REPORT Capstone Project FML

Full Name	Net ID	Class	Professor
Aya El Mir	ae2195	FML	Pascal Wallisch

Directives on the report:
Write a brief report (1-2 pages) as to how you built your model, how you made your design choices (why you did what you did) and addressing how you handled the challenges below.

- ***Make sure to state your final AUC at the bottom of the report and please include a plot of the ROC curve in your report.***
- ***As dimensionality reduction will be critical, please also include a visualization of the genres as clusters in the lower dimensional space (lower than the dimensionality of the original data) you uncovered and include a comment as to what you think about this space/clustering.***
- ***Also make sure to comment on what you think is the most important factor that underlies your classification success.***

Approach:

Data Preparation:

- Checking for outliers

Checking for outliers is important because outliers can have a significant impact on the performance and accuracy of your models. By detecting and addressing outliers, you can ensure that your models are not unduly influenced by extreme values, leading to more reliable and accurate predictions of the music genre.

- Handling missing data

There is randomly missing data, e.g. some of the durations of some of the songs are missing, as well as some of the auditory feature values. There are not many missing values, but we have to handle them somehow, either by imputation or by removing the missing data in some reasonable way.
My approach is to first detect the NaNs in the dataset, they are only 0.17% of the whole dataset, thus, they can be dropped. Indeed, when we count the number of '?' values in the dataset, we find a larger percentage (10%) which corresponds to 4980 values in the tempo column, therefore, this time we do imputation by replacing these values with the mean.

- Standardise all noncategorical variables

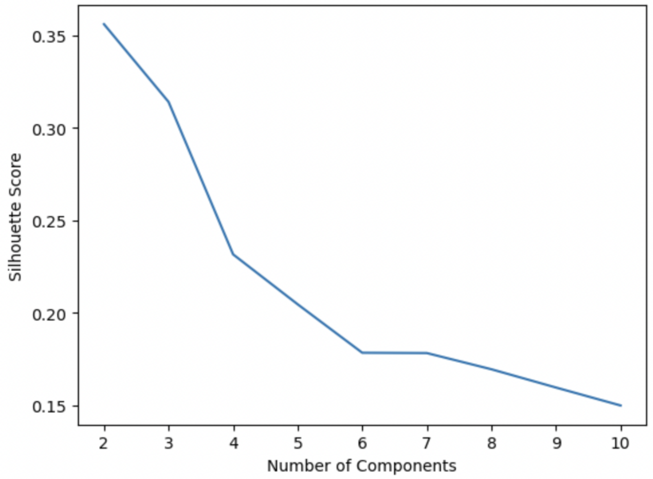
We need according to the prompt to standardize the acoustic features which are unlikely to be normally distributed.
Therefore, the approach I am taking is using StandardScaler() function on all the noncategorical features to ensure they are all following a normal distribution.

- Variable encoding: Transformation of the features in string format and categorical format into numerical data to be useful

To transform text-based features into numerical data, I use a technique called one-hot encoding. One-hot encoding converts categorical variables into binary vectors that can be used in machine learning algorithms. Indeed, it is a technique that preserves the categorical information, handles categorical variables with multiple levels in our case with different music genres, prevents feature dominance, and avoids incorrect assumptions which all improve the model performance.

Exploratory Data Analysis: Visualizing the genres as clusters in a lower-dimensional space

- Dimensionality reduction
  - PCA



- Plotting the Silhouette Plot

In order to determine the optimal number of clusters and aid in comparing different clustering algorithms or parameter settings. Silhouette Plot is a useful tool for understanding the structure and characteristics of the clusters formed by the algorithm and assessing the overall performance of the clustering task.

