naïve Bayes Classifier

Submitted by:
Jothi Prakash (19EC39023)
Tushar Kishore Bokade (19CS30011)

Theory

Naive Bayes Classifier:

A naive Bayes classifier is an algorithm that uses Bayes’ theorem to classify objects. Naive Bayes classifiers assume naive independence between attributes of data points. Popular uses of naive Bayes classifiers include spam filters, text analysis and medical diagnosis.

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

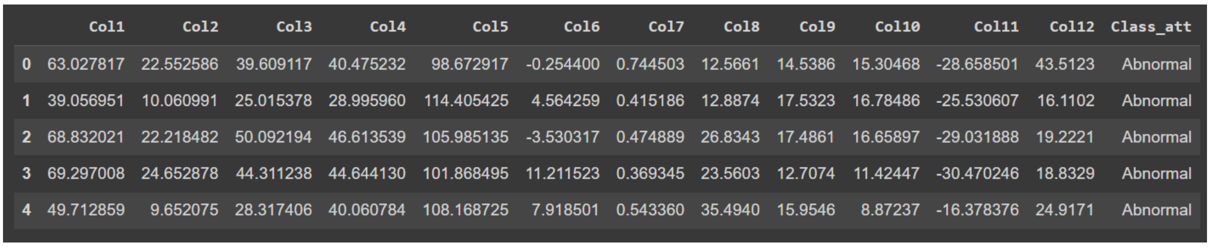
Laplace Correction:

In order to tackle the problem of zero probability in the Naive Bayes algorithm we add alpha to the probability, this is called Laplace Correction.

In our case, whenever the Guassian probability distribution function gives probability less than 0.0001, we add alpha (=0.0001) to the probability.

