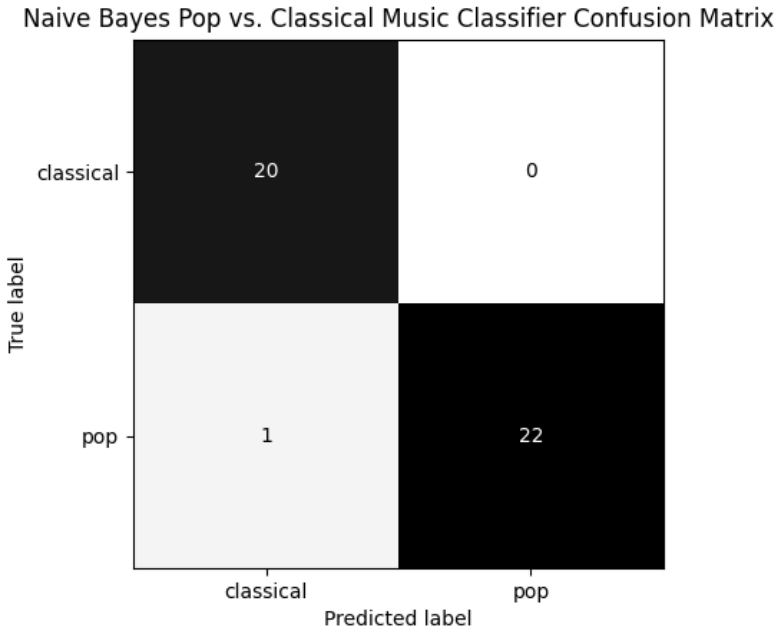
COMP30027 Machine Learning Project 1

Jaicquinn Strafonda, 1268470

Task 1. Pop vs. classical music classification [8 marks]

1. *Compute and report the accuracy, precision, and recall of your model (treat "classical" as the "positive" class). [3 marks]*



Naive Bayes Pop vs. Classical Music Classifier performed with an accuracy of 0.977, precision of 0.952 and recall of 1.0, noting that the classical music genre was treated as the 'positive' class.

2. *For each of the features X (spectral_centroid_mean, harmony_mean, tempo), plot the probability density functions P(X|Class = pop) and P(X|Class = classical).*
If you had to classify pop vs. classical music using just one of these three features, which feature would you use and why? Refer to your plots to support your answer. [5 marks]

To classify music using some feature X, an instance is classified as the class with the largest probability with respect to the instance's feature X value. However, there is a possibility that the instance belongs to another class.
This error rate is the probability of incorrectly labelling an instance. Error rate of the classifier is computed by the area between the x-axis and the minimum of P(X|Class = pop) & P(X|Class = classical). Note the error rate for each of the figures below is clearly marked by the red region.

Figure 1: Distributions of classes for feature spectral_centroid_mean

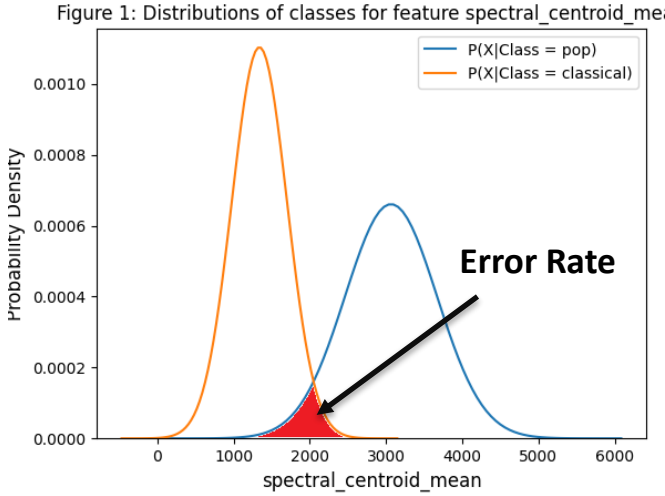


Figure 2: Distributions of classes for feature harmony_mean

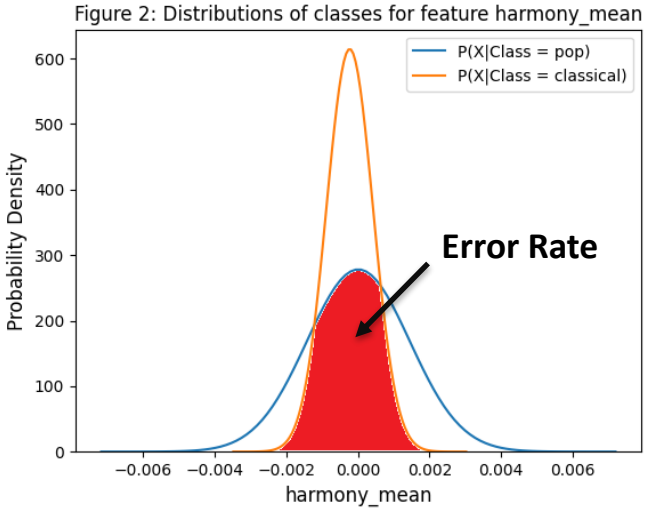
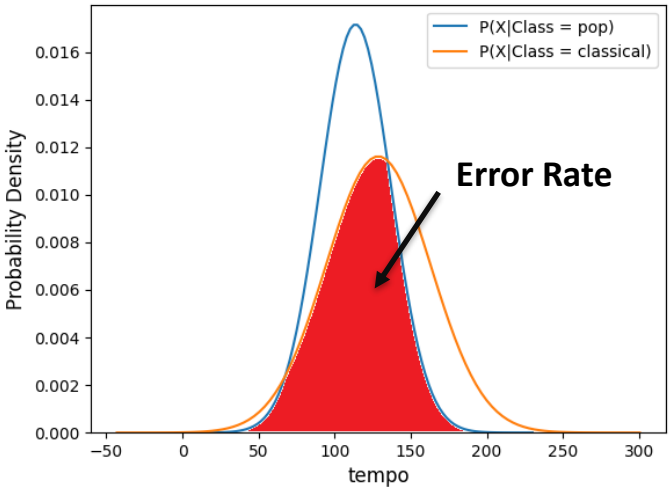


