COMP30027 Assignment 1 Written Report

Part 1

Question 1

Several performance metrics can be calculated by counting true positives, true negatives, false positives, and false negatives and analyzing from the confusion matrix (Figure 1). By classifying "classical" as the true label, I will report my model's accuracy, precision, recall, and f1 score. My model has an accuracy of over 97.6%, a precision of over 95%, a recall of 100%, and an f1 score of around 0.976. Such results showed that my model has a solid and robust performance in classifying classical and pop music. With over 40 test cases, only one instance was classified to the wrong label – a piece of pop music being classified as classical.

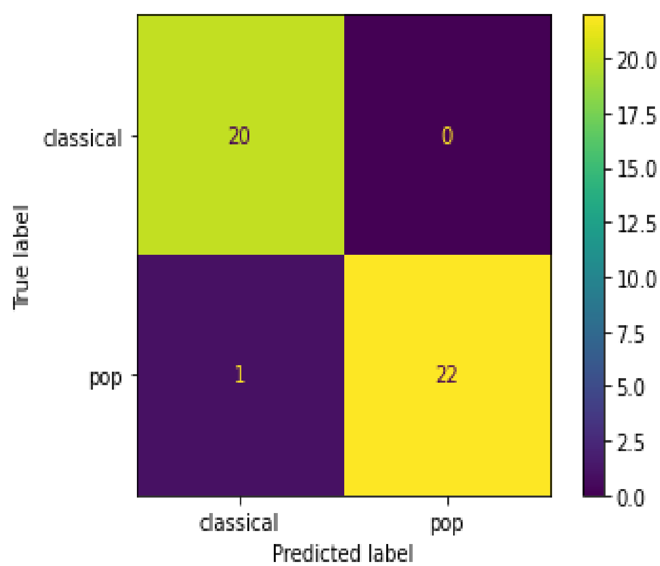

Figure 1

With high accuracy, my model ensured that over 97% of the predicted labels were correct. The high precision and perfect recall also ensure that 95% of the predicted classical music is accurate and can detect all classical music in a given dataset.

Therefore, based on the performance metrics of my model, it is confident to make an evidential conclusion: my model has an overall high performance on classifying pop and classical music with no bias in results. However, the model should be used cautiously if a user wants to ensure all detected classical music are correct, as the model might generate false positives.

Question 2

Spectral centroid mean (SCM) will be the best choice.

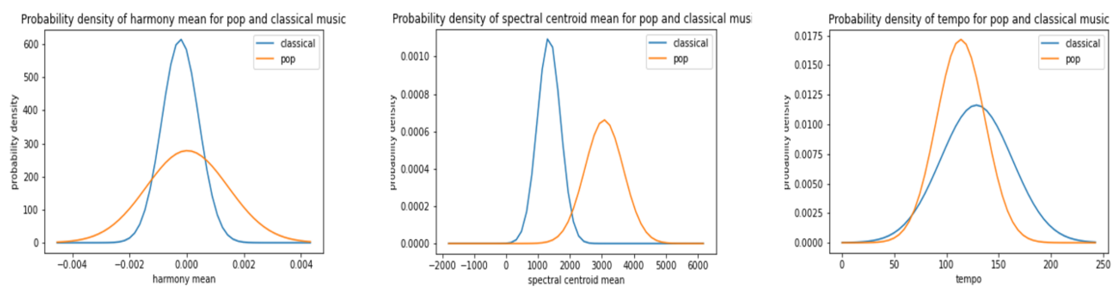

Figure 2

The above graphs (Figure 2) show that SCM has a minimal probability density function (pdf) overlap between the two classes. In contrast, the other two features have high overlap in pdfs.

The overlap of the pdfs can create potential errors: although the probability of a class is higher, which the model will choose, the likelihood of the other class is still present. For example, if we select tempo, since the probability of an instance being pop music is higher between tempo=60 and tempo=140, the model will classify the instance as pop. However, although the probability of pop is high, the chance of the instance being classical music is also not low. In this case, if we only classify instances using tempo, many classical instances will likely be classified as pop, generating false negatives. On the contrary, if we choose SCM, there are only overlaps between SCM=1500 and SCM=2000, which only consists of a tiny proportion of the range of the feature.

