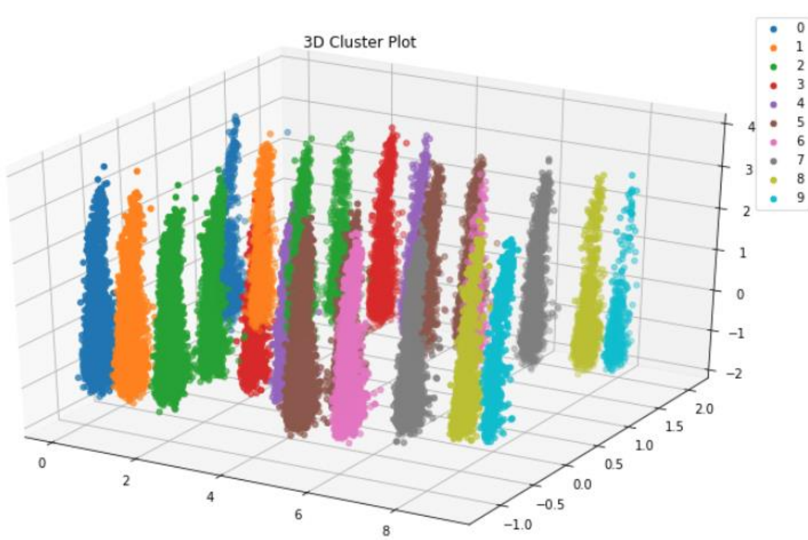
I did a train/test split so that the testing set consists of 500 randomly picked songs for one genre, and 5000 songs in total (since there are 10 genres). The training set has the rest 45000 songs.

For data cleansing, I found that there are 5 NaN values for every column. Therefore, I dropped these NaN so that the dataset becomes exactly 50000 rows. There are also missing values denoted as “?” in column “tempo” of object data type. Take column “tempo”, for example. I checked that the valid values in the “tempo” are all positive. Thus, I decided to replace all missing values with string “0” first and cast the data type from object to float. I then replaced all 0’s in “tempo” with the mean column value (since missing values are replaced as 0, this would give us the exact mean of the column) and replaced any missing values in other numerical columns with the column mean respectively. I also dropped columns “instance_id”, “artist_name”, “track_name”, “obtained_data” for the time being.

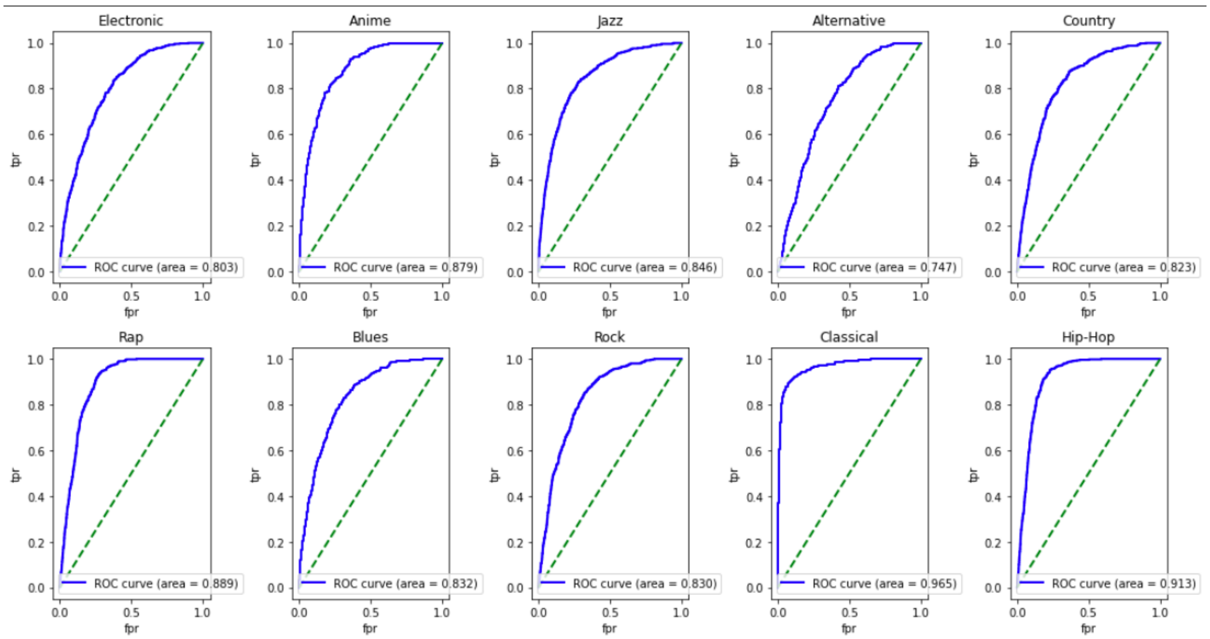
In terms of dimensional reduction, I one-hot encoded the categorical variables like “mode”, “key” using get_dummies function. I used StandardScaler to compute z-scores for features in the training data and mapped those to the testing data for consistency. Since categorical variables have values either 0 or 1, StandardScaler would keep them intact. I applied



PCA on the training set and looked for eigenvalues that are greater than 1, which are factors that contain more amount of information than a single variable. I found that there are 14 eigenvalues greater than 1, so I kept the first 14 columns of PCs. A 3D visualization of the genres as clusters is given above. We can see that for each genre of music, there are also clear distinctions, as there are 2 or 4 (green and brown) sticks of the same color. This shows diversity across genres. Among jazz and country music, they each have 4 clusters for each genre.