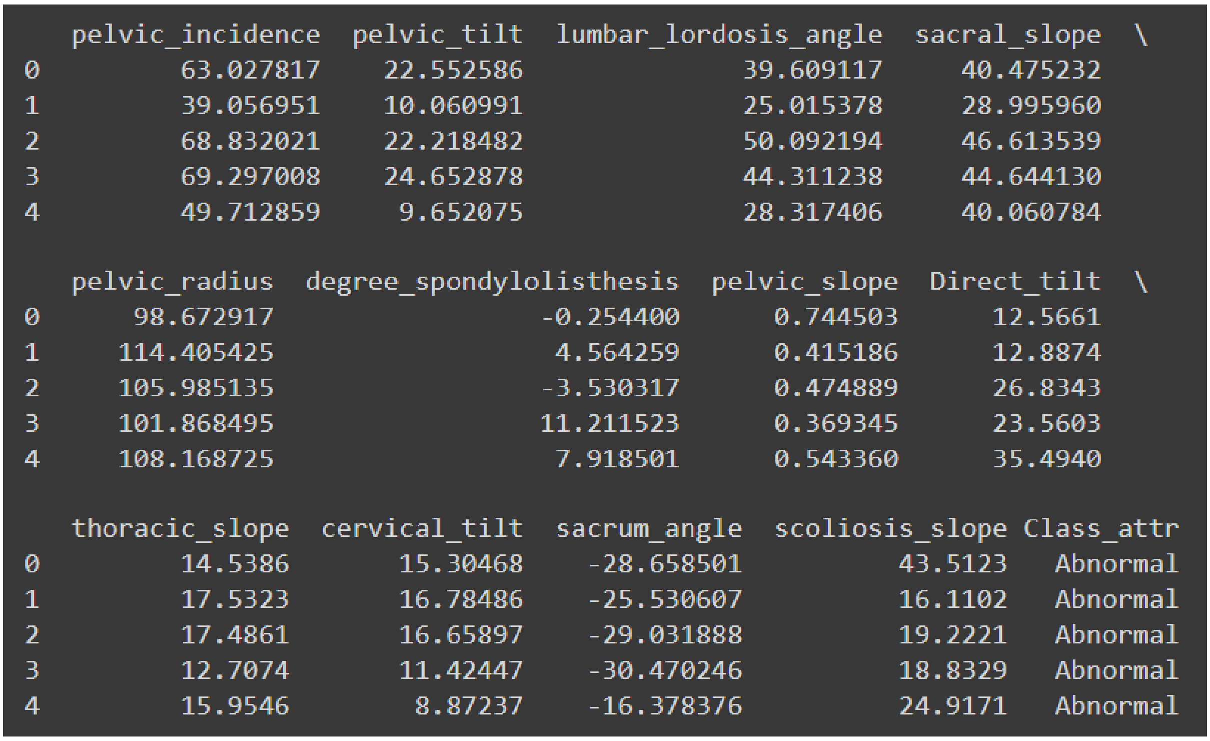
Experimental Procedure

1. Import the required libraries
2. Read the input from the given file



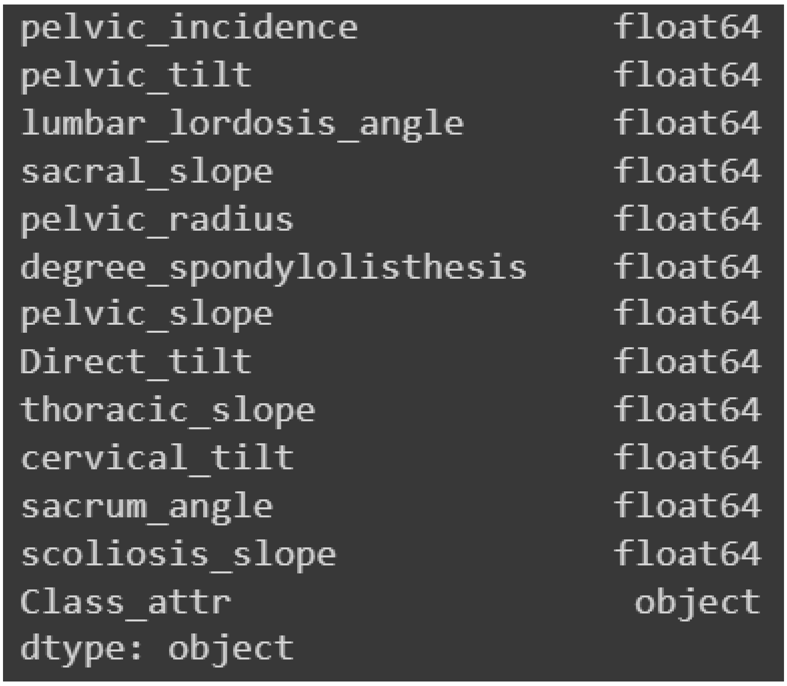
	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9	Col10	Col11	Col12	Class_att
0	63.027817	22.552586	39.609117	40.475232	98.672917	-0.254400	0.744503	12.5661	14.5386	15.30468	-28.658501	43.5123	Abnormal
1	39.056951	10.060991	25.015378	28.995960	114.405425	4.564259	0.415186	12.8874	17.5323	16.78486	-25.530607	16.1102	Abnormal
2	68.832021	22.218482	50.092194	46.613539	105.985135	-3.530317	0.474889	26.8343	17.4861	16.65897	-29.031888	19.2221	Abnormal
3	69.297008	24.652878	44.311238	44.644130	101.868495	11.211523	0.369345	23.5603	12.7074	11.42447	-30.470246	18.8329	Abnormal
4	49.712859	9.652075	28.317406	40.060784	108.168725	7.918501	0.543360	35.4940	15.9546	8.87237	-16.378376	24.9171	Abnormal