- Computing Eigenvalues

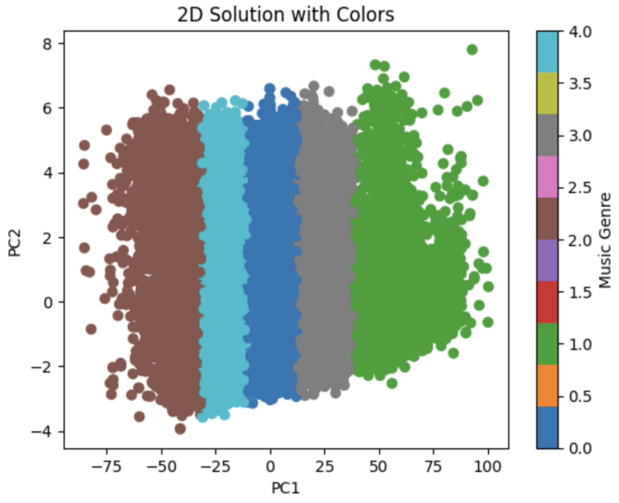
Also to find the optimal numbers of components for PCA. We find that the number of components to keep (number of Eigenvalues above 1) is 5 and indeed we find that they explain 0.9935030119669158 of the variance in the data.

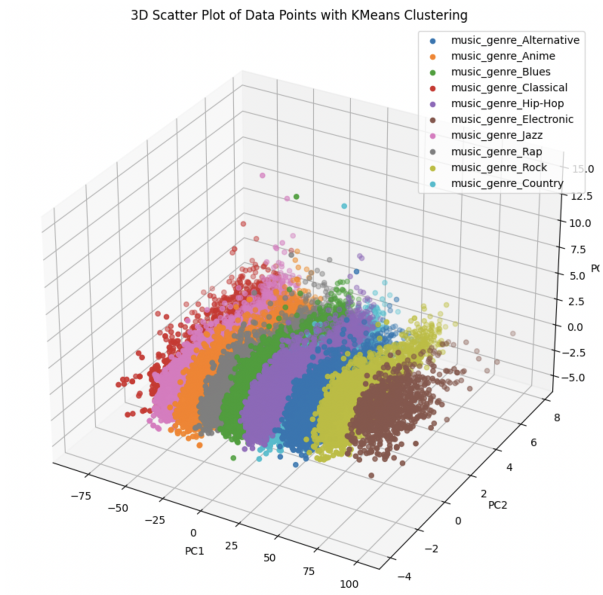
```
Variance explained by first 5 principal components: 0.9935030119669158
```

- ***Visualization of the genres as clusters in the lower dimensional space (lower than the dimensionality of the original data)***

Projecting into 2D (2 components)

Projecting into 3D (3 components)





- ***What do you think about this space/clustering?***

Looking at the 2D and 3D representation we still see almost perfect differentiation of clusters for music genres which makes sense as we already have the labels in the dataset. However, the clustering is not perfect and this is why we need to do more complex methods to be able to get good clustering of each class with respect to other classes and well as most importantly see how we can predict music genre from the given features; leading us to Model Development

***Model Development and Evaluation of the AUC with* hyperparameter tuning:**

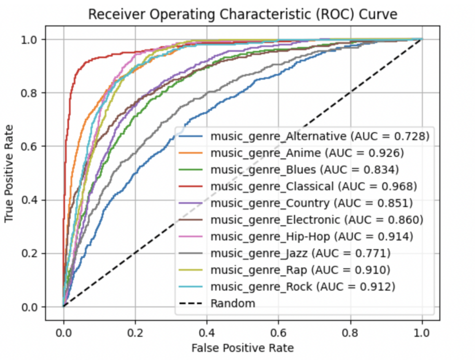
Here we are actually getting into the goal of the assignment which is to get as the highest of AUC as possible which indicates a strong performance of your classification model in distinguishing between different music genres.

- Train Test Split

According to the capstone project requirements we need to make sure to do the following train/test split: For *each* genre, use 500 randomly picked songs for the test set and the other 4500 songs from that genre for the training set. So the complete test set will be 5000x1 randomly picked genres (one per song, 500 from each genre). Use all the other data in the training set and make sure there is no leakage. Indeed, I was able to get X_train, y_train, X_test, and y_test, corresponding to the training and test sets with X having all the features (31) and y having the 10 different music genres the dataset has.

