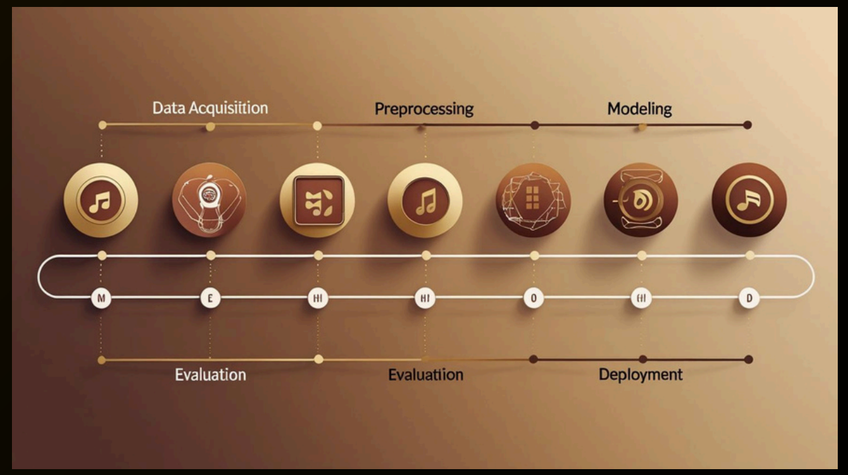
MUSIC GENRE CLASSIFICATION USING MACHINE LEARNING

Built By: Niat Kahsay & Marwah Faraj

August 2025

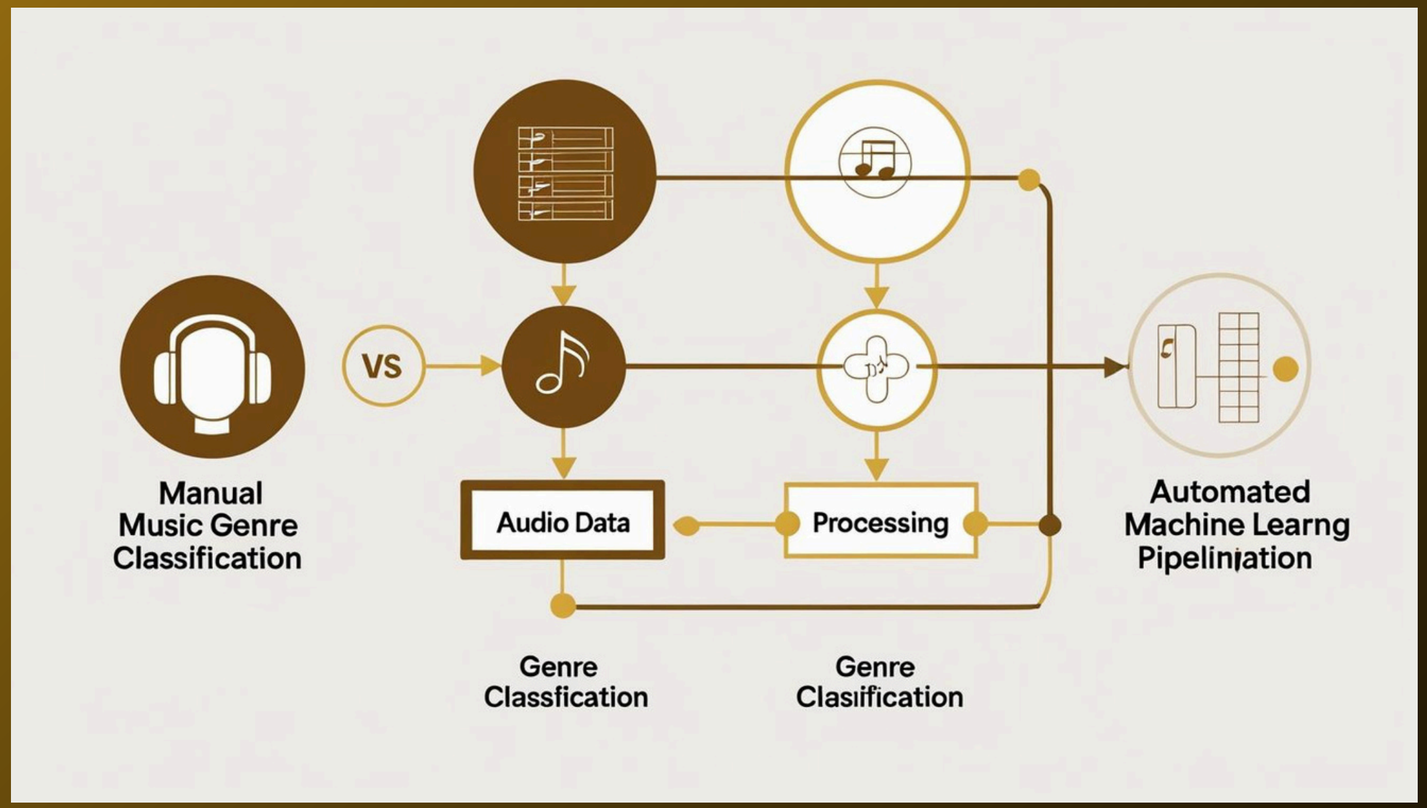
ABSTRACT OVERVIEW

- This study evaluates the performance of classical ML algorithms (SVM, Random Forest, Gradient Boosting, XGBoost), Deep learning and AutoML methods (AutoGluon) for music genre classification.
- Dataset consists of labeled audio tracks with extracted features including tempo, loudness, danceability, energy, and spectral characteristics.
- Primary objective: optimize classification accuracy while analyzing feature importance.
