Jiaxin Liu

Prof. Jacob Koehler

Data bootcamp

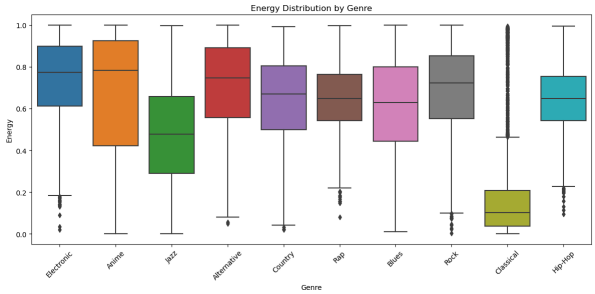
12 May 2025

Introduction: The objective of this project was to predict the genre of a song based solely on its audio features. To achieve this, I began by exploring the dataset through exploratory data analysis (EDA), followed by preprocessing and dimensionality reduction. After preparing the data, I trained and compared three distinct classification models: Neural Network, Random Forest, and Gradient Boosting. Among these models, Gradient Boosting performed the best, achieving a test accuracy of 58.04% and a macro-average AUC of 0.9318. These findings underscore both the opportunities and challenges associated with audio-based genre classification, particularly in differentiating stylistically similar genres such as Hip-Hop and Rap. Conversely, more distinctive genres, like Classical and Anime, were relatively easier to classify.

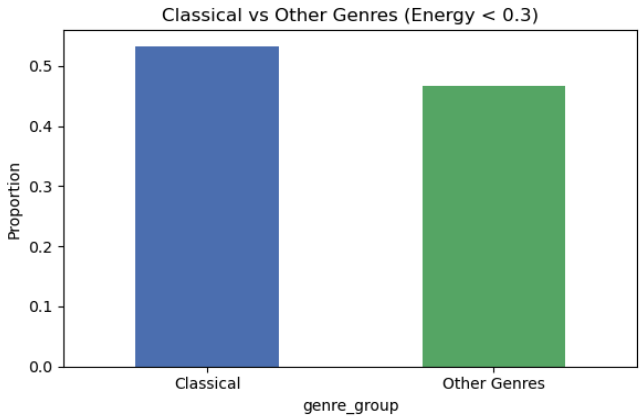
Data Description: The dataset comprises audio features extracted from 50,000 songs through the Spotify API, with equal representation (5,000 songs each) across ten diverse genres, including Classical, Hip-Hop, Pop, Anime, Electronic, and Jazz. Each song is represented by metadata and numerical audio features such as acousticness, danceability, energy, and tempo, alongside categorical features like key and mode. The dataset contains 18 columns, including Metadata: instance_id, artist_name, track_name, obtained_date, popularity, and numerical audio features: acousticness, danceability, duration_ms, energy, instrumentalness, liveness, loudness, speechiness, tempo, valence, and categorical audio features: key and mode. The target variable is

music_genre, which includes 10 music genres. For data cleaning, I used dataset.isnull().any(axis=1) to test how many rows have empty values, and then I checked which rows have empty value, I found that rows 10000, 10001, 10002, 10003, and 10004 are fully empty rows, so I used dataset.dropna(how="all") to drop them. Duration cannot be less than or equal to 0, so I replaced those durations (which <=0) with the duration median. The keys were numerically encoded from musical notations (“C”, “C#’, “D”, etc.) into integers ranging from 0 to 11, and the mode column was converted to a binary feature is_major (1: major, 0: minor). Irrelevant columns like instance_id, artist_name, track_name, and obtained_date were removed as they do not contribute to the predictive power of genre classification. Labels (music_genre) were encoded using LabelEncoder, and only numerical features were scaled using StandardScaler to prepare for PCA and model training. A stratified split ensured that 500 samples from each genre were included in the test set and the other 4500 songs were included in the training set, preserving class balance.