3. Renaming the columns



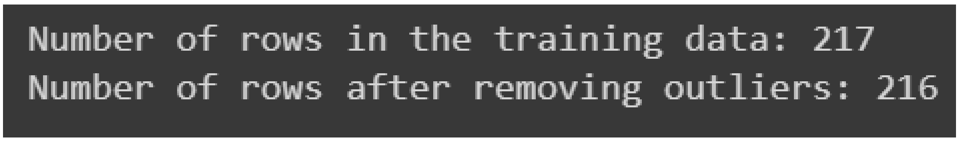
```
   pelvic_incidence  pelvic_tilt  lumbar_lordosis_angle  sacral_slope  \
0         63.027817    22.552586              39.609117     40.475232   
1         39.056951    10.060991              25.015378     28.995960   
2         68.832021    22.218482              50.092194     46.613539   
3         69.297008    24.652878              44.311238     44.644130   
4         49.712859     9.652075              28.317406     40.060784   

   pelvic_radius  degree_spondylolisthesis  pelvic_slope  Direct_tilt  \
0      98.672917                 -0.254400      0.744503      12.5661   
1     114.405425                  4.564259      0.415186      12.8874   
2     105.985135                 -3.530317      0.474889      26.8343   
3     101.868495                 11.211523      0.369345      23.5603   
4     108.168725                  7.918501      0.543360      35.4940   

   thoracic_slope  cervical_tilt  sacrum_angle  scoliosis_slope Class_attr  
0         14.5386       15.30468    -28.658501          43.5123   Abnormal  
1         17.5323       16.78486    -25.530607          16.1102   Abnormal  
2         17.4861       16.65897    -29.031888          19.2221   Abnormal  
3         12.7074       11.42447    -30.470246          18.8329   Abnormal  
4         15.9546        8.87237    -16.378376          24.9171   Abnormal  
```

4. Printing data types



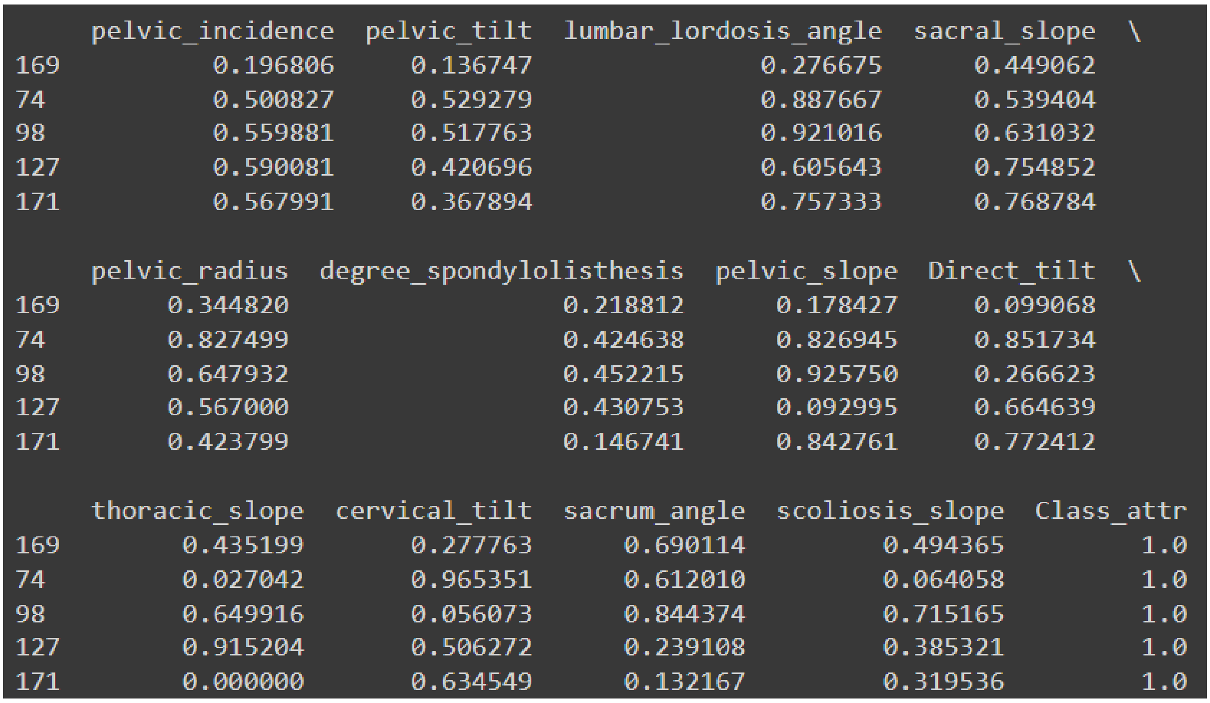
```
pelvic_incidence            float64
pelvic_tilt                 float64
lumbar_lordosis_angle       float64
sacral_slope                float64
pelvic_radius               float64
degree_spondylolisthesis    float64
pelvic_slope                float64
Direct_tilt                 float64
thoracic_slope              float64
cervical_tilt               float64
sacrum_angle                float64
scoliosis_slope             float64
Class_attr                   object
dtype: object
```

