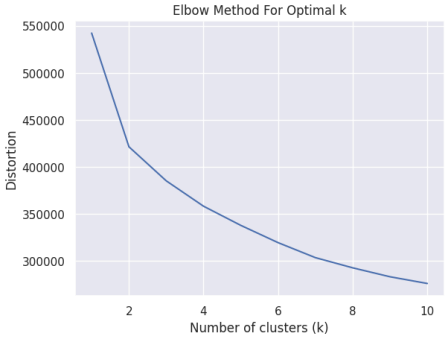
Mimi Chen
FML Capstone Project (Classification)

Data Processing

To begin, I dropped the instance ID and obtained date because they are irrelevant to the music genre. In addition, I also dropped the track name and artist name because they are not relevant audio features. I also performed one-hot encoding on the 'key' column and mapped the mode column to 1 and 0 (since it is a binary feature). To ensure there is no reference column and the features are linerarly independent, I dropped the key_A column after one-hot encoding. Then, I replaced instances where tempo = "?" with NaN. Since there were almost 5000 instances in the data where the tempo was missing, I decided to impute rather than drop the data. I split my data into test and train, with 500 randomly selected songs from each genre. I also standardized the numerical columns using StandardScaler.

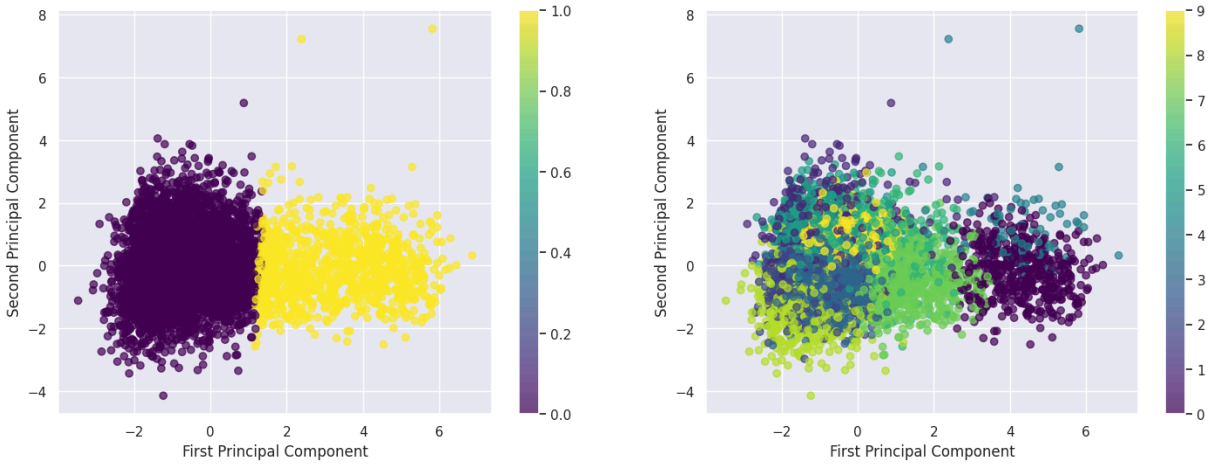
Dimensionality Reduction with PCA

For dimensionality reduction, I applied Principal Component Analysis on the data. Following the Kaiser Criterion, there are only three Principal Components with an eigenvalue above 1, but they only explain 50.45% of the variance. On the other hand, retaining the first 6 Principal Components explains more than 70% of the variance and retaining the first 10 Principal Components explains more than 90% of the variance.



K-Means Clustering

I used the elbow method to evaluate the optimal number of clusters for K-Means clustering. Looking at the graph it seems like 2 is the optimal number of clusters, but given that there are 10 target classes, I also chose to evaluate the clustering results for 10 clusters.

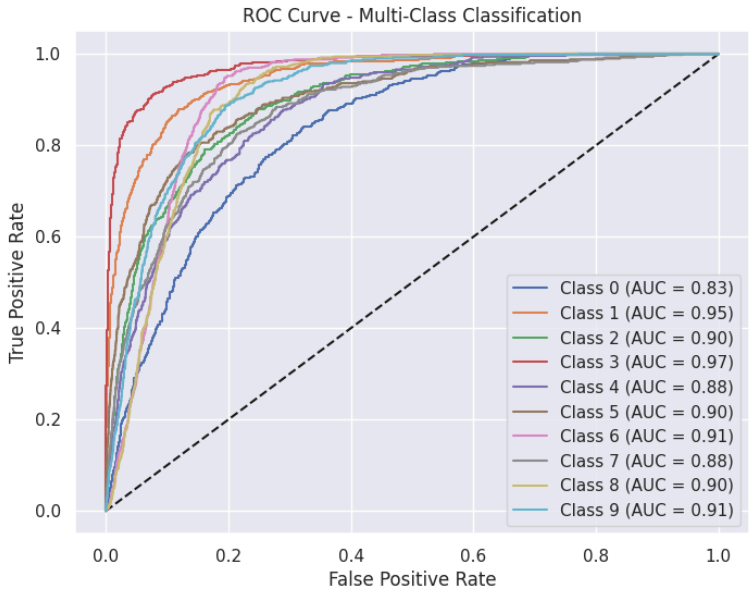


K-Means Clustering with 2 vs10 clusters

There is a pretty distinct difference when we cluster with only 2 clusters, while the clusters overlap quite a bit when we use 10 clusters.

XGBoost Classification

For classification I used XGBoost (Extreme Gradient Boosting), converting true class labels to a binary form and computing the ROC-AUC scores between the binary true labels and predicted probabilities. Then I calculated the ROC-AUC score across all classes and calculated the macro-average ROC-AUC score to access the model's overall performance, yielding a macro-average of **0.90242093.** I used 8 Principal Components because they explain more than 80% of the variance.



ROC-AUC Curve: Multi-Class Classification

Extra Credit

I used the built in feature importance attribute to determine which features were the most important in the classification model. Popularity seems to be the most important featrure when it comes to classification, and out of the first 8 components, energy was the least important. In addition, I used a correlation matrix to evaluate the correlation coefficients between the variables.