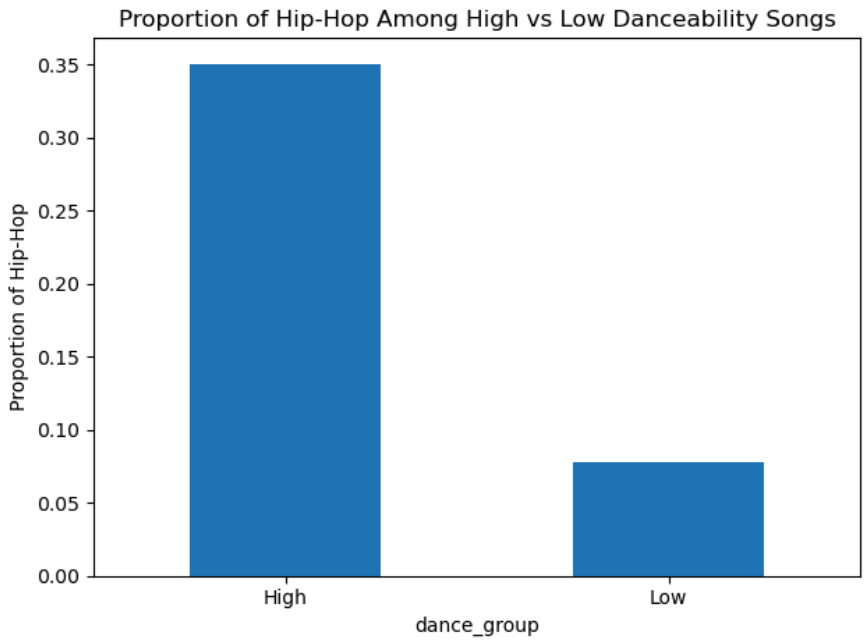
The exploratory data analysis revealed distinct genre-level patterns. I plotted 4 plots, the first one is Boxplot of energy by genre. This boxplot illustrates the distribution of energy levels across different music genres. Classical and Jazz exhibit the lowest median energy and tightest ranges, while genres like Hip-Hop and EDM show significantly higher energy levels and broader distributions. This reinforces the idea that energy is a genre-sensitive feature, which can be leveraged in genre classification algorithms, playlist generation, or mood-based music recommendations. For instance, when users seek calm or high-intensity music, energy can serve as a reliable filter to distinguish between genres that align with those moods or activities.



the second plot is Genre breakdown for low-energy songs (energy < 0.3). The bar plot comparing Classical music to all other genres among songs with energy below 0.3 reveals that Classical alone accounts for over 50% of low-energy tracks, while all other genres combined make up less than half. This strong concentration underscores Classical music's intrinsic association with low-energy characteristics such as soft dynamics, slower tempo, and acoustic instrumentation. The insight is significant not just musically, but also for practical applications: in mood-based recommendation systems or playlist generation, Classical emerges as the most statistically reliable genre for calm and relaxing audio content. Moreover, this genre skew highlights the potential value of using genre–energy interactions as engineered features in predictive modeling, especially when classifying or recommending songs based on audio profiles.



The third plot is for test Is Hip-Hop more likely to be high-danceability compared to other genres. This plot reveals a strong association between the Hip-Hop genre and high danceability. Among songs with a danceability score above 0.8, Hip-Hop accounts for approximately 35% of the tracks, while it represents only around 7% among low-danceability songs. This sharp contrast indicates that Hip-Hop is structurally optimized for movement and rhythm, often featuring steady beats and engaging tempos. This insight can directly inform the design of music recommendation algorithms or playlist curation engines — for example, if a user seeks high-danceability tracks (e.g., for a party or workout), the system can prioritize Hip-Hop tracks with high confidence. Additionally, it supports the idea of including genre–feature interaction terms (e.g., danceability × genre) when building predictive models for genre classification or music mood tagging.