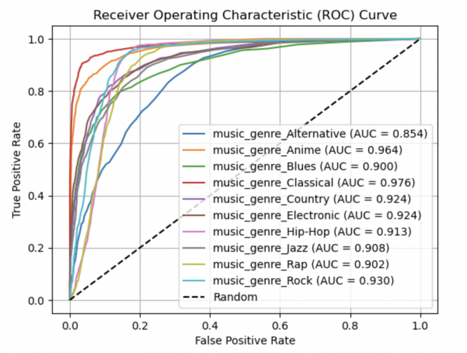
- Model choice:

The approach taken for each model is to loop through each genre and you get AUCs for each genre (corresponding to training for that genre against the others) and then make a huge plot with the ROC curve and the AUC for each genre against the other genres.



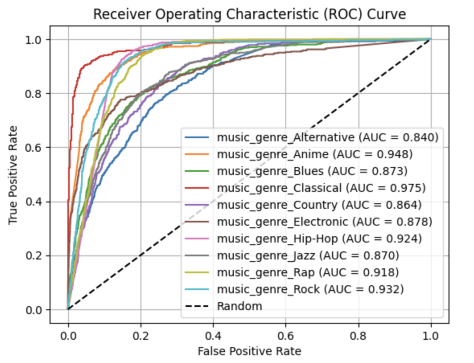
- Logistic Regression

Logistic regression is a linear model that can work well for binary or multi-class classification problems. It's relatively simple, interpretable, and can handle both numerical and categorical features. If you believe that the relationship between the audio features and the genre is approximately linear, logistic regression could be a good choice.



- Random Forest

Decision trees, random forests, and gradient boosting methods (e.g., XGBoost or LightGBM) are popular choices for classification tasks. These models can capture non-linear relationships and interactions between features. They can handle both numerical and categorical features, and they often provide feature importance, which can help you understand which audio features are most important for classification.



- Neural Networks

In order to explore more complex models, we are also considering neural networks. Deep learning models, such as multi-layer perceptron (MLP) or convolutional neural networks (CNN), have been used for music genre classification with promising results. However, they typically require more computational resources and larger datasets to train effectively.

- Model Evaluation:

When the AUC is high, close to 1, it suggests that the classifier has a high true positive rate and a low false positive rate, indicating a strong ability to

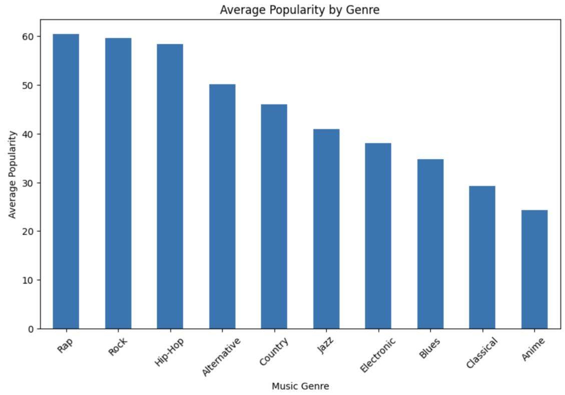
correctly classify positive instances (correctly identify a music genre) while minimizing misclassifications (assigning a music genre when it's not present).
This means that my model is performing well in distinguishing between different music genres.
In summary, a high AUC in your project implies that your SVM model has a high accuracy and reliability in classifying music genres, providing a solid foundation for the successful classification of songs into their respective genres. As can be seen from the plots, **the best AUC was 0.976 for Random Forest.**

Conclusion:

Make sure to comment on what you think is the most important factor that underlies your classification success.
Obtaining a high AUC of 0.976 from the Random Forest model suggests that the model is performing exceptionally well in distinguishing between different genres of music. The main reason behind this is the feature importance in Random Forest, indeed, Random Forest calculates the importance of each feature in the classification task. It ranks the features based on their contribution to the model's accuracy. By considering the most informative features, the model can effectively separate different genres, resulting in a high AUC. Additionally, I made sure to do the right preprocessing techniques to be able to accurately utilize the data set which also plays a huge role in the high performance.

Extra Credit:

One interesting analysis is to explore the relationship between the popularity of songs and their corresponding genres. I, therefore, examine whether certain genres tend to have higher popularity ratings compared to others.



Another aspect we could look at further for extra credit is the genre of Clustering Visualization. Since I have performed clustering on your data, I could visualize the clusters in a visually appealing way. For example, I could further use other dimensionality reduction techniques like t-SNE or UMAP to visualize the clusters and observe if there are any distinct patterns or relationships between genres.