- Results indicate XGBoost outperforms others with an F1-score of 58%, driven primarily by tempo, loudness, and energy.



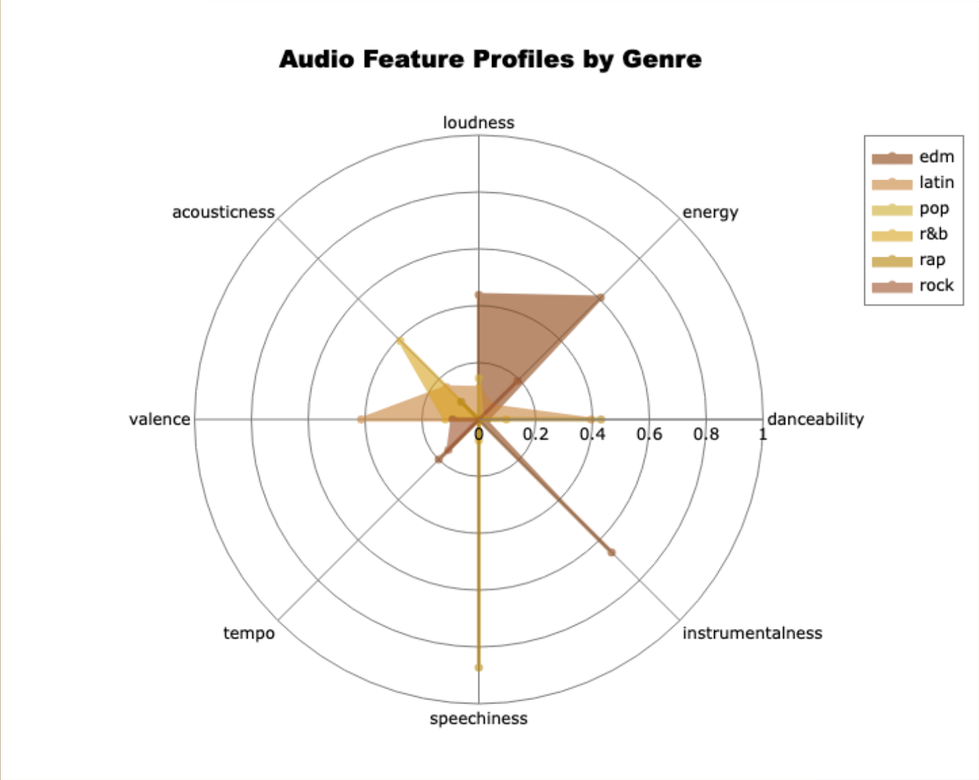
Motivation

- Manual music genre tagging is subjective and inconsistent across platforms.
- Growth of digital music libraries requires scalable, automated classification systems.
- Accurate genre classification improves recommendation systems, playlist curation, and copyright management.
- This research focuses on leveraging structured audio features (no raw waveform processing) for fast, explainable predictions.
- Visual: Schematic comparing manual tagging vs. automated ML pipeline.