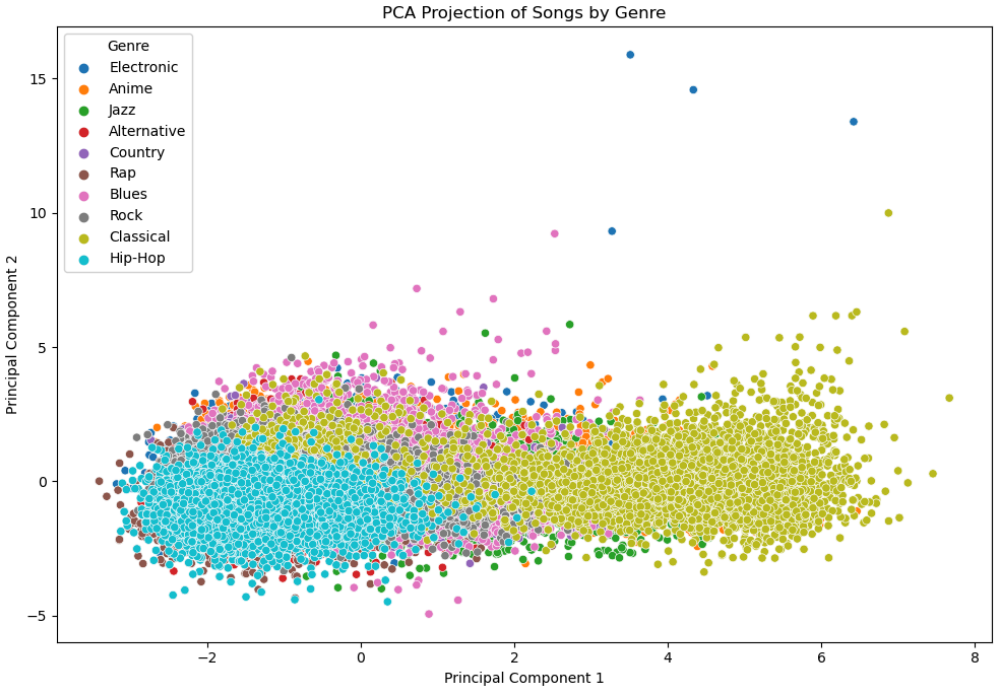


The fourth plot is about the proportion of High-Speechiness Songs (Speechiness > 0.66): Hip-Hop vs Other Genres. This plot focuses on songs with high speechiness (speechiness >

0.66), a range typically associated with spoken-word content such as rapping or dialogue. Among these tracks, Hip-Hop dominates, comprising the majority of songs in this segment. This sharp concentration illustrates that Hip-Hop is uniquely characterized by high vocal presence and rhythmic speech, distinguishing it from more instrumentally focused genres. This insight can guide genre classification efforts and speech-related audio tagging — for instance, a track with high speechiness is much more likely to be Hip-Hop, and this pattern can be used to improve automated genre prediction, voice activity detection, or even assist music platforms in flagging podcast-like content in song databases.

Models and Methods: Handling the complexity and high dimensionality of audio features necessitated a structured approach, beginning with dimensionality reduction. Therefore, Principal Component Analysis (PCA) was applied to simplify the feature space, facilitating easier interpretation and visualization. PCA's resulting 2D scatter plot clearly indicated genre-specific clustering. Notably, Hip-Hop and Electronic formed distinct, identifiable clusters, while Pop, Anime, and Country demonstrated moderate overlap. However, the Rock and Blues genres were dispersed widely, making them difficult to visually differentiate. This observation aligns with the expectation that certain genres share acoustic similarities, posing inherent challenges in genre classification.

Building upon insights from the exploratory analysis and PCA, I selected and implemented three distinct classification algorithms to compare performance systematically.



After that, I implemented three different classification models. The first one is Neural Network with a three-layer network consisting of two hidden layers 64 and 32 neurons respectively (using ReLU activations and a final output layer for the 10 genre classes, I chose ReLU activations because it is computationally efficient and vanishing gradient problem, and I trained using the Adam optimizer with a learning rate of 0.0005 over 80 epochs. The learning rate of 0.0005 is to avoid converging too quickly, allowing the network to gradually minimize the loss. I chose Neural Network because music features like energy, danceability, and tempo are likely to have non-linear patterns. Also, this dataset has 10 genres, neural networks are good at handling multiple output classes, especially when paired with softmax activation in the output layer. The training loss curve indicates the network successfully learned the data patterns, because there is a rapid initial loss reduction followed by convergence to a stable level. The second one is Random Forest Classifier, which is an ensemble model with 100 decision trees. The large number of trees

helps to reduce variance, making the model more accurate. I chose a random forest classifier because it can handle noisy data very well, and it can capture complex relationships without overfitting as easily as single decision trees. The third one is Gradient Boosting Classifier, which builds sequential trees to correct prior mistakes. I created it with 100 boosting iterations, 100 estimators, a learning rate of 0.1, and a max depth of 3. 100 estimators can offer a good balance between performance and overfitting; more estimators could lead to overfitting, especially in a high-dimensional dataset. The max depth of 3 can avoid overfitting. I chose a gradient boosting classifier because it focuses on reducing bias and often yields high performance.

