# Classification Task for Genre

## 1 Question 4

**Research Question:** Can the genre of a track be accurately predicted using its acoustic features with machine learning models in PySpark?

We will explore the genre classification problem presented in the question, which we believe will hold significant relevance for real world applications such as music recommendation systems. The challenge involves leveraging audio features, such as danceability, energy, and loudness, to predict the genre of a track. We hypothesize that specific acoustic features are strong indicators of a track's genre. We expect that machine learning models can effectively employ these features to classify tracks into their respective genres, revealing their significance in genre prediction across various music classes.

### 1.1 Exploratory Data Analysis (EDA)

Before diving into model fitting, performing exploratory data analysis (EDA) is essential to thoroughly understand our dataset's structure. This involves checking data types for accuracy, calculating basic statistics for numerical and categorical fields, and assessing missing values to determine the need for data cleaning or imputation. This process ensures that our models are built on complete and correct information.

We analyze distributions, skewness, outliers in numerical features, and category frequencies in categorical features to aid in feature selection. Correlation analysis is conducted to identify potential multicollinearity among numerical predictors and to examine how features correlate with genres, revealing distinct patterns that could be crucial for genre classification. Visualization tools like Matplotlib or Seaborn are employed to create plots, enhancing our understanding of the data's characteristics and anomalies. This comprehensive approach not only refines our dataset but also helps select significant variables, effectively supporting our research question.

**Key Findings:**

- Many numerical variables in the dataset are stored as strings and require conversion to numerical formats for analysis.

- Certain categorical variables like 'key', 'mode', and 'time_signature' are recorded numerically. However, some values are outliers and do not correspond to any known categories. Due to the limited variation and unclear significance of these variables, we opted not to include them in our model. This decision is based on the assumption that their impact on genre prediction is minimal and the difficulty in handling the anomalous values.

- The dataset exhibits missing values in key fields; specifically, 'danceability' has 11 missing entries and 'energy' has one. Given the substantial size of our dataset, we decided to omit these incomplete records from our analysis.

- A small number of 'genre' entries, such as "210613," were identified as non-standard and unrecognizable as genre names. Out of a total of 9,198 genre entries, 36 were such non-standard types. Considering their negligible representation, these entries were excluded from further analysis, refining our dataset to 9,162 entries across recognized genres including pop, blues, electronic, hip-hop, jazz, rock, reggae, and classical. It is worth noting that each class is well-balanced with a substantial number of observations in each category.

- The histogram and correlation matrix of the features are displayed in Figures **??**. The histogram indicates that some variables are heavily skewed and contain outliers, which supports the choice of a tree-based model for classification because such models are not significantly affected by these issues. The correlation matrix shows that most features do not have strong correlations with each other, implying that they are likely important and should be kept for predictive modeling.

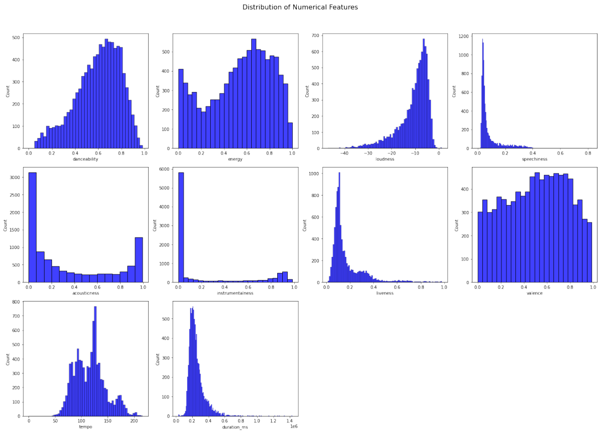

**Figure 1:** Histogram

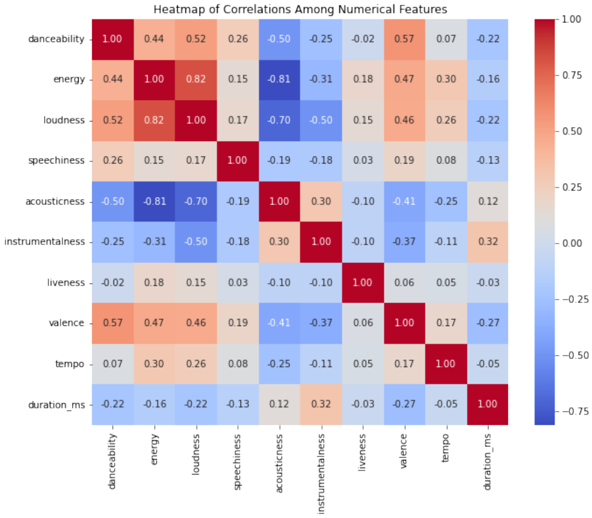

**Figure 2:** Heatmap

## 1.2 Modelling

In our quest to identify the most suitable model for our genre classification challenge, we considered a diverse array of approaches ranging from logistic regression, a linear-based classification method, to more complex techniques such as neural networks. Each model offers distinct advantages and potential drawbacks based on the specifics of the task and the characteristics of our dataset.

After thorough analysis and deliberation informed by the insights gained during our exploratory data analysis (EDA), we decided to implement a tree-based model, specifically the Random Forest algorithm. The decision was influenced by several factors. First, logistic regression, while effective for binary classification, becomes less straightforward when dealing with multi-class problems like ours, which includes eight distinct genre categories. Although logistic regression can be adapted to multi-class settings through strategies like one-vs-rest (OvR) or multinomial logistic regression, these approaches complicate the model without necessarily providing the best results for our needs.