Dataset & Features

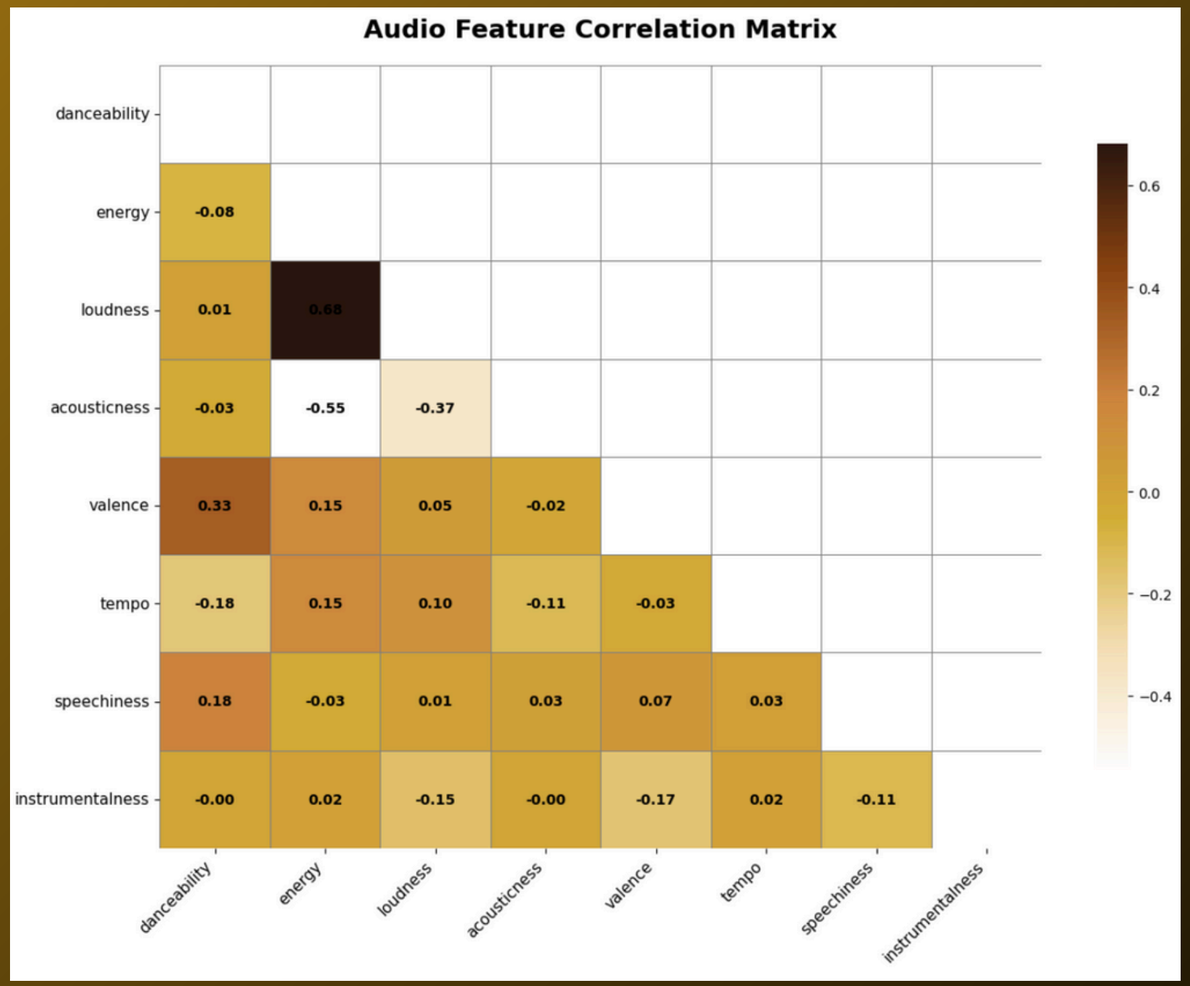
- Dataset: https://www.kaggle.com/datasets/rodolfofigueroa/spotify-12m-songs.
- Rows: 32833, Features: 23.
- Key features: tempo, loudness, danceability, energy, valence, acousticness, instrumentalness, speechiness.
- Target variable: music genre (10 classes).



Radar chart showing average feature profile by genre.