Results and Interpretation: For evaluating their performance, I used accuracy and macro-average AUC for qualitative comparison. The accuracy can tell me how many correct predictions, and AUC is included because accuracy alone can be misleading in multi-class classification, especially when some genres are more difficult to distinguish than others. Accuracy simply measures the proportion of correct predictions, but it does not reveal how well the model ranks predictions or handles uncertainty between similar classes. The macro-average AUC, on the other hand, evaluates the model's ability to distinguish each class from the others across all possible thresholds, treating each class equally regardless of frequency. Also, I plotted the Confusion Matrix and the ROC Curve for visualization comparison. I created a table to summarize the performance of those models.

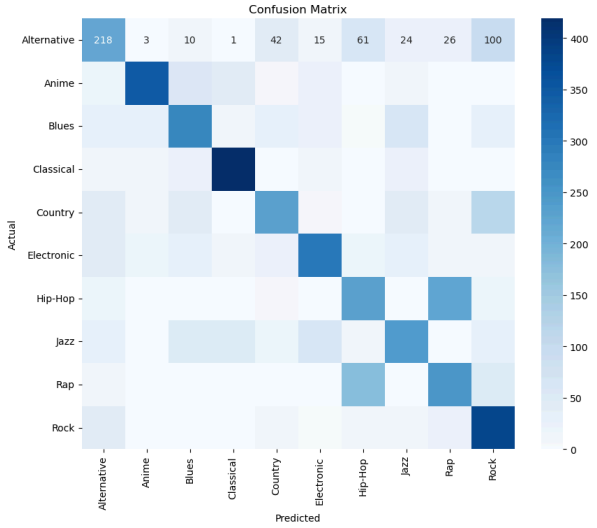
Model	Accuracy	Macro-average AUC
Neural Network	0.5756	0.9267
Random Forest	0.5396	0.9153

Gradient Boosting	0.5804	0.9318

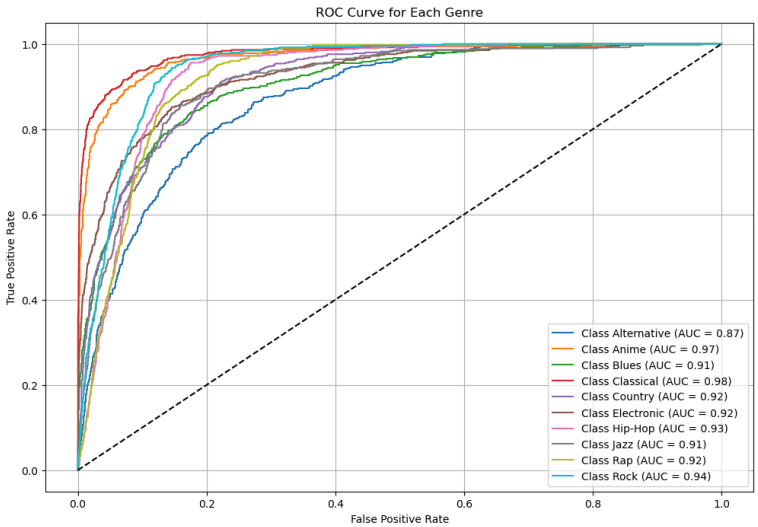
Neural Network: The Neural Network achieved an accuracy of 57.56% and a macro-average AUC of 0.9267, indicating solid overall performance in distinguishing between music genres. The training loss curve shows a rapid initial decline followed by a stable plateau, suggesting that the model successfully minimized the loss and converged without overfitting. However, the presence of noise and the vertical lines across epochs indicate fluctuations in mini-batch gradients, which might have hindered further optimization.

The confusion matrix reveals that the Neural Network performs especially well on clearly defined genres such as Classical, with a strong diagonal value, aligning with prior EDA insights where Classical showed distinctive low energy and minimal overlap with other genres. Anime and Rock also display relatively accurate classification. However, the model struggles to distinguish between Rap and Hip-Hop, with frequent misclassifications between these two genres. This reflects the acoustic similarity captured in the features — both genres share high energy, speechiness, and rhythmic patterns, as confirmed by the EDA. The confusion between these categories signals a limitation in the model's ability to parse subtle stylistic nuances when relying solely on quantitative audio features. The ROC curve supports these findings, with Classical achieving the highest AUC (~0.98), showing that the model is highly confident and correct in separating this genre from others. Conversely, Alternative music exhibits the lowest AUC (~0.87), indicating difficulty in establishing clear genre boundaries, likely due to its broader stylistic variability and overlap with multiple genres like Rock and Blues.