Figure 3: Distributions of classes for feature tempo



As the test dataset is balanced, it follows that better performing classifiers have lower error rates. From visually analysing the area of the red region associated with Figure 1 against that of Figure 2 or Figure 3, the probability of error is minimised when classifying by spectral centroid mean. Hence would classify by the spectral centroid mean feature as such a classifier should be more accurate than a classifier based on either harmony mean or tempo.

Task 2: Naïve Bayes 10-way Classifier

Note my project is an individual submission, so the one question I have selected is Q6 – missing data.

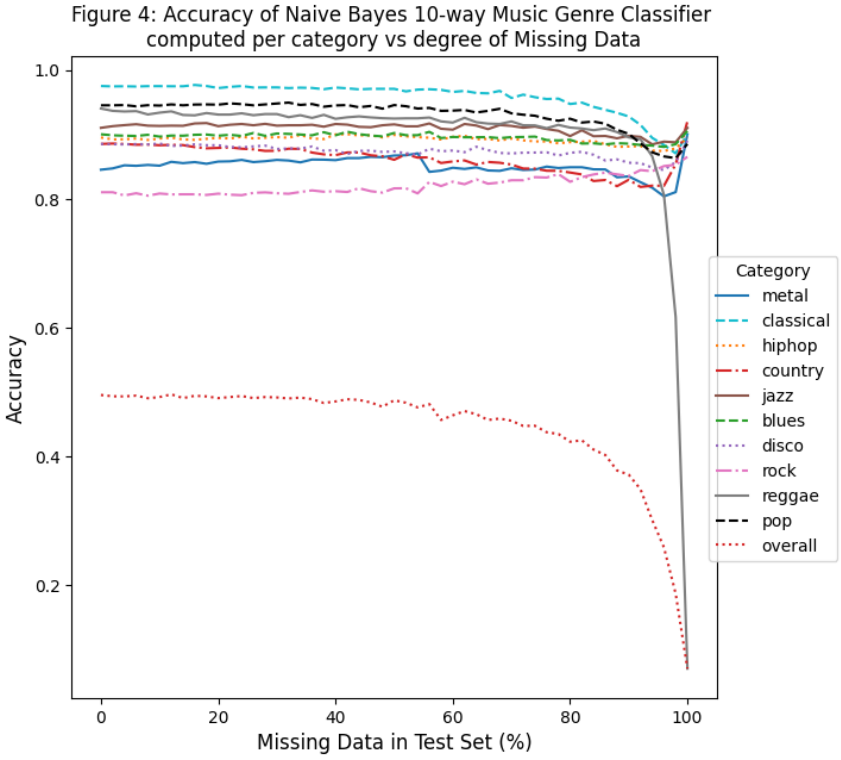
Q6. *Modify your naïve Bayes model to handle missing attributes in the test data. Randomly delete some attributes from the provided test set to test how robust your model is to missing data. In your write-up, evaluate how your model's performance changes as the amount of missing data increases. [4 marks]*

From Table 1, the classifier is more likely to assign the label reggae for higher percentages of missing data in the test dataset because reggae has the largest prior probability in the training dataset.