5. Encoding categorical variables (Replacing Abnormal with 1 and Normal with 0)
6. Dividing data into training data and testing data in 70:30 ratio.
7. Removing Outliers

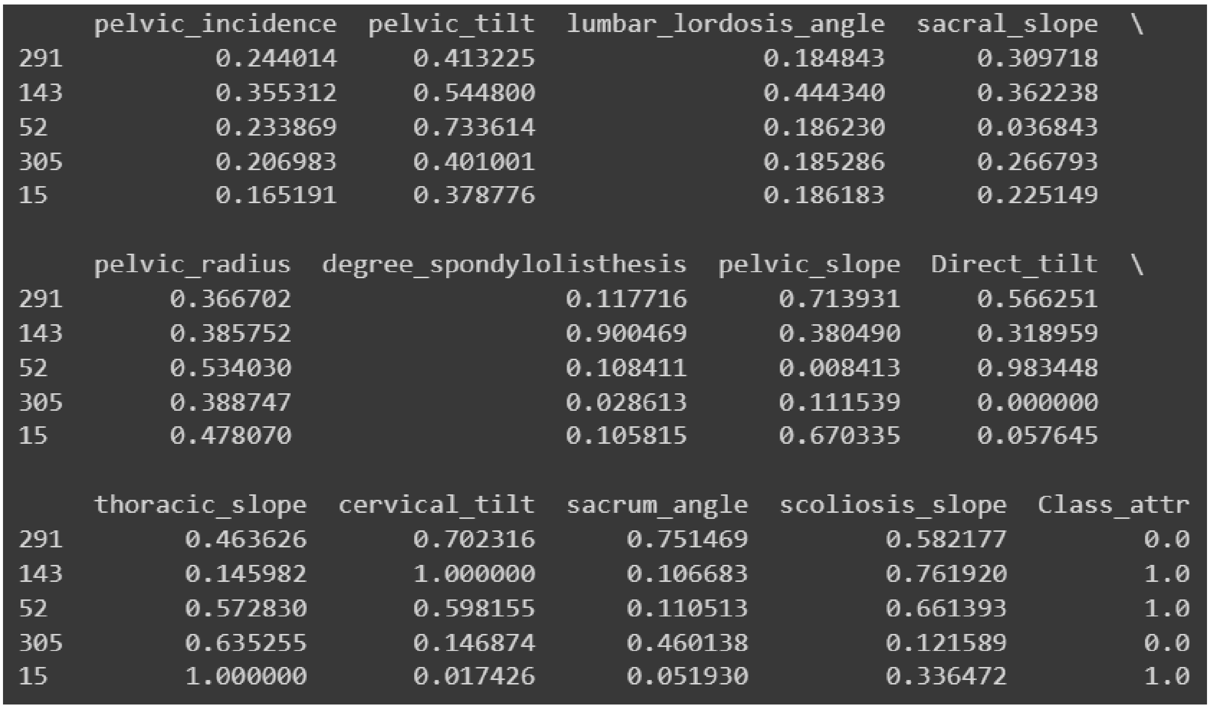


```
Number of rows in the training data: 217
Number of rows after removing outliers: 216
```