Taken together, while the Neural Network is able to capture non-linear relationships among features and perform reasonably well in genre classification, its limitations become apparent when genres have overlapping acoustic profiles. This suggests a need for additional feature engineering — possibly incorporating temporal or lyrical features — to help the model better distinguish between nuanced categories like Rap and Hip-Hop.



(Confusion matrix of Neural Network)

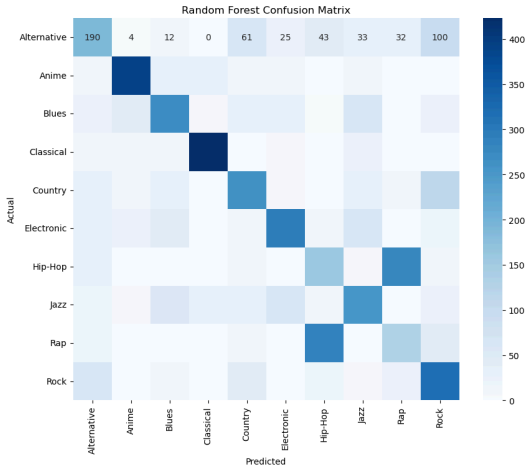


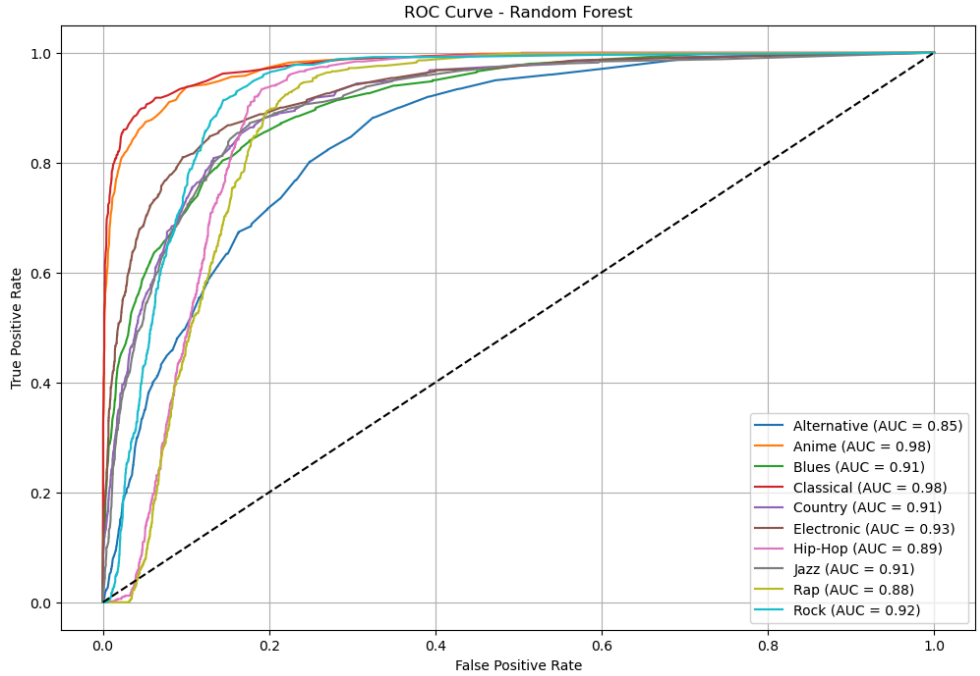
Random Forest: The Random Forest model achieved an accuracy of 53.96% and a macro-average AUC of 0.9153, which, while slightly lower than the Neural Network and Gradient Boosting, still reflects reasonable performance in a 10-class classification task.

The confusion matrix shows that the Classical genre is predicted with the highest accuracy, evidenced by the darkest cell along the diagonal. This aligns with findings from EDA: Classical music's consistently low energy and minimal rhythmic variation make it an outlier, allowing tree-based models to isolate it with fewer decision splits. Other relatively well-predicted genres include Rock and Anime, likely due to distinct tempo and energy patterns. However, the model exhibits significant confusion between Hip-Hop and Rap, as seen by high off-diagonal values between these two classes. This again supports earlier observations from the EDA: while Hip-Hop and Rap differ culturally, they share similar acoustic features such as high speechiness and rhythmic intensity, which leads to ambiguity in feature space. The Random Forest, being a collection of non-linear but axis-aligned decision trees, may struggle to draw nuanced boundaries when genres occupy overlapping regions. The ROC curves confirm these observations. Classical and Anime have sharp curves with high AUCs (~0.98 and ~0.97), indicating the model is confident in identifying them correctly. In contrast, Alternative and Hip-Hop have more gradual curves (AUC ~0.87–0.91), reflecting the model's uncertainty and weaker separability for those genres.