Table 1: Prior Probabilities in the Training Set and Testing Dataset

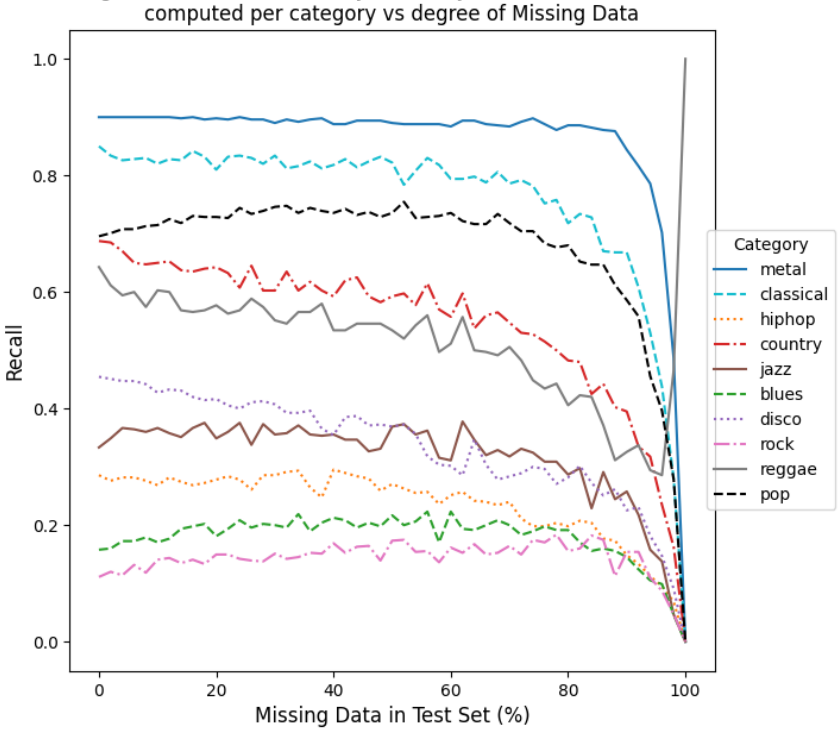
Prior Probabilities	**blues**	**classical**	**country**	**disco**	**hiphop**	**jazz**	**metal**	**pop**	**reggae**	**rock**
Training Dataset	0.10125	0.1	0.105	0.0975	0.09875	0.1025	0.1	0.09625	0.1075	0.09125
Testing Dataset	0.095	0.1	0.08	0.11	0.105	0.09	0.1	0.115	0.07	0.135

Further note from Table 1 that in the testing dataset, 7% of the instances have the true label reggae. Hence in Figure 4, when all the test dataset is missing, the overall accuracy and accuracy with respect to reggae is 0.07. Meanwhile the accuracy computed per the other class labels is high.



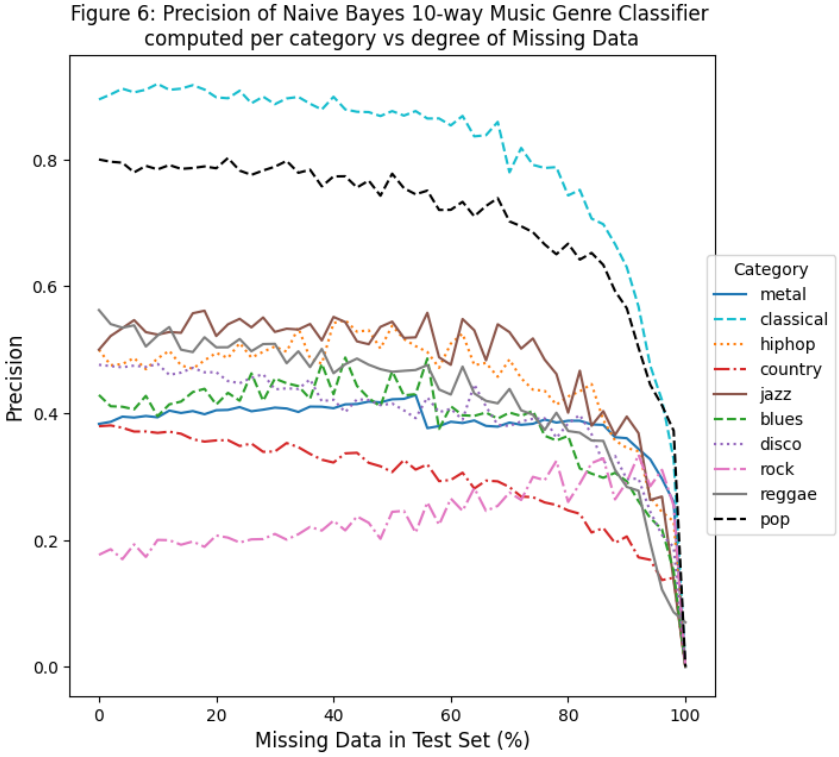
The recall does not significantly deviate between the missing data percentage being 0% to around 40%. Most classes do portray some low degree of loss in recall within this range, but remarkably the recall for blues, rock and metal happen to slightly improve.
When all the testing dataset is missing, recall is 1 for reggae and 0 for all the other classes, due to all instances being classified as reggae.

Figure 5: Recall of Naive Bayes 10-way Music Genre Classifier computed per category vs degree of Missing Data



From Figure 6, the precision of the classifier does not significantly change until more than 50% of the testing dataset is missing. For higher percentages of missing data, the precision quickly approaches 0 for all classes but reggae, reflecting that more and more instances are incorrectly classified as reggae.

Figure 6: Precision of Naive Bayes 10-way Music Genre Classifier computed per category vs degree of Missing Data



From the accuracy, recall and precision analysis, the 10-way Naïve Bayes Classifier does not significantly change in performance until we begin to have around 40% of the testing dataset missing. However, from Figure 4, the overall accuracy when none of the dataset is missing is only 0.495. Consequently, such a classifier is quite unlikely to be an acceptable, practical solution in the real world.