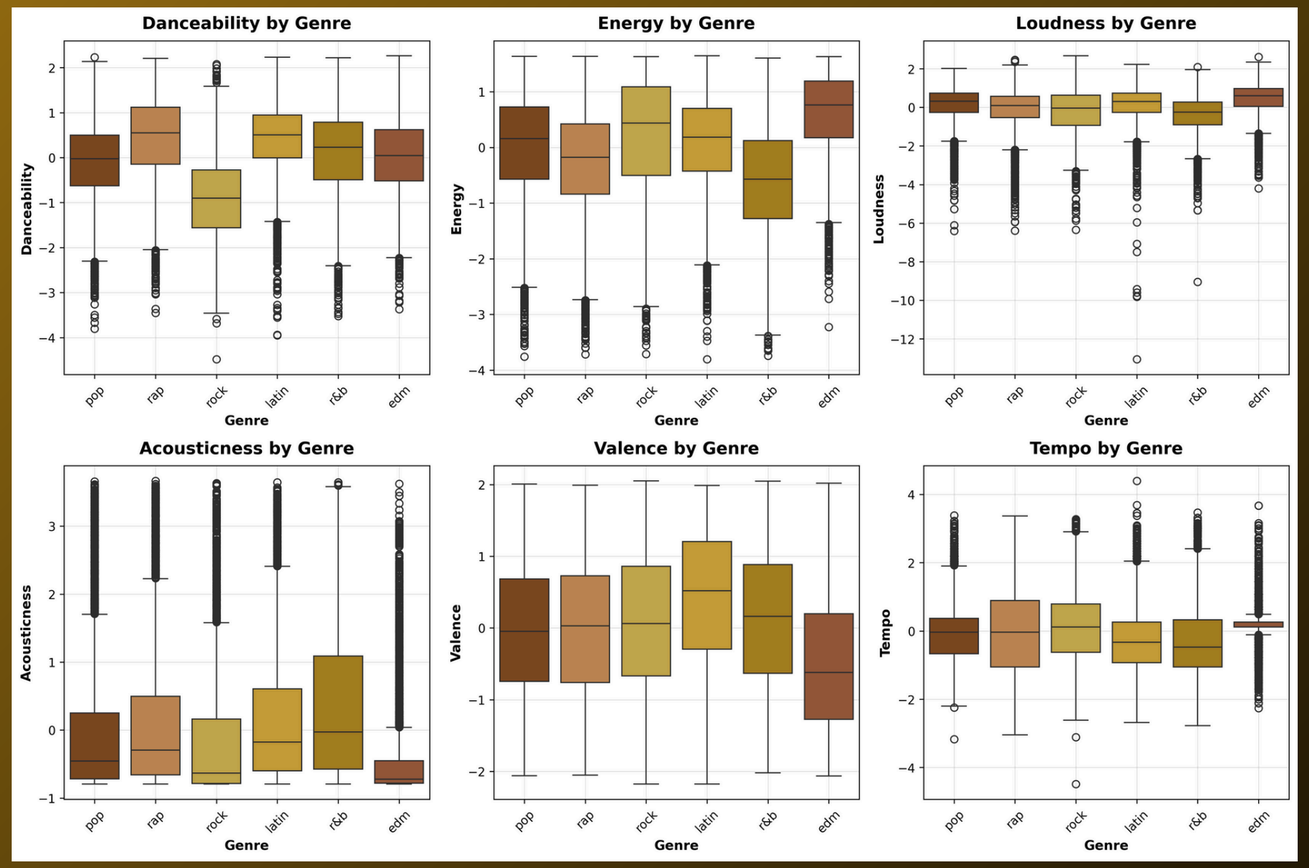
EDA

- Strong positive correlation between energy and loudness.
- High variance in tempo across genres.
- Overlapping distributions between Pop and Latin.



EDA

- EDM showed the highest energy, loudness, and tempo, with low acousticness.
- R&B had the highest acousticness and lower energy levels.
- Latin had the highest valence, while Rock showed wide variability in several features.
- Pop and Latin had overlapping danceability ranges, suggesting classification overlap.
- All features contained notable outliers, especially in tempo and loudness.



