Sakib Ahmed

CSCI-UA 473: Intro to Machine Learning

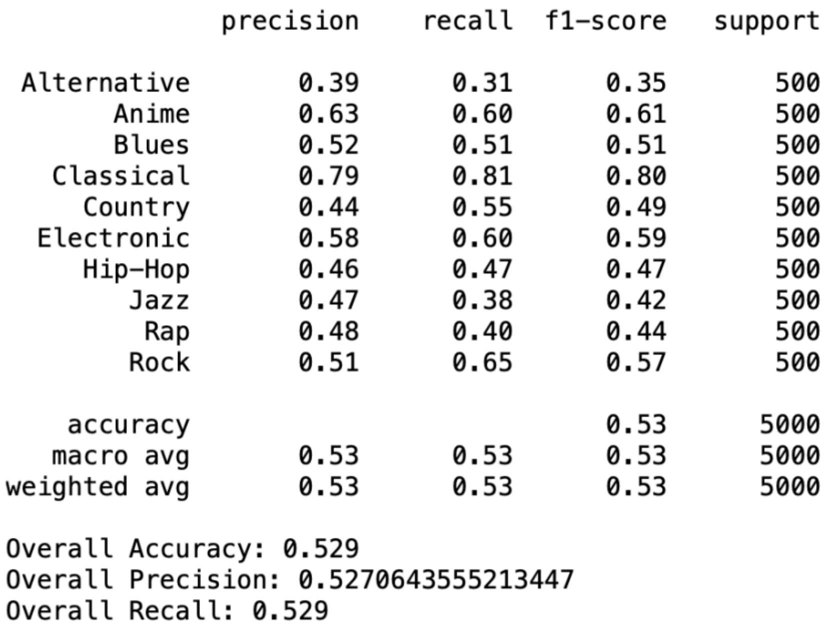
Classification Capstone Project

I built a logistic regression model with factor analysis of mixed data (FAMD) dimensionality reduction and clustering. First, I loaded the data from “musicData.csv” into two different dataframes: data and dataNumerical. Next, I dropped the features “instance_id”, “artist_name”, “track_name”, “obtained_date” from both dataframes. After doing this, I changed dataNumerical[“music_genres”] from a categorical to a numerical format, ordered from 0 to 9. The next step was to one-hot encode the “mode” and “key” features in both dataframes, as well as drop any rows with missing data. Next, I created multiple smaller dataframes for each musical genre and split them into a testing and training set, such that the test set for each genre would contain 500 entries; I followed this by concatenating these test and training sets into a larger test and training set, respectively. I standardized these dataframes and ran a logistic regression with the parameters “multi-class=‘auto’” and “solver=‘lbfgs’”. The logistic regression model on the dataframe with categorical music genres was used to calculate the by-genre, macro average, and weighted average for precision, recall, f1-score. The dataframe with numerical musical genres was used to create a confusion matrix, heatmap and plot the ROC curves for each music genre; also, the ROC curve was used to calculate the AUC score with a OneVsRestClassifier. I then performed FAMD on the categorical training dataframe and visualized the results plotted on the first two principal components. From this visualization, I identified 7 distinct clusters and added this as an additional feature to the data. With these updated dataframes, I performed a logistic regression again in the same way as earlier, recording the noticeably greater values for the by-genre, macro average, and weighted average of precision, recall, f1-score, the updated ROC curve, confusion matrix, heatmap, and the improved AUC score.

