Student Declaration of Authorship

HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

Course code and name:	F21DL: Data Mining and Machine Learning
Type of assessment:	**Group**
Coursework Title:	Part 1. Data Analysis and Bayes Nets
Student Name:	Arshati Ajay Marchande,
Student ID Number:	H00382093

Declaration of authorship. By signing this form:

- **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.

- I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.

- I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism

Student Signature *(type your name):* *Arshati*

Date: *22/10/2023*

Student Declaration of Authorship

HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

Course code and name:	F21DL: Data Mining and Machine Learning
Type of assessment:	**Group**
Coursework Title:	Part 1. Data Analysis and Bayes Nets
Student Name:	Asmitha Krishnakumar
Student ID Number:	H00376043

Declaration of authorship. By signing this form:

- **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.

- I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.

- I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism

Student Signature *(type your name):* *Asmitha*

Date: *22/10/2023*

Student Declaration of Authorship

HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

Course code and name:	F21DL: Data Mining and Machine Learning
Type of assessment:	**Group**
Coursework Title:	Part 1. Data Analysis and Bayes Nets
Student Name:	Gauri Revankar
Student ID Number:	H00373987

Declaration of authorship. By signing this form:

- **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.

- I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.

- I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism

Student Signature *(type your name):* *Gauri*

Date: *22/10/2023*

Student Declaration of Authorship

HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

Course code and name:	F21DL: Data Mining and Machine Learning
Type of assessment:	**Group**
Coursework Title:	Part 1. Data Analysis and Bayes Nets
Student Name:	Pooja Sheni Meledath
Student ID Number:	H00386700

Declaration of authorship. By signing this form:

- **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.

- I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.

- I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism

Student Signature *(type your name):* *Pooja*

Date: *22/10/2023*

Student Declaration of Authorship

HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

Course code and name:	F21DL: Data Mining and Machine Learning
Type of assessment:	**Group**
Coursework Title:	Part 1. Data Analysis and Bayes Nets
Student Name:	Prasitha Prasanna Naidu
Student ID Number:	H00379641

Declaration of authorship. By signing this form:

- **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.

- I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.

- I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism

Student Signature *(type your name):* *Prasitha*

Date: *22/10/2023*

DMML Coursework Part 1 - REPORT

22.10.2023

Arshati Ajay, Asmitha Krishnakumar, Gauri Revankar, Pooja Sheni, Prasitha Naidu

Group-5

Github Link:
https://github.com/poojameledath/F20DL_Group_5_DMML_Portfolio

Group Contribution

Data Visualization and Exploration & Report - Prasitha and Arshati

Data Preprocessing & Report - Pooja and Asmitha

Feature Selection & Report - Pooja and Asmitha

Cross Validation and Naive Bayes Classifier & Report - Gauri

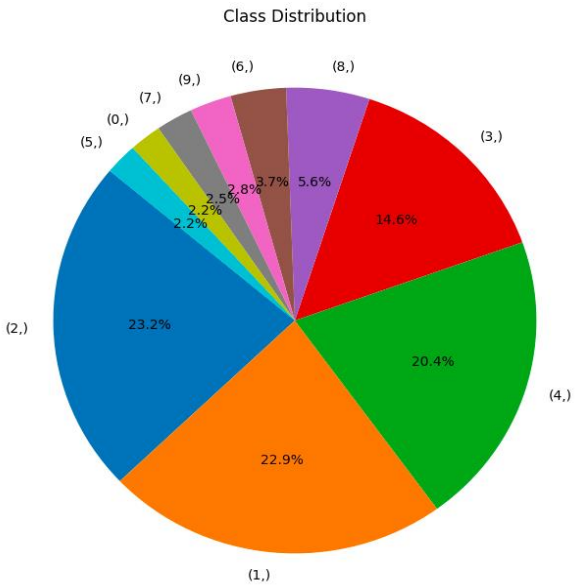
Confusion matrix & Report - Prasitha and Arshati

ROC curves & Area & Report - Prasitha and Arshati

Data Visualization and Exploration

```
Row ranges for each class label in x_train:
Class 0: Rows 0 to 209
Class 1: Rows 210 to 2429
Class 2: Rows 2430 to 4679
Class 3: Rows 4680 to 6089
Class 4: Rows 6090 to 8069
Class 5: Rows 8070 to 8279
Class 6: Rows 8280 to 8639
Class 7: Rows 8640 to 8879
Class 8: Rows 8880 to 9419
Class 9: Rows 9420 to 9689
```

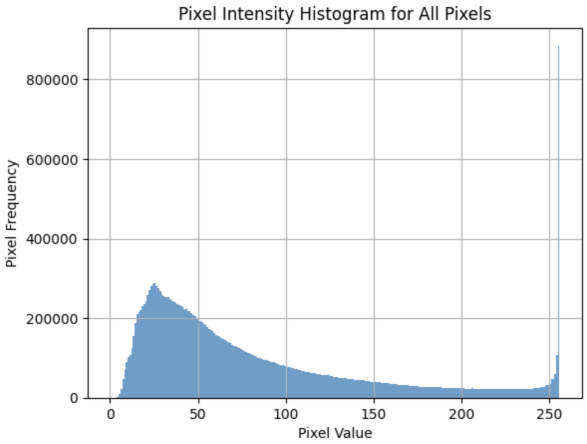
We executed the code to generate the row ranges for each class label in x_trainl.



The resulting pie chart represents the distribution of class labels in the dataset, showing the proportion of each class label relative to the whole dataset.

Data Preprocessing

During the initial phase of our preprocessing, we examined the dataset given using various functions available in the Pandas library, such as head, shape and info, to comprehend the dataset better.



To understand the distribution of the pixels, we plotted a Pixel Intensity Histogram. From this, we could infer that the highest pixel frequency lay between 0 to 50, and there is a sharp increase in 250. The pixel range from 0 to 50 indicates darker regions and low exposure in the images.