8. Normalise training data and testing data

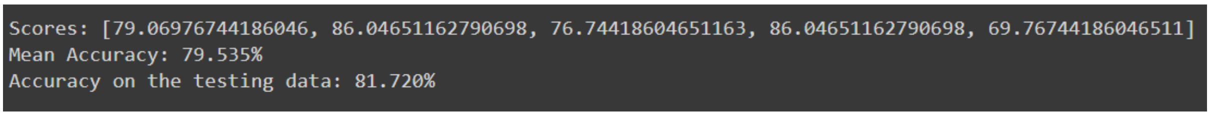


```
     pelvic_incidence  pelvic_tilt  lumbar_lordosis_angle  sacral_slope  \
169          0.196806     0.136747               0.276675      0.449062   
74           0.500827     0.529279               0.887667      0.539404   
98           0.559881     0.517763               0.921016      0.631032   
127          0.590081     0.420696               0.605643      0.754852   
171          0.567991     0.367894               0.757333      0.768784   

     pelvic_radius  degree_spondylolisthesis  pelvic_slope  Direct_tilt  \
169       0.344820                  0.218812      0.178427     0.099068   
74        0.827499                  0.424638      0.826945     0.851734   
98        0.647932                  0.452215      0.925750     0.266623   
127       0.567000                  0.430753      0.092995     0.664639   
171       0.423799                  0.146741      0.842761     0.772412   

     thoracic_slope  cervical_tilt  sacrum_angle  scoliosis_slope  Class_attr  
169        0.435199       0.277763      0.690114         0.494365         1.0  
74         0.027042       0.965351      0.612010         0.064058         1.0  
98         0.649916       0.056073      0.844374         0.715165         1.0  
127        0.915204       0.506272      0.239108         0.385321         1.0  
171        0.000000       0.634549      0.132167         0.319536         1.0  
```



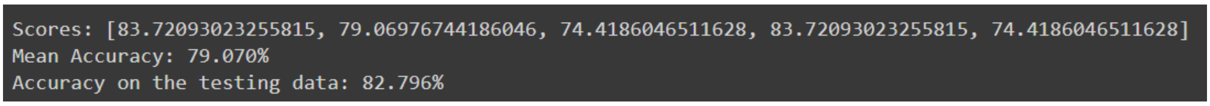
```
     pelvic_incidence  pelvic_tilt  lumbar_lordosis_angle  sacral_slope  \
291          0.244014     0.413225               0.184843      0.309718   
143          0.355312     0.544800               0.444340      0.362238   
52           0.233869     0.733614               0.186230      0.036843   
305          0.206983     0.401001               0.185286      0.266793   
15           0.165191     0.378776               0.186183      0.225149   

     pelvic_radius  degree_spondylolisthesis  pelvic_slope  Direct_tilt  \
291       0.366702                  0.117716      0.713931     0.566251   
143       0.385752                  0.900469      0.380490     0.318959   
52        0.534030                  0.108411      0.008413     0.983448   
305       0.388747                  0.028613      0.111539     0.000000   
15        0.478070                  0.105815      0.670335     0.057645   

     thoracic_slope  cervical_tilt  sacrum_angle  scoliosis_slope  Class_attr  
291        0.463626       0.702316      0.751469         0.582177         0.0  
143        0.145982       1.000000      0.106683         0.761920         1.0  
52         0.572830       0.598155      0.110513         0.661393         1.0  
305        0.635255       0.146874      0.460138         0.121589         0.0  
15         1.000000       0.017426      0.051930         0.336472         1.0  
```

9. Using k split cross validation and applying naive bayes algorithm



```
Scores: [79.06976744186046, 86.04651162790698, 76.74418604651163, 86.04651162790698, 69.76744186046511]
Mean Accuracy: 79.535%
Accuracy on the testing data: 81.720%
```

10. Applying Laplace correction



```
Scores: [83.72093023255815, 79.06976744186046, 74.4186046511628, 83.72093023255815, 74.4186046511628]
Mean Accuracy: 79.070%
Accuracy on the testing data: 82.796%
```

Results

1. Applying naive bayes classifier on the given data with 5 fold cross validation split on the training data gives an average accuracy between 70% - 85%.
2. The accuracy for a single fold can go upto 95%.
3. When using the Laplace correction on the same training and testing data, the accuracy deviates by 0% - 5%.