COMP30027 Machine Learning Asst1 Report

Liangdongfang Xu 1174154

Task 1. Pop vs. classical music classification

1.
The accuracy is 0.9767441860465116
The precision is 0.9523809523809523
The recall is 1.0

2.
The graphs on the right side (Figure 1) are three density plots of spectral centroid mean, harmony mean and tempo. Each plot consists a blue curve and a red curve, which corresponds to pop and classical respectively.

I may choose spectral centroid mean to be the attribute X. The reason to choose one of the three attributes is that we want one attribute to best distinguish two labels (pop and classical). We don't want some values of the attribute results into a difficulty for separating two labels, that is to find the graph which has the minimum area of overlaps between the blue and red curves. By using visual approximation, we will certainly exclude the temp as it apparently has a big overlapping, and the overlapping area of the spectral centroid mean is similar to a triangle (1000 * 0.0002 / 2 = 0.1) while the area of the harmony mean is roughly a rectangle (0.00025 * 1000 = 0.25). We would like to choose the one with the minimum area, that is spectral centroid mean.

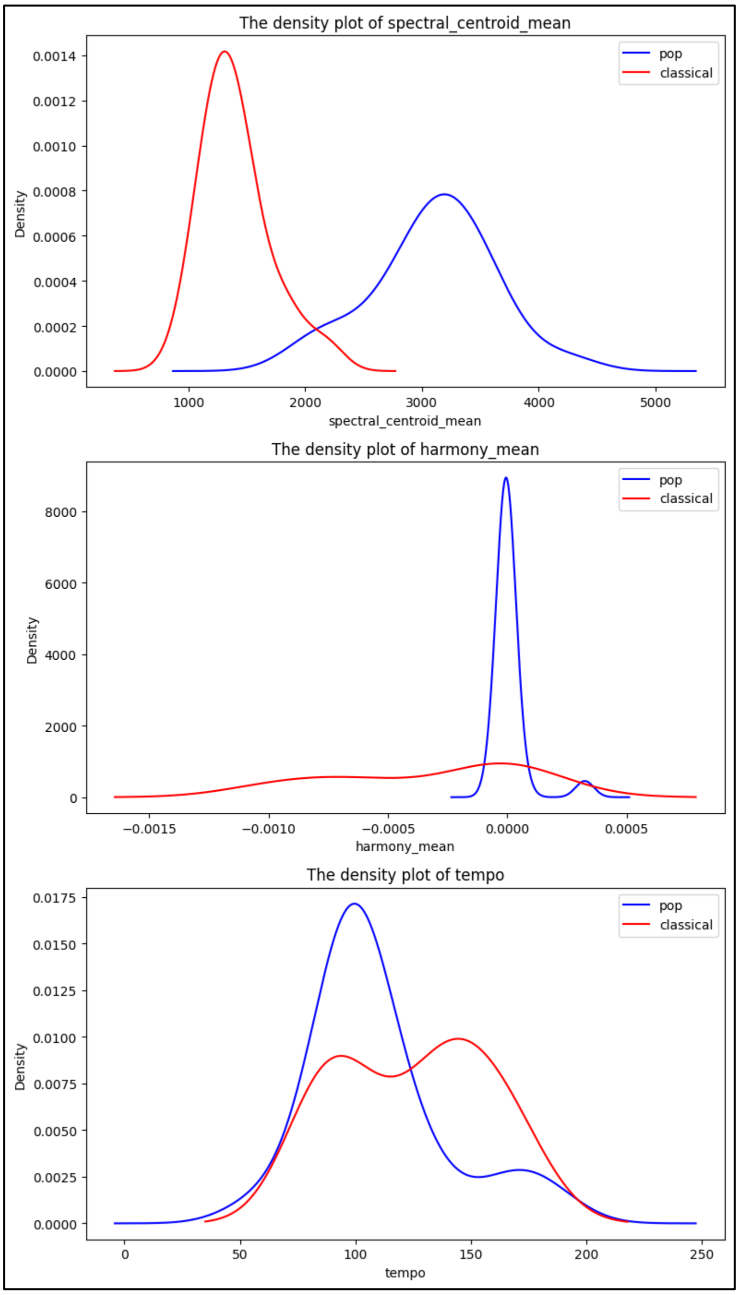
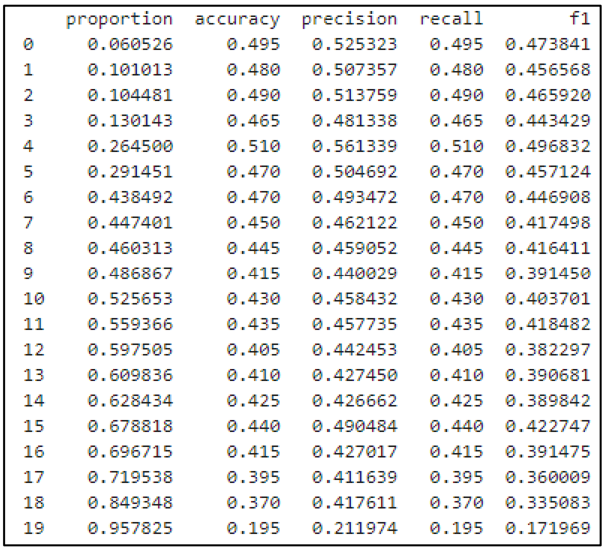