Next, I did a clustering step and included the labels in the dataset as a new feature. It would also boost our classification performance potentially. Elbow method and Silhouette method both showed that the optimal number of clusters is 12. Thus, I included this feature for every row of data in both training and testing sets.

For the multi-class classification problem, the neural network would be able to identify whether the classification was correct or not. Since models with activation function yields better performance than those without generally, I considered cases with activation functions. I built feedforward neural network of one-hidden and two-hidden layers (with ReLU, Tanh, Sigmoid respectively) and assessed their accuracy. The input size is 15 features, and the output size is 10,



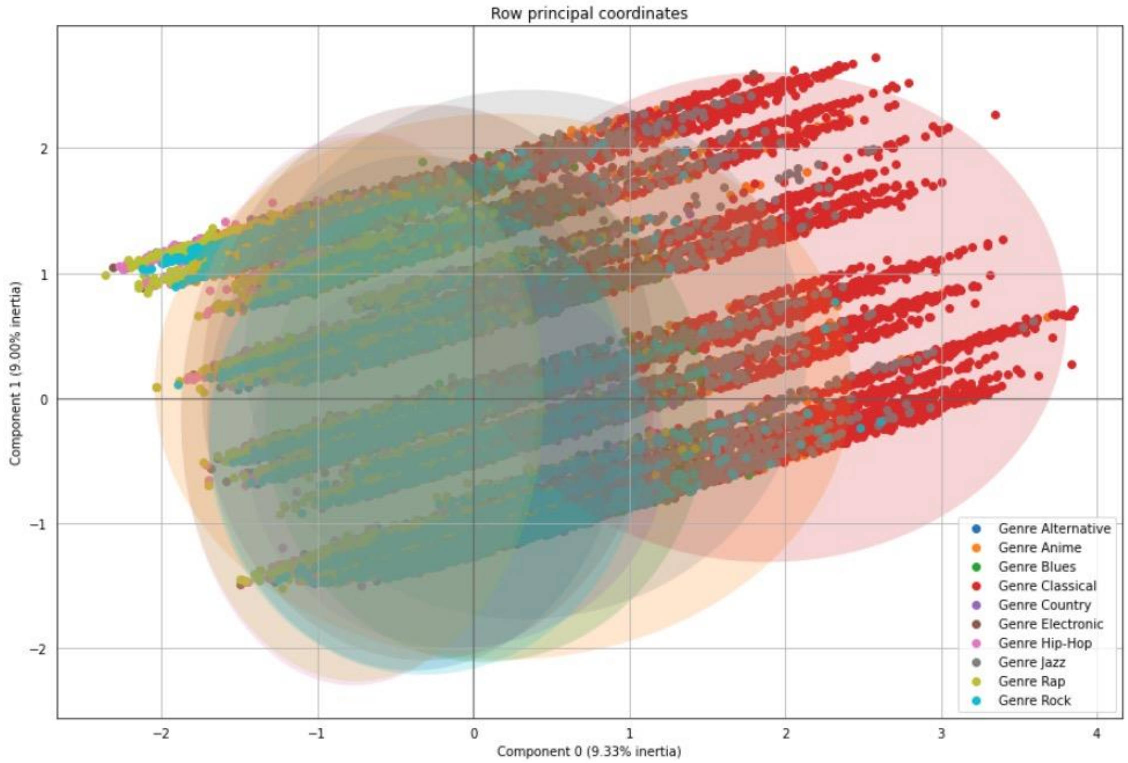
as there are 10 music genres. At the output layer, I applied a Softmax activation function to get the probabilities of the classes for future plotting (ROC curve) and calculations (AUC). Since it's a multiclass classification problem, I chose to use mean of One-vs-rest AUC score to assess model performance. I used lists to store the models and dictionaries to store the correspondence for better comparison.

I found that two-hidden layer ReLU has the highest mean OvR AUC score among all six models, at about 0.866. I also tried 3-hidden layer ReLU model, but the AUC is lower than 0.866, so I think **0.866** is the highest value I can get (there are fluctuations in terms of AUC scores, and 0.866 is the highest I get after multiple runs of code). I plotted the ROC curves of each music genre for the 2-hidden layer ReLU model.

Overall, I think the most important factor for classification is dimensional reduction, as it uses less features to incorporate as much information as possible of the entire dataset (25 columns after one-hot encoding). Adding clustering labels is also crucial to boost the performance of the model.

Some non-trivial observations would be that prior to using PCA, I tried to use FAMD (factor analysis of mixed data), as it works well with categorical variables. After dimensional reduction via FAMD, I noticed that there is huge overlapping between genres. However, classical music seems to be somewhat distant from all other genres based on the 2D plot, indicating that it's easier to categorize the specific genre (see next page for the plot).

```
defaultdict(list,
            {'one_layer': {'relu': 0.8392835777777778,
              'sigmoid': 0.8059914666666665,
              'tanh': 0.8245668888888888},
             'two_layer': {'relu': 0.8464959555555556,
              'sigmoid': 0.7834593333333333,
              'tanh': 0.8373469555555555}})
```



FAMD Dimensional Reduction 2D Plot

Final AUC: 0.866

Clustering technique is referenced from https://towardsdatascience.com/cluster-then-predict-for-classification-tasks-142fdfdc87d6.

ROC curves plotting is referenced from https://towardsdatascience.com/multiclass-classification-evaluation-with-roc-curves-and-roc-auc-294fd4617e3a.