Overall, while Random Forest handles noisy features well and does not overfit easily, its limited expressiveness compared to neural networks or boosting methods may restrict its ability to capture complex interactions, particularly important in distinguishing between acoustically

similar genres. Enhancements such as feature interactions (e.g., speechiness × tempo) or deeper feature engineering could improve performance, especially in resolving genre pairs with high confusion rates.



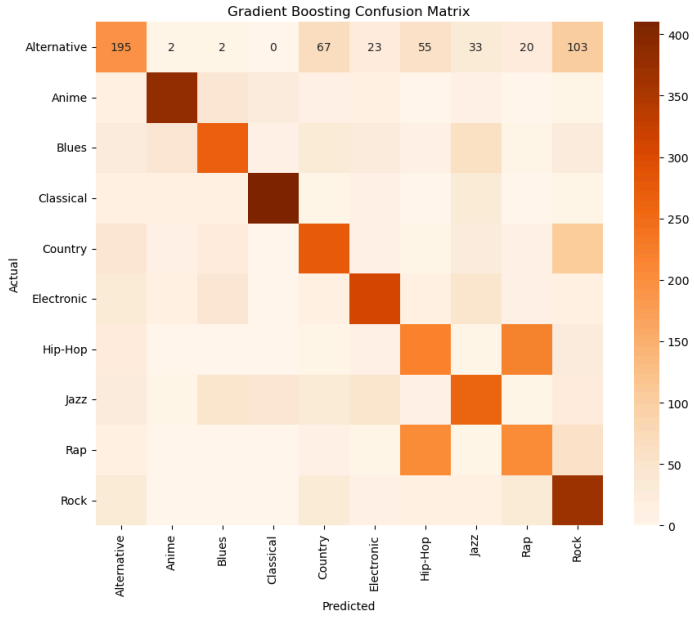


Gradient Boosting: The Gradient Boosting Classifier delivered the best overall performance, with a test accuracy of 58.04% and a macro-average AUC of 0.9318, outperforming both the Neural Network and Random Forest models. This strong performance reflects Gradient

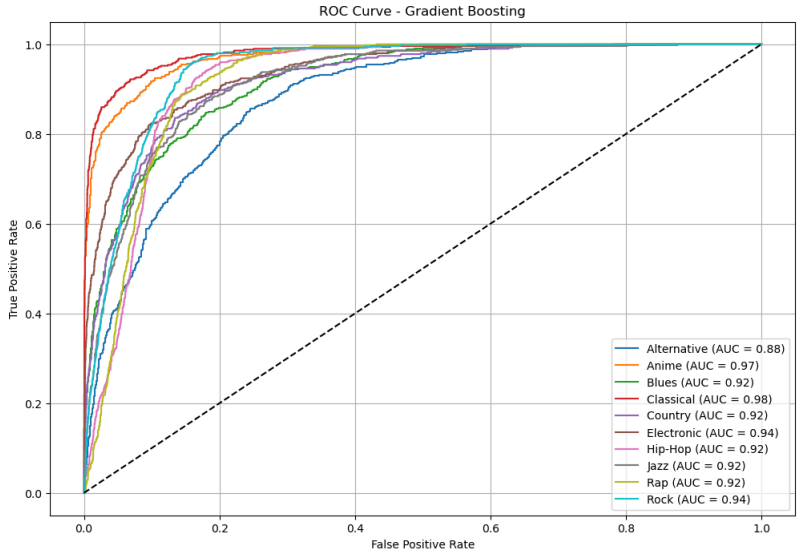
Boosting's ability to iteratively correct the errors of prior trees and capture complex, non-linear patterns in the dataset.