Figure 1

Task 2. 10-way music genre classification

Q6 Dealing with missing values.
My function will randomly generate 20 proportional values from 0 to 1. A function called delete_value_by_proportion will be called to delete some values of attributes of the training data frame based on the proportion given. Then this modified data frame will passed into the predict_missing_value function, which simply substitute $\frac{1}{number\ of\ attributes + 1}$ into the missing values. In this case, the number of attributes is 57, excluding "filename" and "label". Then the evaluate function will compare the training result and the actual result and output the accuracy, precision, recall and f1 score. Results are shown on (Figure 2.1). Notably, since proportions are randomly generated, every time the running result will be differently, and (figure 2.1) is a one-time result.



```
    proportion  accuracy  precision  recall        f1
0     0.060526     0.495   0.525323   0.495  0.473841
1     0.101013     0.480   0.507357   0.480  0.456568
2     0.104481     0.490   0.513759   0.490  0.465920
3     0.130143     0.465   0.481338   0.465  0.443429
4     0.264500     0.510   0.561339   0.510  0.496832
5     0.291451     0.470   0.504692   0.470  0.457124
6     0.438492     0.470   0.493472   0.470  0.446908
7     0.447401     0.450   0.462122   0.450  0.417498
8     0.460313     0.445   0.459052   0.445  0.416411
9     0.486867     0.415   0.440029   0.415  0.391450
10    0.525653     0.430   0.458432   0.430  0.403701
11    0.559366     0.435   0.457735   0.435  0.418482
12    0.597505     0.405   0.442453   0.405  0.382297
13    0.609836     0.410   0.427450   0.410  0.390681
14    0.628434     0.425   0.426662   0.425  0.389842
15    0.678818     0.440   0.490484   0.440  0.422747
16    0.696715     0.415   0.427017   0.415  0.391475
17    0.719538     0.395   0.411639   0.395  0.360009
18    0.849348     0.370   0.417611   0.370  0.335083
19    0.957825     0.195   0.211974   0.195  0.171969
```

Figure 2.1

(Figure 2.2) is a scatter plot with line of best fit based on the result of (Figure 2.1). It is clear that when the proportion of missing value going up, the accuracy/precision/recall/f1 is going down, that is the model's performance of dealing with missing value is decreasing. Observing that when the proportion is close to 1, the evaluation factors are about 0.2, which is not good, but better than randomly guessing. Another interesting point is that only three lines are on the graph (Figure 2.2). This is because that accuracy and recall are same for this prediction.

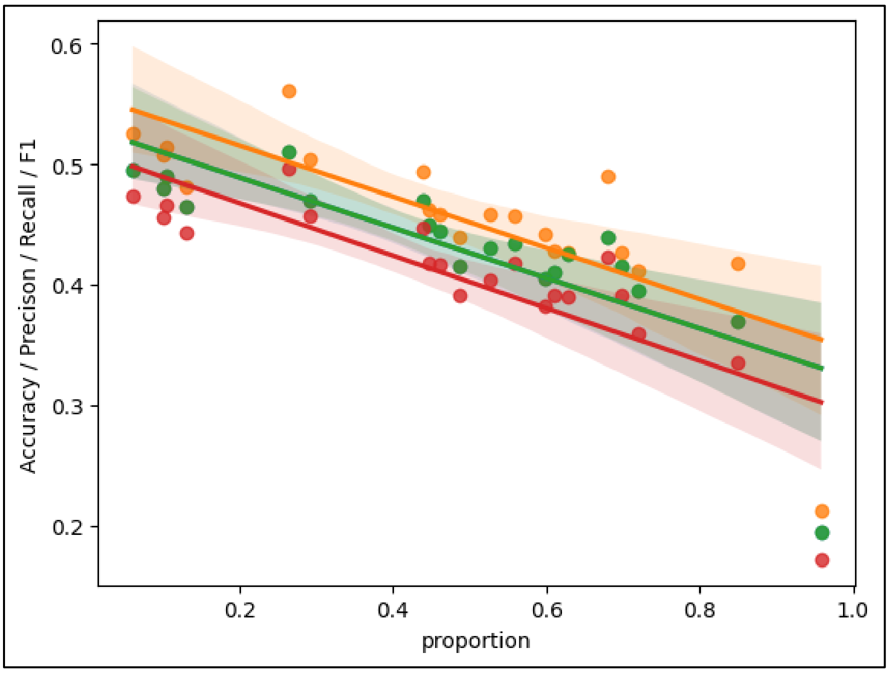


Figure 2.2