Image Enhancement

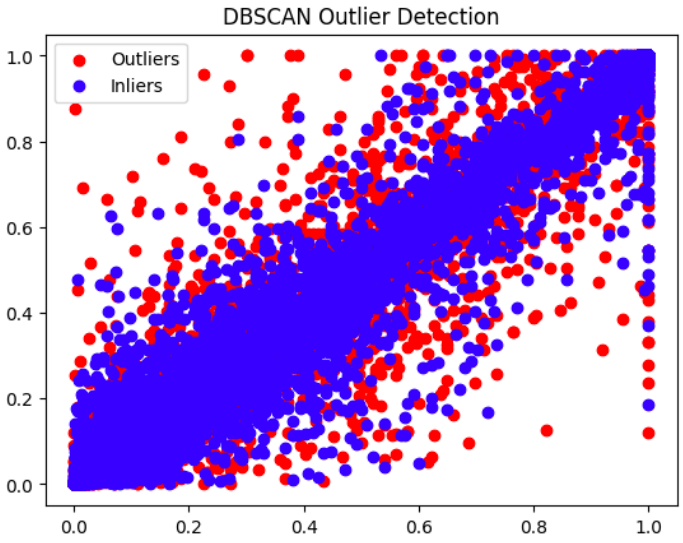
To improve the contrast of the greyscale images, we applied several image enhancement techniques, such as Histogram Equalization and Gamma Correction. Histogram equalization redistributes the pixel intensities more evenly across the image, resulting in an increase in brightness. Furthermore, Gamma Correction deals with a non-linear power law transformation to the pixels, making the images more perceptually uniform. By applying Gamma Correction on top of Histogram Equalisation, we had more control over the enhancement as we could fine-tune the parameter to achieve better contrast.

Normalizing, Missing Value and Outlier Mining

We normalized the dataset by dividing all the values by 255. By scaling pixel values, we could work effectively with our large dataset.

To verify whether the dataset has any null values, we checked using a Pandas function, `isna()`, provided to identify and handle missing values. After applying it, it became evident that no missing values exist within the dataset.

We used a clustering algorithm Density-Based Spatial Clustering of Applications with Noise (DBSCAN), to identify outliers in the dataset. DBSCAN mainly clusters the data points closely packed together in a high-density region. The data points that are not part of any grouping are considered outliers by DBSCAN. We were able to identify 3680 outliers in the dataset. Below is the visualization of the outliers in the dataset.



Feature Selection

To get the top 5, top 10 and top 20 features for each class, we used the SelectKBest library provided by scikit-learn. SelectKBest ranks the features according to their importance and retains the top features. We store all the unique features in top5, top10, and top20 datasets.

Cross Validation

In order to assess the performance and generalization ability of our chosen model, we have used K-Fold Cross Validation. In this type of cross validation, the training data is divided into k subsets (folds). The model is trained on k-1 folds and tested on the remaining fold. This process is repeated k times and the results are averaged.

Below are the cross validation evaluation results of the three models (the figures below show the accuracy of the model):

1) Gaussian NB

```
top 5 features:
Average cross-validation score: 0.5248
top 10 features:
Average cross-validation score: 0.5028
top 20 features:
Average cross-validation score: 0.5078
Complete train dataset:
Average cross-validation score: 0.3641
```

2) Multinomial NB

```
top 5 features:
Average cross-validation score: 0.5864
top 10 features:
Average cross-validation score: 0.5809
top 20 features:
Average cross-validation score: 0.5702
Complete train dataset:
Average cross-validation score: 0.3170
```

3) Complement NB

```
top 5 features:
Average cross-validation score: 0.5908
top 10 features:
Average cross-validation score: 0.5512
top 20 features:
Average cross-validation score: 0.5208
Complete train dataset:
Average cross-validation score: 0.3562
```

Naive Bayes Classifier

There are 3 Naive Bayes Classification Models used in the coursework:

1) Gaussian Naive Bayes

2) Multinomial Naive Bayes
3) Complement Naive Bayes

The evaluation metrics (Accuracy, Precision, Recall, Mean Absolute Error) after training the models are given as below:

1) Gaussian Naive Bayes

```
#training model with top 5 feature dataset
training_with_diff_datasets_GNB(df_5, train_y_df)
```

```
Training set score: 0.5614
Test set score: 0.0447

Model Accuracy Score : 0.0447
Precision Score : 0.0447
Recall Score : 0.0447
Mean Absolute Error : 3.1472
```

```
#training model with top 10 feature dataset
training_with_diff_datasets_GNB(df_10, train_y_df)
```

```
Training set score: 0.5448
Test set score: 0.0718

Model Accuracy Score : 0.0718
Precision Score : 0.0718
Recall Score : 0.0718
Mean Absolute Error : 2.9036
```

```
#training model with top 20 feature dataset
training_with_diff_datasets_GNB(df_20, train_y_df)
```

```
Training set score: 0.5461
Test set score: 0.1272

Model Accuracy Score : 0.1272
Precision Score : 0.1272
Recall Score : 0.1272
Mean Absolute Error : 2.2055
```

```
#training model complete train dataset
training_with_diff_datasets_GNB(normalized_df, train_y_df)
```

```
Training set score: 0.4359
Test set score: 0.1476

Model Accuracy Score : 0.1476
Precision Score : 0.1476
Recall Score : 0.1476
Mean Absolute Error : 1.6117
```

2) Multinomial Naive Bayes

```
#training model with top 5 feature dataset
training_with_diff_datasets_MNB(df_5, train_y_df)
```

```
Training set score: 0.6080
Test set score: 0.5919

Model Accuracy Score : 0.5919
Precision Score : 0.5919
Recall Score : 0.5919
Mean Absolute Error : 1.1243
```

```
#training model with top 10 feature dataset
training_with_diff_datasets_MNB(df_10, train_y_df)
```

```
Training set score: 0.6057
Test set score: 0.5107

Model Accuracy Score : 0.5107
Precision Score : 0.5107
Recall Score : 0.5107
Mean Absolute Error : 1.3243
```

```
#training model with top 20 feature dataset
training_with_diff_datasets_MNB(df_20, train_y_df)
```

```
Training set score: 0.5947
Test set score: 0.4832

Model Accuracy Score : 0.4832
Precision Score : 0.4832
Recall Score : 0.4832
Mean Absolute Error : 1.1883
```

```
#training model with complete train dataset
training_with_diff_datasets_MNB(normalized_df, train_y_df)
```

```
Training set score: 0.3686
Test set score: 0.2822

Model Accuracy Score : 0.2822
Precision Score : 0.2822
Recall Score : 0.2822
Mean Absolute Error : 1.6340
```