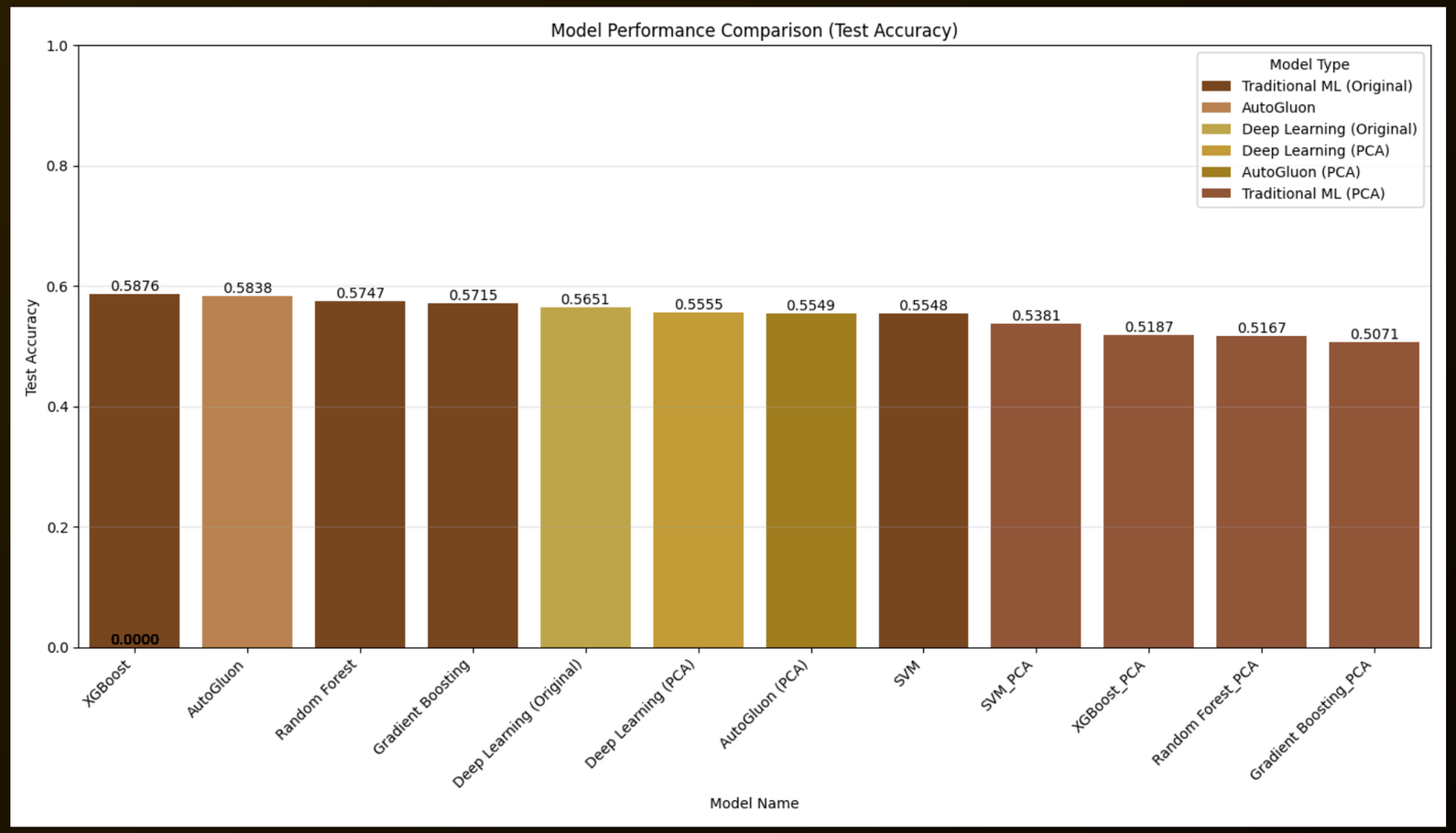
Models Tested

- Traditional ML: XGBoost, Random Forest, Gradient Boosting, Support Vector Machines.
- Deep Learning: Dense Neural Network with 3 hidden layers (ReLU activations, Dropout 0.3).
- AutoML: AutoGluon to optimize hyperparameters automatically.
- Hyperparameter tuning:
  - XGBoost: tuned max_depth, n_estimators, learning_rate.
  - Random Forest: tuned n_estimators, max_features.
  - DL model: tuned batch size & learning rate with Adam optimizer.

	Model Type	Model Name	Test Accuracy	Weighted F1 Score
0	Traditional ML (Original)	Random Forest	0.5747	NaN
1	Traditional ML (Original)	Gradient Boosting	0.5715	NaN
2	Traditional ML (Original)	XGBoost	0.5876	NaN
3	Traditional ML (Original)	SVM	0.5548	NaN
4	Traditional ML (PCA)	Random Forest_PCA	0.5167	0.5127
5	Traditional ML (PCA)	Gradient Boosting_PCA	0.5071	0.5055
6	Traditional ML (PCA)	XGBoost_PCA	0.5187	0.5159
7	Traditional ML (PCA)	SVM_PCA	0.5381	0.5350
8	Deep Learning (Original)	Deep Learning (Original)	0.5651	0.5940
9	Deep Learning (PCA)	Deep Learning (PCA)	0.5555	0.5509
10	AutoGluon	AutoGluon	0.5838	0.5825
11	AutoGluon (PCA)	AutoGluon (PCA)	0.5549	0.5497