Therefore, choosing SCM will yield a better model performance as classes are more disguisable with this feature and will result in fewer potential errors.

Part 2

Question 6

The model is tested from 0% missing value (original dataset) to around 90% random missing data, with three iterations. The results are as follows:

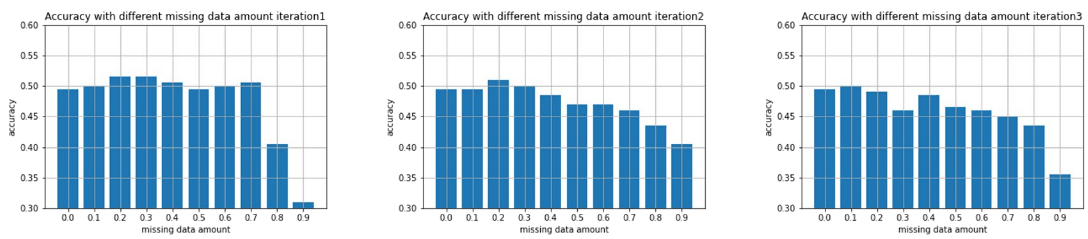

Figure 3

Figure 4

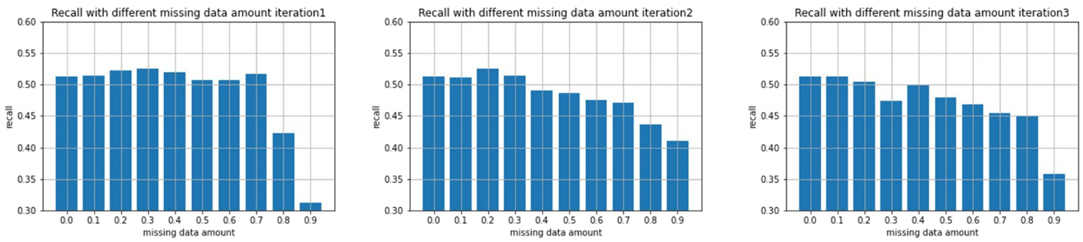

Figure 5

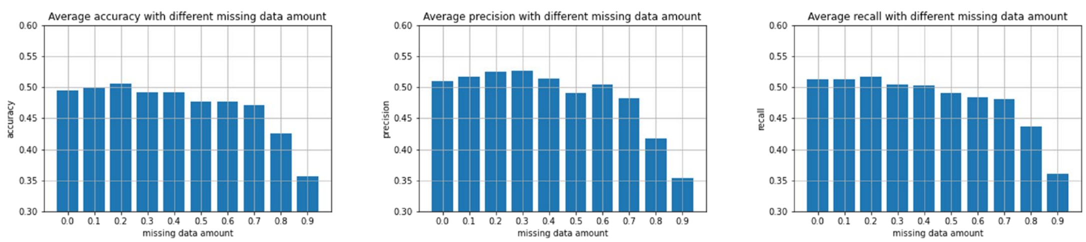

Figure 6

The graphs above show that my model's performance in all metrics varied between the amount of missing data in all iterations (Figure 3-5). Though variations appear across iterations, the model's performance showed the same trend as in averages for all metrics – a slight increase with low levels of missing values but decreases when the number of missing values increases from then on, especially when the number of missing values is exceptionally high (80% and 90%).

The performance increase with a low missing data percentage could result from removing dependent attributes and removing attributes with high error rates. Since my model is a gaussian Naïve Bayes (GNB) model, the model assumes data are in gaussian distribution and are independent. However, such assumptions might not hold in our dataset as no validation was progressed. By randomly removing some attributes, some dependencies might distinguish, resulting in increased independence between features and fewer overweighed attributes. Another reason could be that attributes with high pdf overlaps between classes were removed, as discussed in Q2.

The dramatic performance drop with high missing value rates could result from lacking information. Removing 80% - 90% of features will cause the model to infer results based on only one to two attributes. However, those features could be misleading, and there are no other features to balance out misleading features, causing poor performance.

Overall, a GNB model is robust against low levels of missing values but will perform poorly with higher levels of missing values.