3) Complement Naive Bayes

```
#training model with top 5 feature dataset
training_with_diff_datasets_CNB(df_5, train_y_df)
```

```
Training set score: 0.6052
Test set score: 0.5233

Model Accuracy Score : 0.5233
Precision Score : 0.5233
Recall Score : 0.5233
Mean Absolute Error : 1.2557
```

```
#training model with top 10 feature dataset
training_with_diff_datasets_CNB(df_10, train_y_df)
```

```
Training set score: 0.5712
Test set score: 0.4524

Model Accuracy Score : 0.4524
Precision Score : 0.4524
Recall Score : 0.4524
Mean Absolute Error : 1.2492
```

```
#training model with top 20 feature dataset
training_with_diff_datasets_CNB(df_20, train_y_df)
```

```
Training set score: 0.5469
Test set score: 0.4207

Model Accuracy Score : 0.4207
Precision Score : 0.4207
Recall Score : 0.4207
Mean Absolute Error : 1.3016
```

```
#training model with complete train dataset
training_with_diff_datasets_CNB(normalized_df, train_y_df)
```

```
Training set score: 0.3882
Test set score: 0.2968

Model Accuracy Score : 0.2968
Precision Score : 0.2968
Recall Score : 0.2968
Mean Absolute Error : 1.4278
```

This shows that the models train best and give the best accuracies when trained with the dataset containing the top 5 features. The accuracies decrease when trained with top 10, top 20 and complete training dataset. This shows that in this case, feature selection enhances the accuracy of the models.

The Multinomial Naive Bayes Model gives the best accuracy of 60.80% and 59.19% with train and test set respectively (when trained with top 5 feature dataset).

The reason why we didn't choose the Bernoulli Naive Bayes Model is because it expects the samples of the dataset to have binary values. The dataset we are working on currently contains continuous values. Hence, it wouldn't be appropriate to use Bernoulli Model in this case.

Confusion Matrix

The confusion matrix for the **Gaussian Naive Bayes** classifier on the **top 5 feature** dataset is as follows:

```
TP (True Positive): [ 0  0  0 67  0  0 71  0  0  0]
TN (True Negative): [3030 2370 2340 2366 2430 3030  322 3030 2940 3000]
FN (False Negative): [   0    0    0  274    0    0 2678    0    0    0]
FP (False Positive): [ 60 720 750 383 660  60  19  60 150  90]
```

```
Sensitivity (true positive rate), for each class: [0.         0.         0.         0.14888889 0.         0.
 0.78888889 0.         0.         0.         ]
Specificity (true negative rate), for each class: [1.         1.         1.         0.89621212 1.         1.
 0.10733333 1.         1.         1.         ]

FP rate: [0.         0.         0.         0.10378788 0.         0.
 0.89266667 0.         0.         0.         ]
FN rate: [1.         1.         1.         0.85111111 1.         1.
 0.21111111 1.         1.         1.         ]
```

The confusion matrix for the **Gaussian Naive Bayes** classifier on the **top 10 feature** dataset is as follows:

```
TP (True Positive): [  0   0   0 162   0   0  60   0   0   0]
TN (True Negative): [3030 2370 2340 2104 2430 3030  668 3030 2940 3000]
FN (False Negative): [   0    0    0  536    0    0 2332    0    0    0]
FP (False Positive): [ 60 720 750 288 660  60  30  60 150  90]
```

```
Sensitivity (true positive rate), for each class: [0.         0.         0.         0.36       0.         0.
 0.66666667 0.         0.         0.         ]
Specificity (true negative rate), for each class: [1.         1.         1.         0.7969697  1.         1.
 0.22266667 1.         1.         1.         ]

FP rate: [0.         0.         0.         0.2030303  0.         0.
 0.77733333 0.         0.         0.         ]
FN rate: [1.         1.         1.         0.64       1.         1.
 0.33333333 1.         1.         1.         ]
```

The confusion matrix for the **Gaussian Naive Bayes** classifier on the **top 20 feature** dataset is as follows:

```
TP (True Positive): [  0   0   0 333   0   0  60   0   0   0]
TN (True Negative): [3025 2370 2340 1312 2430 3030 1636 3030 2940 3000]
FN (False Negative): [   5    0    0 1328    0    0 1364    0    0    0]
FP (False Positive): [ 60 720 750 117 660  60  30  60 150  90]
```

```
Sensitivity (true positive rate), for each class: [0.         0.         0.         0.74       0.         0.
 0.66666667 0.         0.         0.        ]
Specificity (true negative rate), for each class: [0.99834983 1.         1.         0.4969697  1.         1.
 0.54533333 1.         1.         1.        ]

FP rate: [0.00165017 0.         0.         0.5030303  0.         0.
 0.45466667 0.         0.         0.        ]
FN rate: [1.         1.         1.         0.26       1.         1.
 0.33333333 1.         1.         1.        ]
```

The confusion matrix for the **Gaussian Naive Bayes** classifier on the **complete** dataset is as follows:

```
TP (True Positive): [  0   0   0 434   0   0  22   0   0   0]
TN (True Negative): [2958 2366 2340  132 2430 3030 2950 3030 2940 3000]
FN (False Negative): [  72    4    0 2508    0    0   50    0    0    0]
FP (False Positive): [ 60 720 750  16 660  60  68  60 150  90]
```

```
Sensitivity (true positive rate), for each class: [0.         0.         0.         0.96444444 0.         0.
 0.24444444 0.         0.         0.        ]
Specificity (true negative rate), for each class: [0.97623762 0.99831224 1.         0.05       1.         1.
 0.98333333 1.         1.         1.        ]

FP rate: [0.02376238 0.00168776 0.         0.95       0.         0.
 0.01666667 0.         0.         0.        ]
FN rate: [1.         1.         1.         0.03555556 1.         1.
 0.75555556 1.         1.         1.        ]
```

The **Gaussian Naive Bayes** models perform well when trained with the dataset containing top 5 features with the highest sensitivity for some classes. With the inclusion of more features the model's performance remains suboptimal. This implies that, in this case, feature selection enhances accuracy of the model.