The confusion matrix reveals particularly strong predictive accuracy for the Classical and Anime genres, each showing dense diagonal entries. These results are consistent with the exploratory data analysis, which highlighted Classical's distinctively low energy and Anime's unique tempo and mood profiles. The model effectively isolates these genres because their audio features cluster cleanly in the feature space, making them easier to classify with minimal error. On the other hand, the matrix also shows persistent confusion between Hip-Hop and Rap, a recurring pattern across all models. Despite the EDA highlighting speechiness and danceability as distinguishing features, the acoustic similarity between these genres still leads to frequent misclassification. This suggests that even advanced boosting methods face challenges when feature distributions overlap heavily and genre boundaries are more cultural or lyrical than numerical. The ROC curves further illustrate Gradient Boosting's strengths: genres like Classical, Anime, and Rock achieve AUC scores above 0.94, showing excellent class separation and model confidence. Even genres with broader stylistic boundaries, such as Jazz and Blues, maintain high AUCs (~0.91), indicating the model's robustness across diverse genres. What sets Gradient Boosting apart is its ability to focus learning on difficult cases, such as previously misclassified songs, by adjusting weights and error gradients at each iteration. This gives it a competitive edge in balancing bias and variance, which is particularly valuable in multi-class tasks like genre classification. Additionally, the model's interpretability (via feature importance or SHAP values) allows for further refinement by highlighting which features drive decisions for each genre. In summary, Gradient Boosting excels in both accuracy and discriminative power

across genres, especially when genres have distinct acoustic profiles. However, its limitations in classifying overlapping genres like Hip-Hop and Rap suggest a ceiling in what can be achieved using only numerical audio features. Integrating lyrical content, rhythm patterns, or higher-level semantic audio features could further improve genre separation in future iterations.



(Confusion matrix of gradient boosting)



Conclusion: This project explored the task of music genre classification using audio features from 50,000 Spotify tracks across 10 genres. Exploratory data analysis revealed strong genre-specific patterns—Classical music dominated low-energy segments, Hip-Hop was strongly associated with high speechiness and danceability, and Anime exhibited a distinct tempo profile. These insights shaped model selection and offered early evidence for which genres might be easier or harder to classify. Among the models tested, Gradient Boosting achieved the best performance with an accuracy of 58.04% and a macro-average AUC of 0.9318, outperforming the Neural Network and Random Forest classifiers. Its sequential learning process allowed it to capture complex relationships and adjust to misclassified examples effectively. Classical and Anime were predicted with high confidence and precision, consistent with their distinct feature profiles in the dataset. Conversely, Hip-Hop and Rap were frequently confused due to overlapping acoustic characteristics—a trend observed across all models. This limitation highlights that some genre distinctions may not be fully captured by standard audio features alone. The updated results emphasize that while genre classification using audio features is feasible and informative, it faces inherent challenges with stylistically similar genres. Future improvements may require incorporating lyrical analysis, rhythmic patterns, or domain-specific musical knowledge. Such enhancements could help bridge the gap between numerical patterns and cultural or stylistic genre identities.

Next Steps: To improve model performance and address persistent misclassifications, particularly between acoustically similar genres like Hip-Hop and Rap, future analysis should focus on both feature enrichment and model enhancement.

First, a deeper investigation into the confusion matrix using error analysis and misclassification heatmaps could help identify exactly which features are driving classification errors. Feature interaction terms (e.g., speechiness × tempo or danceability × genre) may be added to emphasize differences in borderline cases. Additionally, techniques like SHAP values can offer model interpretability, revealing which features contribute most to predictions for each genre. Second, incorporating domain-specific musical knowledge could yield substantial improvements. For instance, leveraging chord progressions, key modulations, or harmonic structures may help disambiguate genres with similar energy or rhythm patterns but distinct compositional styles. Natural Language Processing (NLP) could also be applied to lyrical content to distinguish between culturally similar genres, like Rap and Hip-Hop, that share acoustic traits but differ thematically. On the modeling side, more advanced architectures such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) could be explored to capture temporal dynamics or spectral patterns within audio signals, especially if raw audio or spectrogram data becomes available. Ensemble methods or stacked models combining boosting and deep learning could also be tested for further gains. For real-world applicability, this system could be integrated into music streaming platforms to assist in automated genre tagging, helping to classify newly uploaded or untagged tracks. It could also support mood-based recommendation systems or playlist generation, where inferred genre plays a role in matching user preferences. Improving genre prediction accuracy, especially for similar genres, would enhance the personalization and discovery experience for end users.