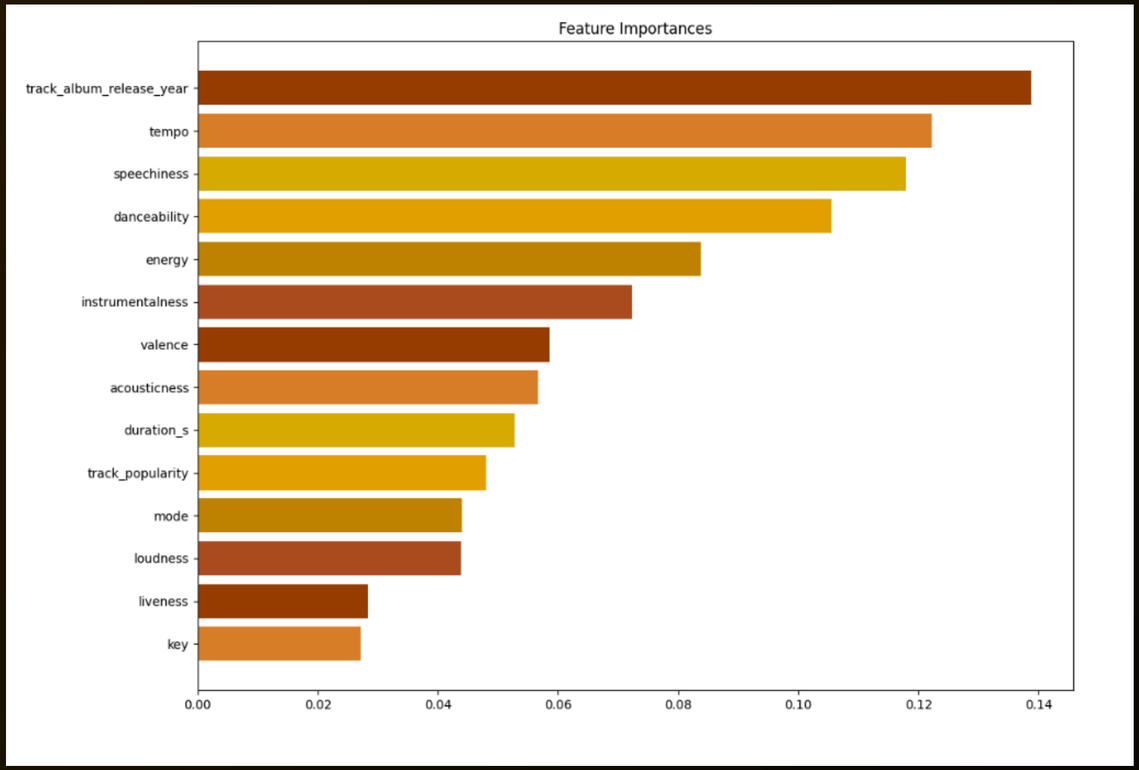
Results Summary

The comparison of model performances demonstrated that training without PCA generally resulted in higher test accuracy. Among all models evaluated, XGBoost without PCA achieved the highest accuracy (0.5876), followed closely by AutoGluon (0.5838) and Random Forest (0.5747). Models trained with PCA consistently showed slightly lower performance, indicating that dimensionality reduction did not enhance predictive accuracy for this dataset.