The confusion matrix for the **Multinomial Naive Bayes** classifier on the **top 5 feature** dataset is as follows:

```
TP (True Positive): [ 17 399 511 289 433  45  15  16  53  51]
TN (True Negative): [2954 2304 2233 2415 2156 2977 2716 2980 2876 2938]
FN (False Negative): [ 76  66 107 225 274  53 284  50  64  62]
FP (False Positive): [ 43 321 239 161 227  15  75  44  97  39]
```

```
Sensitivity (true positive rate), for each class: [0.28333333 0.55416667 0.68133333 0.64222222 0.65606061 0.75
 0.16666667 0.26666667 0.35333333 0.56666667]
Specificity (true negative rate), for each class: [0.97491749 0.9721519  0.9542735  0.91477273 0.8872428  0.98250825
 0.90533333 0.98349835 0.97823129 0.97933333]

FP rate: [0.02508251 0.0278481  0.0457265  0.08522727 0.1127572  0.01749175
 0.09466667 0.01650165 0.02176871 0.02066667]
FN rate: [0.71666667 0.44583333 0.31866667 0.35777778 0.34393939 0.25
 0.83333333 0.73333333 0.64666667 0.43333333]
```

The confusion matrix for the **Multinomial Naive Bayes** classifier on the **top 10 feature** dataset is as follows:

```
TP (True Positive): [ 10 330 477 217 369  32  15  20  39  69]
TN (True Negative): [2946 2324 2154 2290 2226 3001 2604 2995 2894 2864]
FN (False Negative): [ 84  46 186 350 204  29 396  35  46 136]
FP (False Positive): [ 50 390 273 233 291  28  75  40 111  21]
```

```
Sensitivity (true positive rate), for each class: [0.16666667 0.45833333 0.636      0.48222222 0.55909091 0.53333333
 0.16666667 0.33333333 0.26       0.76666667]
Specificity (true negative rate), for each class: [0.97227723 0.98059072 0.92051282 0.86742424 0.91604938 0.99042904
 0.868      0.98844884 0.98435374 0.95466667]

FP rate: [0.02772277 0.01940928 0.07948718 0.13257576 0.08395062 0.00957096
 0.132      0.01155116 0.01564626 0.04533333]
FN rate: [0.83333333 0.54166667 0.364      0.51777778 0.44090909 0.46666667
 0.83333333 0.66666667 0.74       0.23333333]
```

The confusion matrix for the **Multinomial Naive Bayes** classifier on the **top 20 feature** dataset is as follows:

```
TP (True Positive): [ 12 337 477 245 311  21  12  29  30  19]
TN (True Negative): [2834 2287 2118 2209 2326 3019 2700 2906 2848 2966]
FN (False Negative): [196  83 222 431 104  11 300 124  92  34]
FP (False Positive): [ 48 383 273 205 349  39  78  31 120  71]
```

```
Sensitivity (true positive rate), for each class: [0.2        0.46805556 0.636      0.54444444 0.47121212 0.35
 0.13333333 0.48333333 0.2        0.21111111]
Specificity (true negative rate), for each class: [0.93531353 0.9649789  0.90512821 0.83674242 0.95720165 0.99636964
 0.9        0.95907591 0.96870748 0.98866667]

FP rate: [0.06468647 0.0350211  0.09487179 0.16325758 0.04279835 0.00363036
 0.1        0.04092409 0.03129252 0.01133333]
FN rate: [0.8        0.53194444 0.364      0.45555556 0.52878788 0.65
 0.86666667 0.51666667 0.8        0.78888889]
```

The confusion matrix for the **Multinomial Naive Bayes** classifier on the **complete** dataset is as follows:

```
TP (True Positive): [  9 230 239 169 163   7   8  17  30   0]
TN (True Negative): [2671 1941 2072 2134 2304 3030 2731 2910 2806 2993]
FN (False Negative): [359 429 268 506 126   0 269 120 134   7]
FP (False Positive): [ 51 490 511 281 497  53  82  43 120  90]
```

```
Sensitivity (true positive rate), for each class: [0.15       0.31944444 0.31866667 0.37555556 0.2469697  0.11666667
 0.08888889 0.28333333 0.2        0.        ]
Specificity (true negative rate), for each class: [0.88151815 0.81898734 0.88547009 0.80833333 0.94814815 1.
 0.91033333 0.96039604 0.95442177 0.99766667]

FP rate: [0.11848185 0.18101266 0.11452991 0.19166667 0.05185185 0.
 0.08966667 0.03960396 0.04557823 0.00233333]
FN rate: [0.85       0.68055556 0.68133333 0.62444444 0.7530303  0.88333333
 0.91111111 0.71666667 0.8        1.        ]
```

Overall, the analysis indicates that the Multinomial Naive Bayes classifier performs differently for each class resulting in differences in sensitivity and specificity. Furthermore, with inclusion of more features such as top 5, top 10, top 20, and complete training dataset, the model's performance appears to improve for some classes but deteriorate for others.Further fine-tuning of the model or exploring alternative classification algorithms may be necessary to improve classification results, particularly for classes with lower sensitivity.

The confusion matrix for the **Complement Naive Bayes** classifier on the **top 5 feature** dataset is as follows:

```
TP (True Positive): [  0 264 502 307 479   0   1   0  64   0]
TN (True Negative): [3030 2345 2194 2172 1956 3030 2926 3030 2654 3000]
FN (False Negative): [  0  25 146 468 474   0  74   0 286   0]
FP (False Positive): [ 60 456 248 143 181  60  89  60  86  90]
```

```
Sensitivity (true positive rate), for each class: [0.         0.36666667 0.66933333 0.68222222 0.72575758 0.
 0.01111111 0.         0.42666667 0.         ]
Specificity (true negative rate), for each class: [1.         0.98945148 0.93760684 0.82272727 0.80493827 1.
 0.97533333 1.         0.90272109 1.         ]

FP rate: [0.         0.01054852 0.06239316 0.17727273 0.19506173 0.
 0.02466667 0.         0.09727891 0.         ]
FN rate: [1.         0.63333333 0.33066667 0.31777778 0.27424242 1.
 0.98888889 1.         0.57333333 1.         ]
```