On the other hand, neural networks are known for their high accuracy and ability to model complex nonlinear relationships. However, they require extensive computational resources to train, particularly with large datasets like ours. Moreover, neural networks operate as "black boxes," offering limited interpretability regarding how decisions are made, which is a significant drawback for our project. Understanding feature importance and how each variable influences the model is crucial for our analysis, particularly when explaining the results to stakeholders or refining the model based on specific genre characteristics.

Random Forest, a robust ensemble method, addresses many of these concerns. It is well-suited for multi-class classification and is inherently capable of handling feature importance analysis, which allows for greater interpretability of the model. Additionally, Random Forest models tend to be less prone to overfitting and can handle large datasets efficiently, making them ideal for our current application in music genre classification. Moreover, since the trees in a Random Forest split nodes on subsets of features and samples, outliers have less influence compared to other algorithms that might scale according to extreme values. This inherent characteristic of decision trees to isolate outliers in specific branches minimizes their impact on the overall model, and further enhances its applicability to our dataset.

**Hyperparameters Tuning:**

Despite its advantages, ensuring optimal performance of Random Forest in practical applications requires attention to several factors:

- Feature Selection: Excessive number of features can dilute the influence of truly significant predictors, degrade model interpretability, increase the risk of incorporating noise that can mislead the learning algorithm, and substantially increase complexity and potentially lead to longer training times and model overfitting. Nevertheless, the need to reduce dimensionality and improve performance through feature selection might not be urgent in our case—given we only have 10 predictors—it remains a good practice to evaluate the significance and impact of each predictor on the model's accuracy.

- Class Balance: Fortunately, in our situation, each class is already well-balanced, featuring a substantial number of observations per category, which simplifies the modeling process. Otherwise, employing techniques such as stratified sampling is critical to maintain class balance throughout the training process.

- Hyperparameter Tuning: The tuning of hyperparameters is a crucial step to harness the full potential of the Random Forest algorithm. Properly selecting the number of trees ('numTrees') and the maximum depth of the trees ('maxDepth') can significantly influence the model's performance. More trees in the forest generally lead to better performance and lower risk of overfitting, but also increase computational cost. Similarly, a deeper tree might capture more detailed data specificity but could lead to overfitting if not checked with other parameters like minimum samples per leaf ('minSamplesLeaf') or minimum sample split ('minSamplesSplit'). Utilizing grid search and cross validation methods to experiment with these parameters systematically can help in identifying the best combination that enhances prediction accuracy while keeping the computational expense in check. This approach not only refines the model to adapt to specific data traits but also optimizes its predictive accuracy and generalization capability across unseen data.

In addressing our specific classification problem, we have determined that the optimal set of hyperparameters includes setting 'numTrees' to 100 and 'maxDepth' to 20. After extensive testing and validation, these settings have proven to enhance prediction accuracy significantly. For other parameters such as minSamplesLeaf and minSamplesSplit, maintaining the default settings has yielded satisfactory results. This configuration strikes a balance between model complexity and computational efficiency while ensuring robust performance.

## 1.3 Outcomes

By carefully fine-tuning the parameters of our Random Forest model, our goal is to enhance the model's capability to generalize effectively to unseen data while preventing overfitting. We have achieved a modeling accuracy of 0.7346 on the test set, which is noteworthy considering that we've employed a train-test split methodology for model evaluation, ensuring that model training was strictly performed on the training set only.

Additionally, the importance of each feature in the model has been quantitatively assessed, as shown in the accompanying table 1. This analysis reveals that all ten features contribute significantly to the model's predictions, albeit to varying degrees. This distribution of feature importance reinforces our decision to retain all features in the model, as each one adds valuable information that enhances the model's predictive performance.

Finally, we noted that the model excels in predicting genres with distinct attributes such as classical, hip-hop, and electronic music, as evidenced by Figure 3. Additionally, Blues and rock often share overlaps in instrumentals and production techniques, leading to 10.09% of rock being misclassified as pop and 19.10% of pop as rock. Similarly, Reggae and hip-hop share rhythmic and cultural similarities, often blending in modern music, which might explain why 8.74% of hip-hop is predicted as reggae and 14.73% of reggae as hip-hop. This points to the complex interplay of genre characteristics influencing predictive accuracy.

**Table 1:** Feature Importances from Random Forest Model

| Feature | Importance |
|---|---|
| Danceability | 0.1572 |
| Energy | 0.1173 |
| Loudness | 0.0855 |
| Speechiness | 0.0606 |
| Acousticness | 0.1228 |
| Instrumentalness | 0.1419 |
| Liveness | 0.0542 |
| Valence | 0.1028 |
| Tempo | 0.0785 |
| Duration ms | 0.0793 |

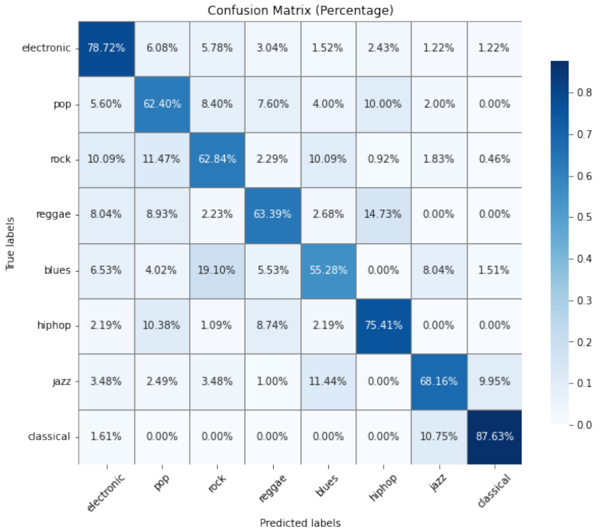

**Figure 3:** Confusion Matrix