Feature Importance

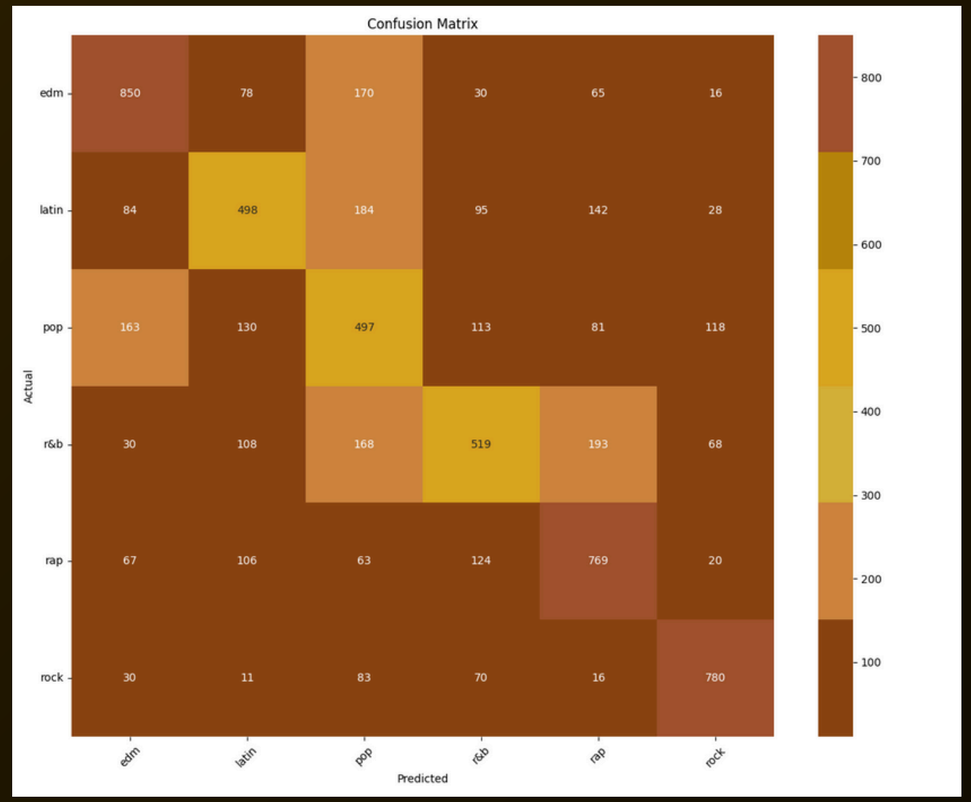
- XGBoost Feature Importance Analysis:
- Top predictors: danceability, energy, tempo.
- Lower-ranked: speechiness, acousticness.
- Interpretation: High-energy and rhythmic features dominate genre differentiation.
- Potential bias: Similar genres share feature patterns, affecting classification boundaries.

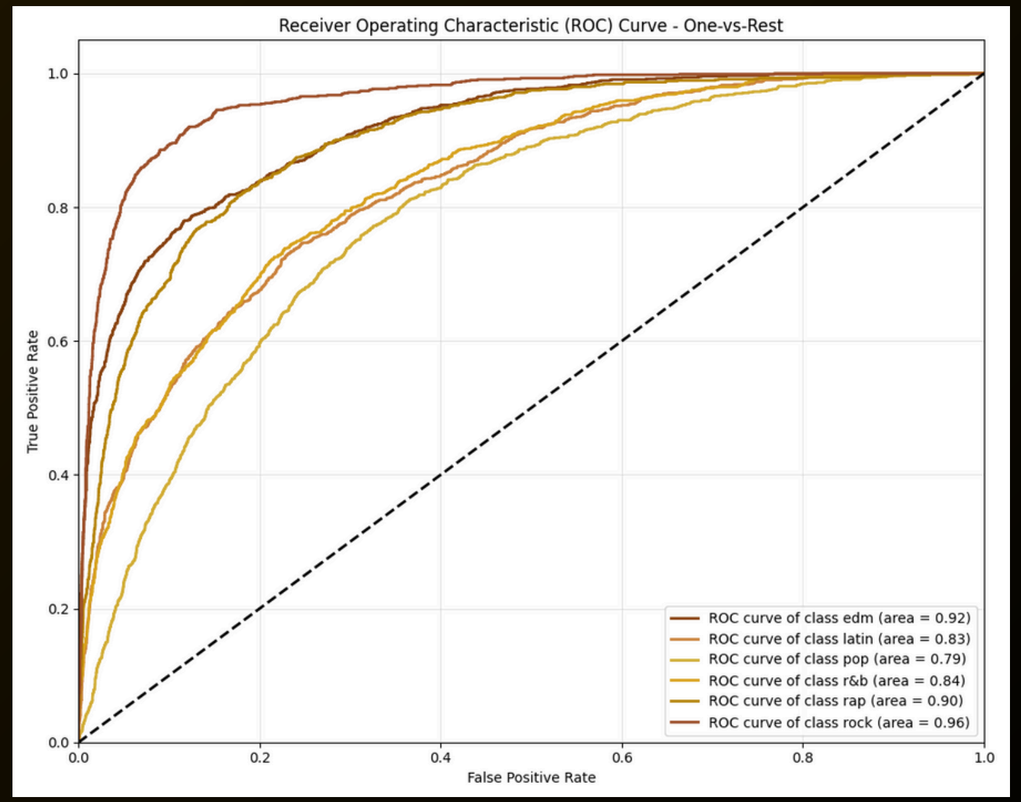


XGBoost feature importance plot.