The confusion matrix for the **Complement Naive Bayes** classifier on the **top 10 feature** dataset is as follows:

```
TP (True Positive): [  0 216 439 333 360   0   0   0  50   0]
TN (True Negative): [3030 2357 2159 1716 2094 3030 2975 3030 2727 3000]
FN (False Negative): [  0  13 181 924 336   0  25   0 213   0]
FP (False Positive): [ 60 504 311 117 300  60  90  60 100  90]
```

```
Sensitivity (true positive rate), for each class: [0.         0.3        0.58533333 0.74       0.54545455 0.
 0.         0.         0.33333333 0.         ]
Specificity (true negative rate), for each class: [1.         0.99451477 0.92264957 0.65       0.8617284  1.
 0.99166667 1.         0.92755102 1.         ]

FP rate: [0.         0.00548523 0.07735043 0.35       0.1382716  0.
 0.00833333 0.         0.07244898 0.         ]
FN rate: [1.         0.7        0.41466667 0.26       0.45454545 1.
 1.         1.         0.66666667 1.         ]
```

The confusion matrix for the **Complement Naive Bayes** classifier on the **top 20 feature** dataset is as follows:

```
TP (True Positive): [  0 215 410 353 290   0   0  11  21   0]
TN (True Negative): [3030 2351 2024 1458 2235 3030 2994 3024 2874 3000]
FN (False Negative): [   0   19  316 1182  195    0    6    6   66    0]
FP (False Positive): [ 60 505 340  97 370  60  90  49 129  90]
```

```
Sensitivity (true positive rate), for each class: [0.         0.29861111 0.54666667 0.78444444 0.43939394 0.
 0.         0.18333333 0.14       0.         ]
Specificity (true negative rate), for each class: [1.         0.99198312 0.86495726 0.55227273 0.91975309 1.
 0.998      0.9980198  0.97755102 1.         ]

FP rate: [0.         0.00801688 0.13504274 0.44772727 0.08024691 0.
 0.002      0.0019802  0.02244898 0.         ]
FN rate: [1.         0.70138889 0.45333333 0.21555556 0.56060606 1.
 1.         0.81666667 0.86       1.         ]
```

The confusion matrix for the **Complement Naive Bayes** classifier on the **complete** dataset is as follows:

```
TP (True Positive): [  0  92 337 341 133   0   0   0  14   0]
TN (True Negative): [3029 2265 1727 1256 2379 3030 3000 3030 2921 3000]
FN (False Negative): [   1  105  613 1384   51    0    0    0   19    0]
FP (False Positive): [ 60 628 413 109 527  60  90  60 136  90]
```

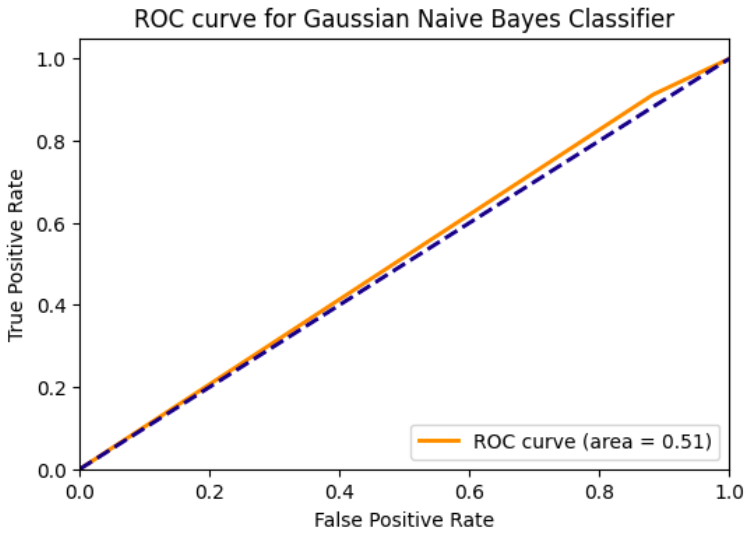
```
Sensitivity (true positive rate), for each class: [0.         0.12777778 0.44933333 0.75777778 0.20151515 0.
 0.         0.         0.09333333 0.        ]
Specificity (true negative rate), for each class: [0.99966997 0.9556962  0.73803419 0.47575758 0.97901235 1.
 1.         1.         0.99353741 1.        ]

FP rate: [3.30033003e-04 4.43037975e-02 2.61965812e-01 5.24242424e-01
 2.09876543e-02 0.00000000e+00 0.00000000e+00 0.00000000e+00
 6.46258503e-03 0.00000000e+00]
FN rate: [1.         0.87222222 0.55066667 0.24222222 0.79848485 1.
 1.         1.         0.90666667 1.        ]
```

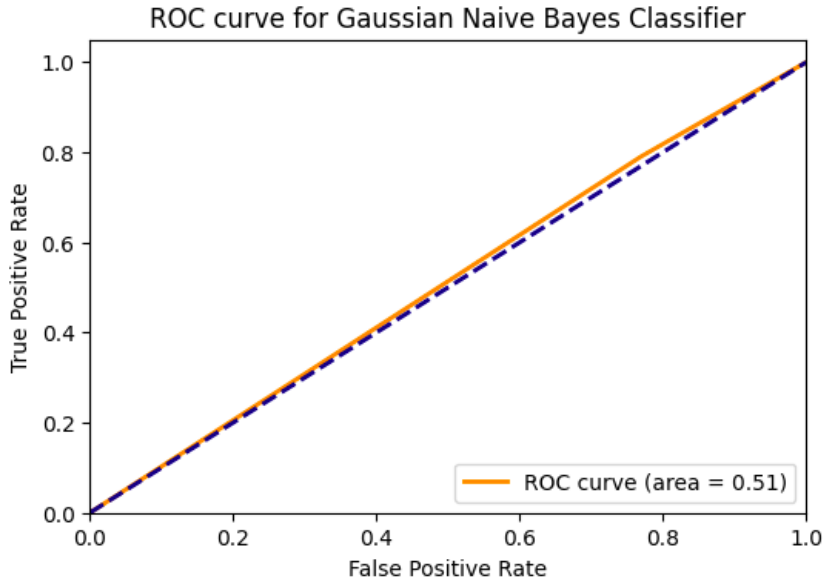
In conclusion, the Complement Naive Bayes classifier exhibits distinct performance variations across different classes, showing differences in sensitivity and specificity..