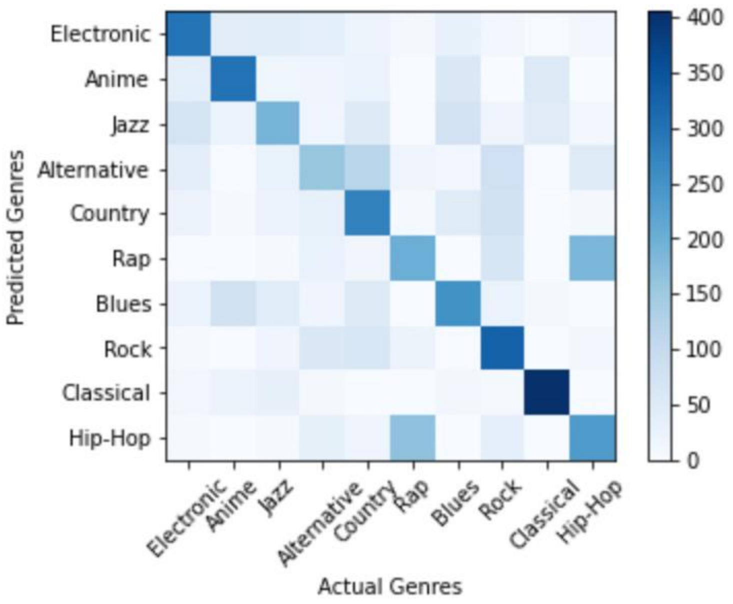
I used a logistic regression model because I recall professor stressing that these are the most used models in actual industry situations, and so I wanted to see for myself how effective they can be with real data. I created two identical dataframes only differing in their "music_genre" format because FAMD only works if a dataframe contains some categorical data, and, as far as I know, ROC curves could only be plotted with a dataframe that contains only numerical data. I made a heatmap to visualize the confusion matrix, because the standard confusion matrix can be difficult to interpret when working with multi-class classification. I chose to do FAMD dimensionality reduction and clustering because it seems to be the only clustering method that properly weighs categorical features with respect to continuous features. Additionally, PCA is a poor algorithm for mixed data and my attempts at performing kMeans with tSNE or MDS would cause a memory overflow and the kernel to die. The FAMD produced great results and made out seven distinct clusters. With this information, I followed the methodology of the article linked by Avinav in the discussion board regarding how to use information from clustering to improve a classification model and added these cluster assignments as an additional feature in my data. After doing this, I found that my results had improved a considerable amount.

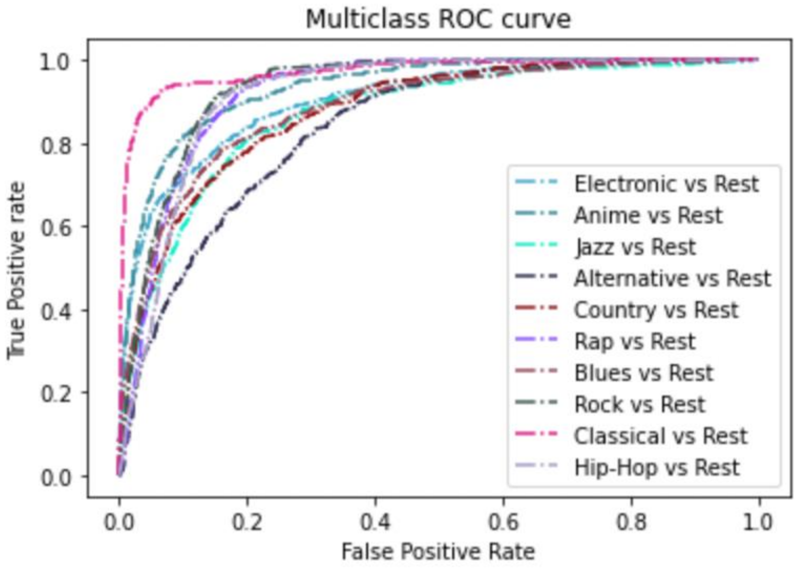
I addressed the missing data, which were about 5000 entries with a '-1.0' value for "duration_ms" and about 5000 entries with a '?' for "tempo" by dropping them from the dataframe. I addressed the acoustic features not following a normal distribution by instead z-score normalizing. I addressed the categorical format of "keys" and "mode" by one-hot encoding them into dummy variables. I addressed the "music_genre" categorical format by turning it into a numerical format. I addressed the avoidance of normalization of categorical features by z-score normalizing only continuous features. I did not consider linguistic properties in my classification.

Performing a logistic regression model without dimensionality reduction and clustering resulted in an AUC score of 0.905 and the following information:

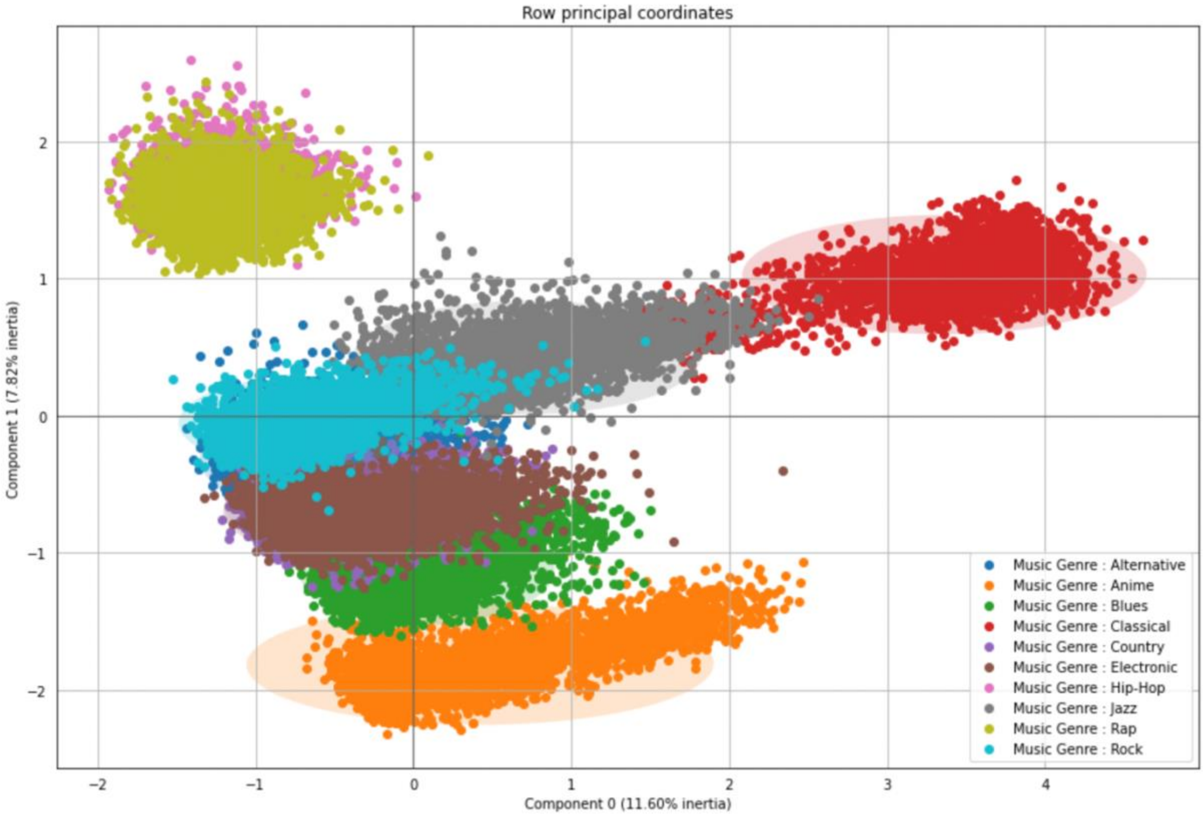


```
              precision    recall  f1-score   support

 Alternative       0.39      0.31      0.35       500
       Anime       0.63      0.60      0.61       500
       Blues       0.52      0.51      0.51       500
   Classical       0.79      0.81      0.80       500
     Country       0.44      0.55      0.49       500
  Electronic       0.58      0.60      0.59       500
     Hip-Hop       0.46      0.47      0.47       500
        Jazz       0.47      0.38      0.42       500
         Rap       0.48      0.40      0.44       500
        Rock       0.51      0.65      0.57       500

    accuracy                           0.53      5000
   macro avg       0.53      0.53      0.53      5000
weighted avg       0.53      0.53      0.53      5000

Overall Accuracy: 0.529
Overall Precision: 0.5270643555213447
Overall Recall: 0.529
```

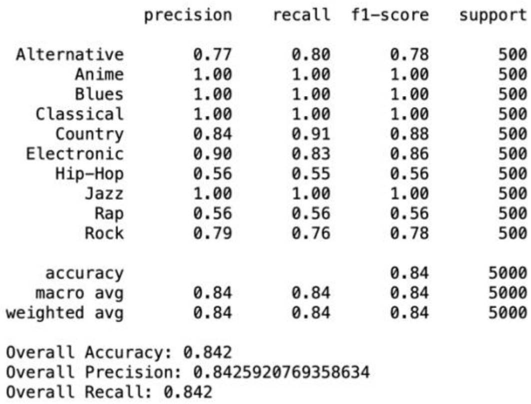




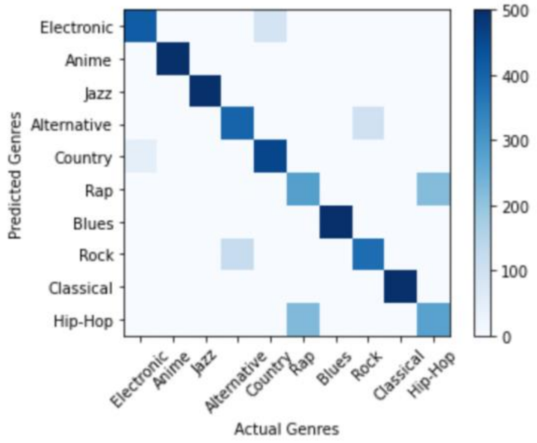
After these initial results, I performed dimensionality reduction and clustering via FAMD with the following visualization:

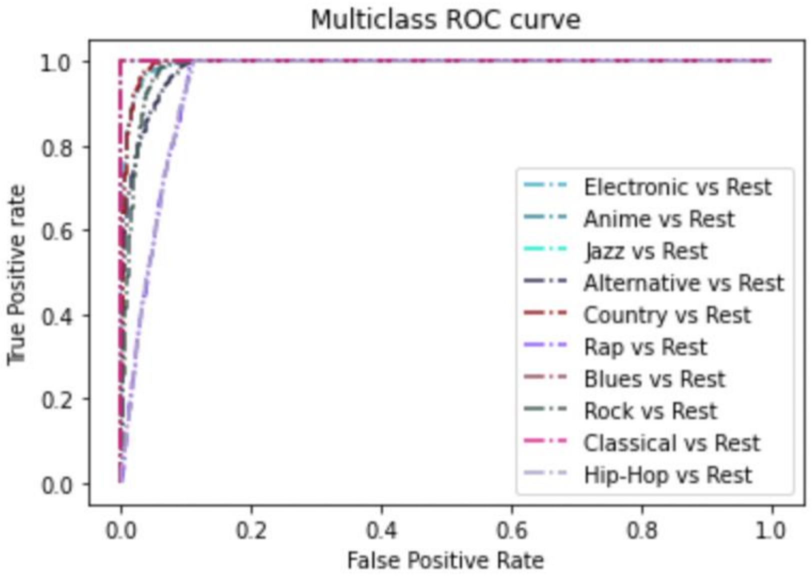


Finally, performing a logistic regression model with dimensionality reduction and clustering via FAMD resulted in an **AUC score of 0.987** and the following information:



```
              precision    recall  f1-score   support

 Alternative       0.77      0.80      0.78       500
       Anime       1.00      1.00      1.00       500
       Blues       1.00      1.00      1.00       500
   Classical       1.00      1.00      1.00       500
     Country       0.84      0.91      0.88       500
  Electronic       0.90      0.83      0.86       500
     Hip-Hop       0.56      0.55      0.56       500
        Jazz       1.00      1.00      1.00       500
         Rap       0.56      0.56      0.56       500
        Rock       0.79      0.76      0.78       500

    accuracy                           0.84      5000
   macro avg       0.84      0.84      0.84      5000
weighted avg       0.84      0.84      0.84      5000

Overall Accuracy: 0.842
Overall Precision: 0.8425920769358634
Overall Recall: 0.842
```





Extra Credit

An interesting observation found in the FAMD is the grouping of certain music genres. First, I was not surprised that hip-hop and rap were so similar but was shocked to see how far they were from every other genre on the FAMD plot. These are personally my favorite music genres and I know that something unique to these two genres is their frequent use of 808s, which is an electronic drum-like sound with a low-end sub bass profile that is rarely found in other genres. Second, initially seeing country and electronic music together didn’t make sense to me at all, but after thinking about some of the biggest edm hits in the last decade, I realized that they all feature an acoustic guitar; some of these songs are Avicii’s “Wake Me Up”, Gazzo’s “Sun Turns Cold”, Kygo’s “It Ain’t Me”. I can’t think of any other genre that uses the acoustic guitar, besides some alternative music, and looking at the FAMD we see that alternative is the only other genre that has some slight overlap with country and electronic. Lastly, seeing rock and alternative together makes a lot of sense to me, as growing up I’ve seen many musical groups blend the line between these two genres. In particular, I listened to a lot of Coldplay growing up and I can recall them being labeled as alternative by some and rock by others throughout the years. A more recent example is Imagine Dragons, another band that can be seen as both rock and alternative. Interestingly enough, both these bands biggest instruments, besides their voice and maybe percussion drums, is most often an electric guitar.