ROC Curve & Area Under ROC

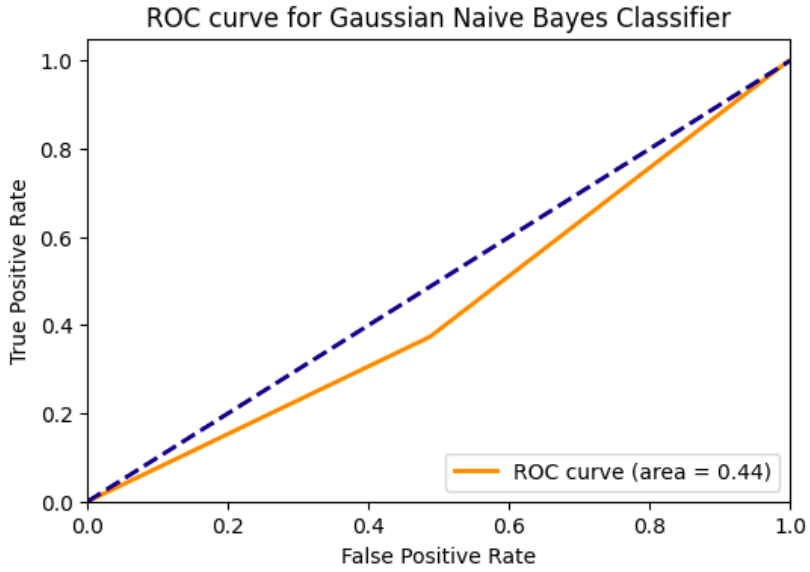
The ROC Curve and Area under the graph for the **Gaussian Naive Bayes** classifier on the **top 5 feature** dataset is as follow:



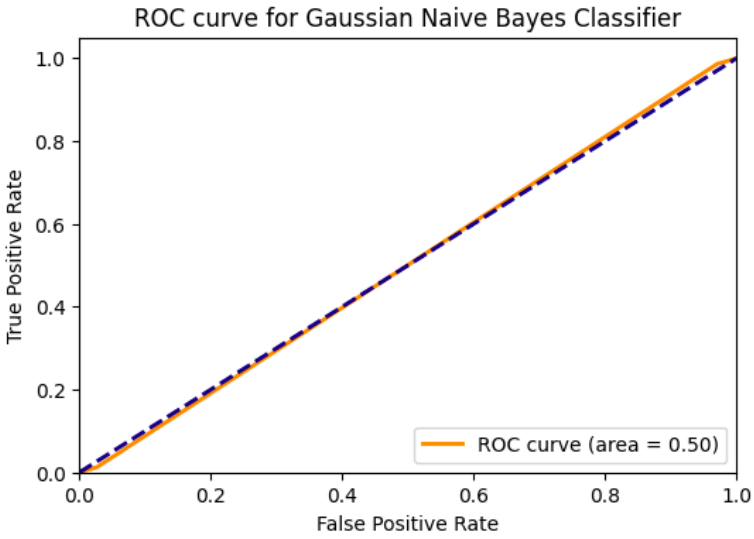
The ROC Curve and Area under the graph for the **Gaussian Naive Bayes** classifier on the **top 10 feature** dataset is as follow:



The ROC Curve and Area under the graph for the **Gaussian Naive Bayes** classifier on the **top 20 feature** dataset is as follow:

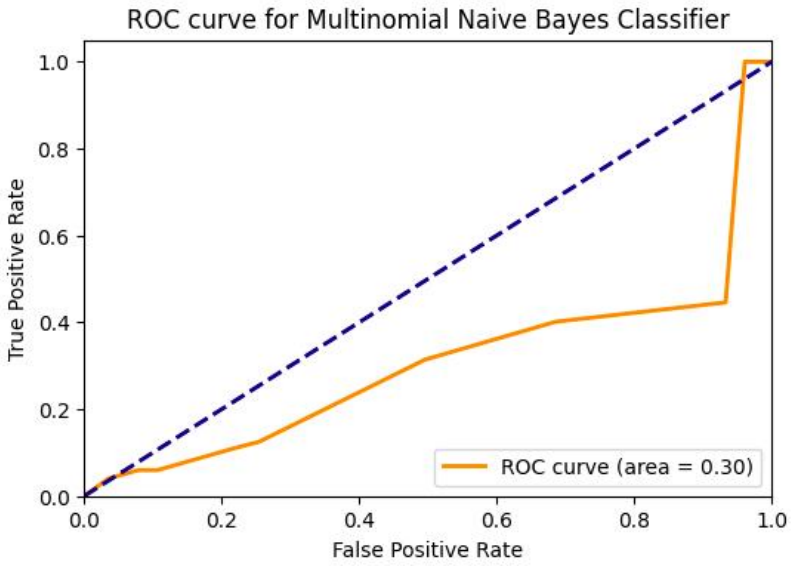


The ROC Curve and Area under the graph for the **Gaussian Naive Bayes** classifier on the **complete** dataset is as follow:

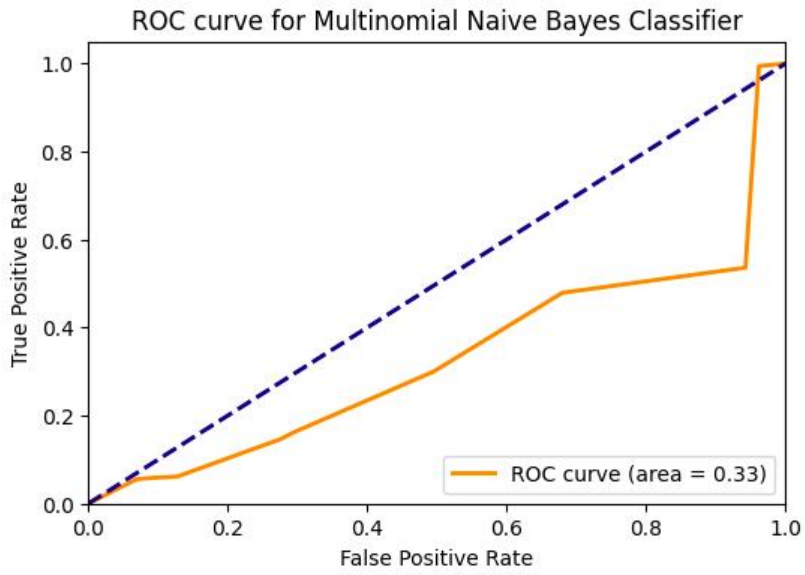


The ROC AUC value for all the feature datasets are reasonably close to each other, with values in the range of 0.44 to 0.51. The ROC AUC values for the top 5 feature dataset is 0.51 and for complete training dataset is 0.50. In the case of the top 10 and top 20 feature datasets, their ROC AUC scores are 0.51 and 0.44, respectively. These scores closely resemble the ROC AUC of the top 5 feature dataset, indicating that there is no significant boost in model performance with the inclusion of additional features.

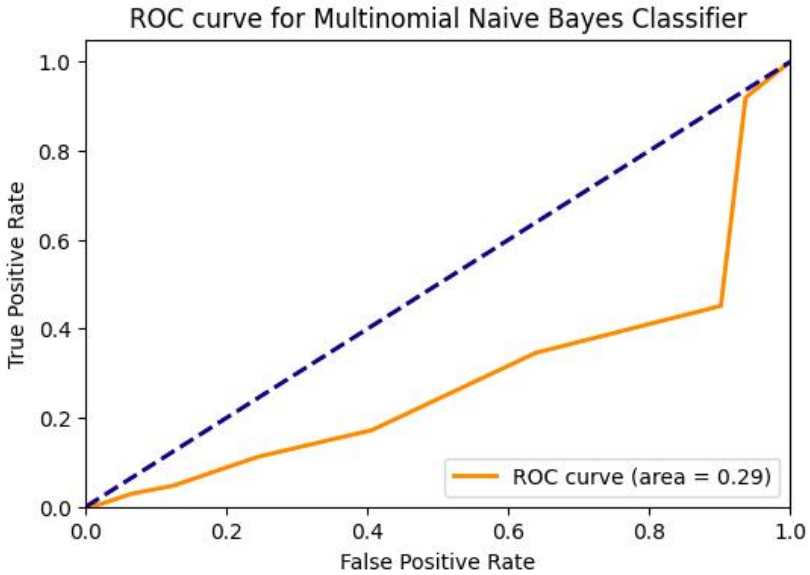
The ROC Curve and Area under the graph for the **Multinomial Naive Bayes** classifier on the **top 5 feature** dataset is as follow:



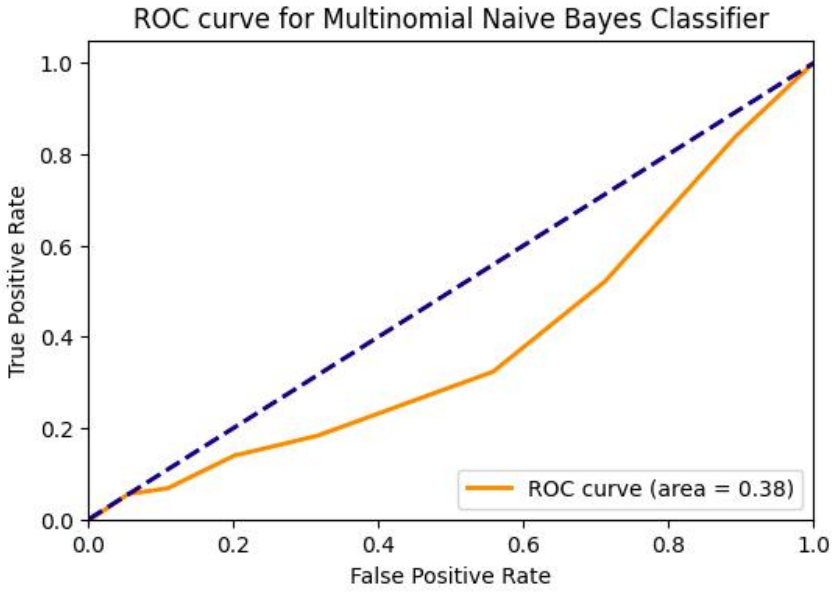
The ROC Curve and Area under the graph for the **Multinomial Naive Bayes** classifier on the **top 10 feature** dataset is as follow:



The ROC Curve and Area under the graph for the **Multinomial Naive Bayes** classifier on the **top 20 feature** dataset is as follow:

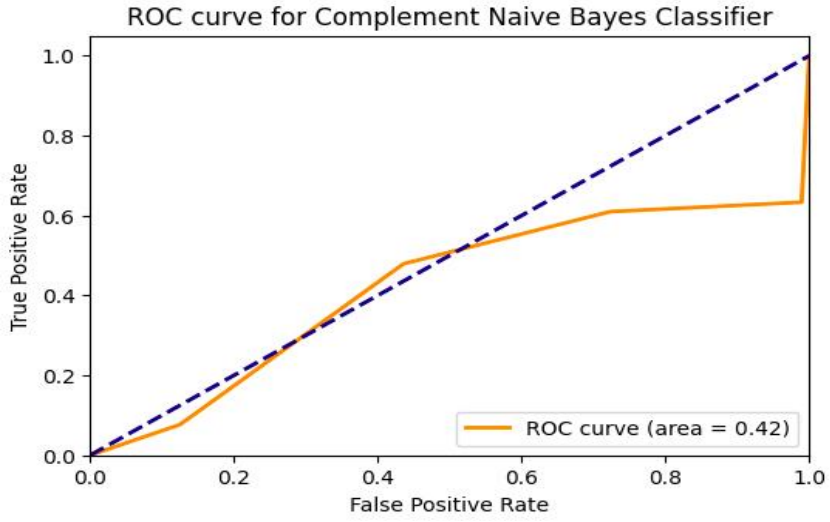


The ROC Curve and Area under the graph for the **Multinomial Naive Bayes** classifier on the **complete** dataset is as follow:

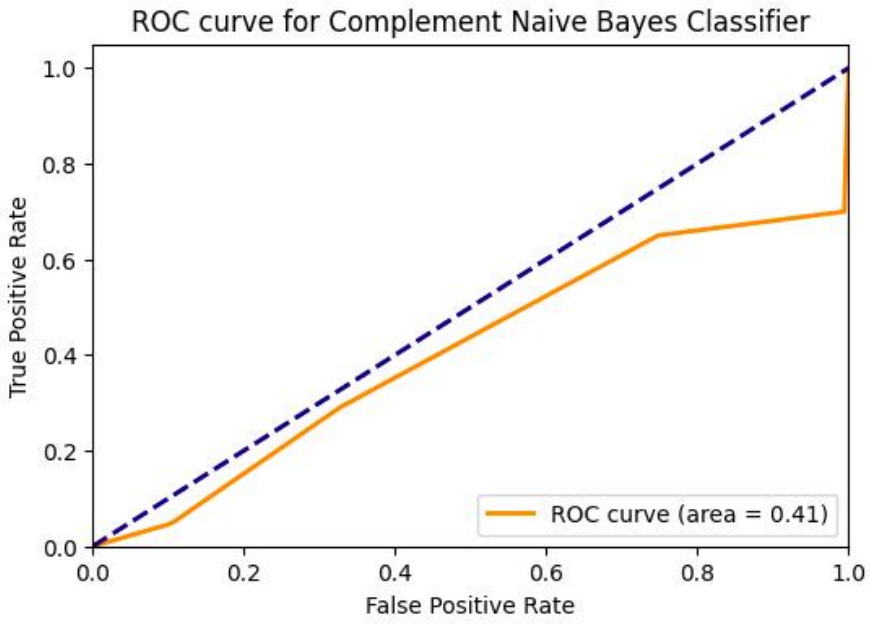


In conclusion, the analysis of the ROC curve and the Area under the Curve (AUC) reveals varying performance levels for the Multinomial Naive Bayes classifier across different feature datasets, with the highest AUC obtained when using the complete dataset.

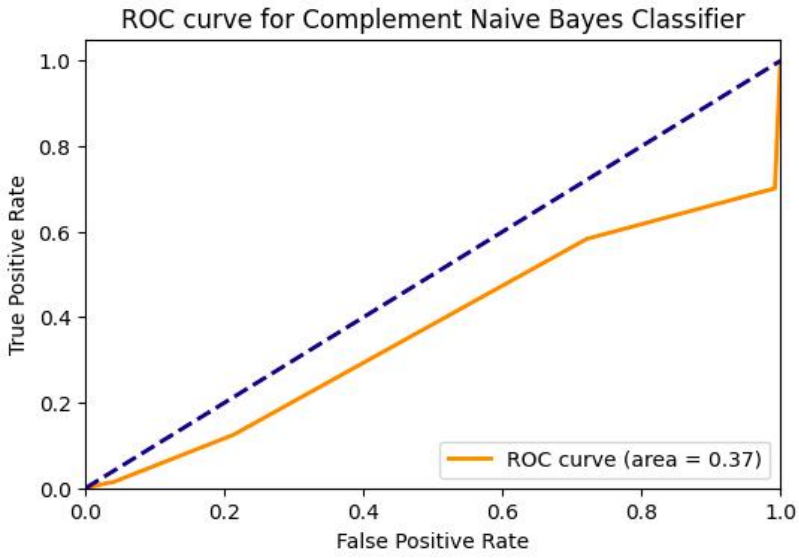
The ROC Curve and Area under the graph for the **Complement Naive Bayes** classifier on the **top 5 feature** dataset is as follow:



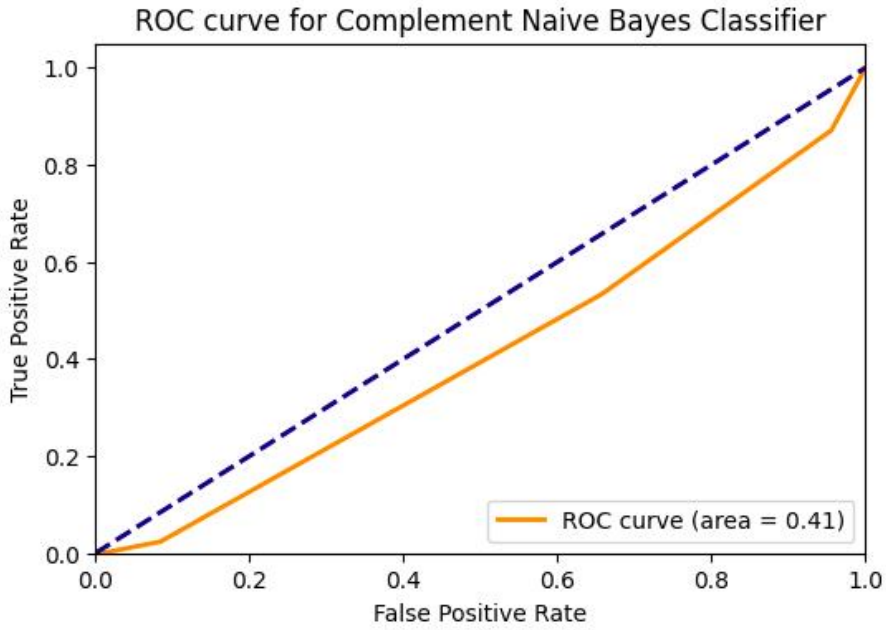
The ROC Curve and Area under the graph for the **Complement Naive Bayes** classifier on the **top 10 feature** dataset is as follow:



The ROC Curve and Area under the graph for the **Complement Naive Bayes** classifier on the **top 20 feature** dataset is as follow:



The ROC Curve and Area under the graph for the **Complement Naive Bayes** classifier on the **complete** dataset is as follow:



The analysis for ROC curve and the Area under the Curve (AUC) for Complement Naive Bayes classifier across different datasets reveals varying performance levels, with the highest AUC achieved with the top 5 feature dataset, making it the most effective configuration among the ones tested.