Model Performance Evaluation

- The confusion matrix showed that most misclassifications occurred between similar genres, particularly Pop and Latin, and Pop and R&B, indicating overlap in feature space.
- High classification accuracy was achieved for EDM and Rock, as reflected by the larger diagonal values in those categories.
- ROC curves indicated strong discriminative ability for Rock (AUC = 0.96) and EDM (AUC = 0.92), while Pop (AUC = 0.79) and Latin (AUC = 0.83) demonstrated lower separability.
- Overall, the models exhibited better performance for classes with more distinct audio features and slightly weaker performance for classes with overlapping characteristics.





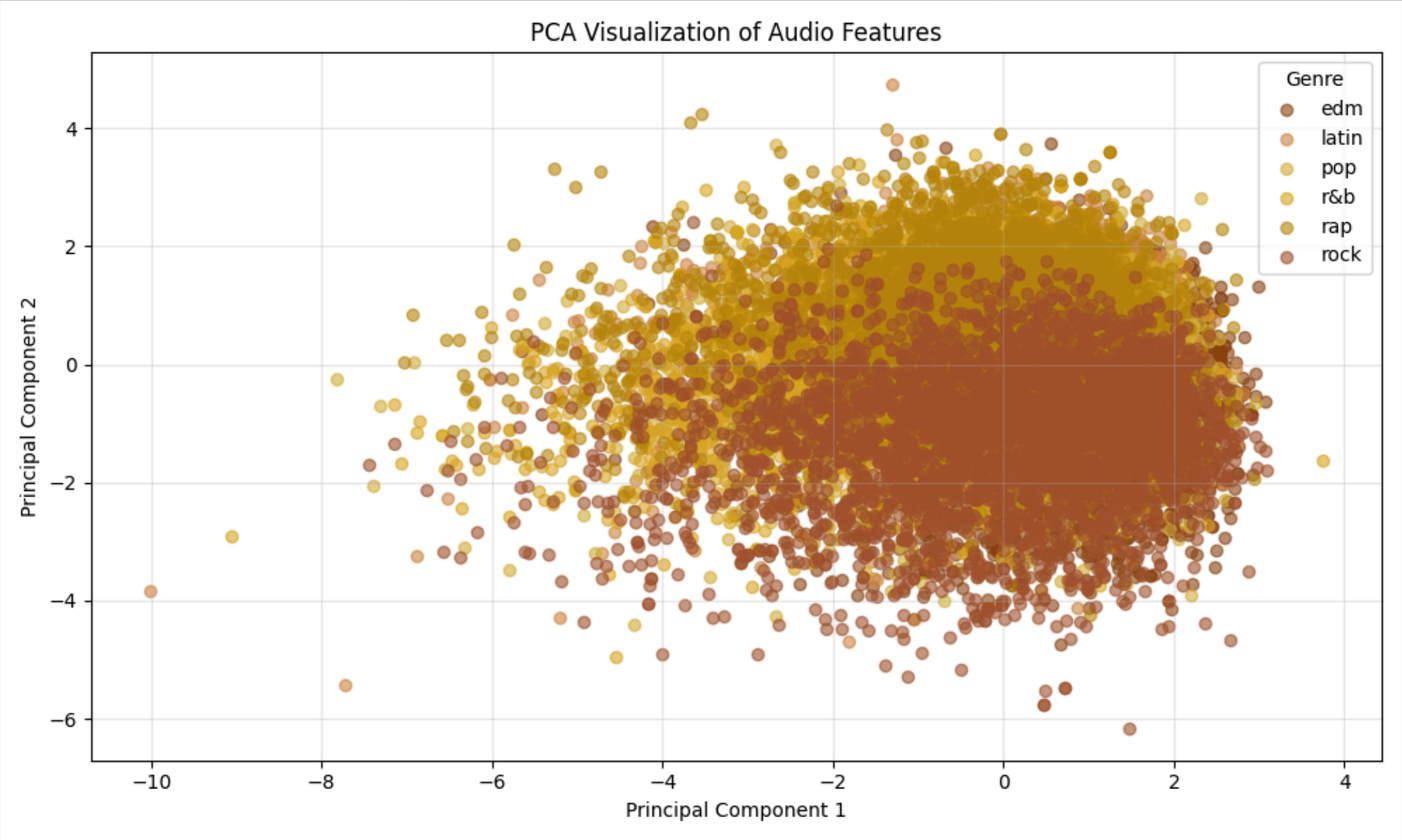
Challenges & Insights

Challenges:

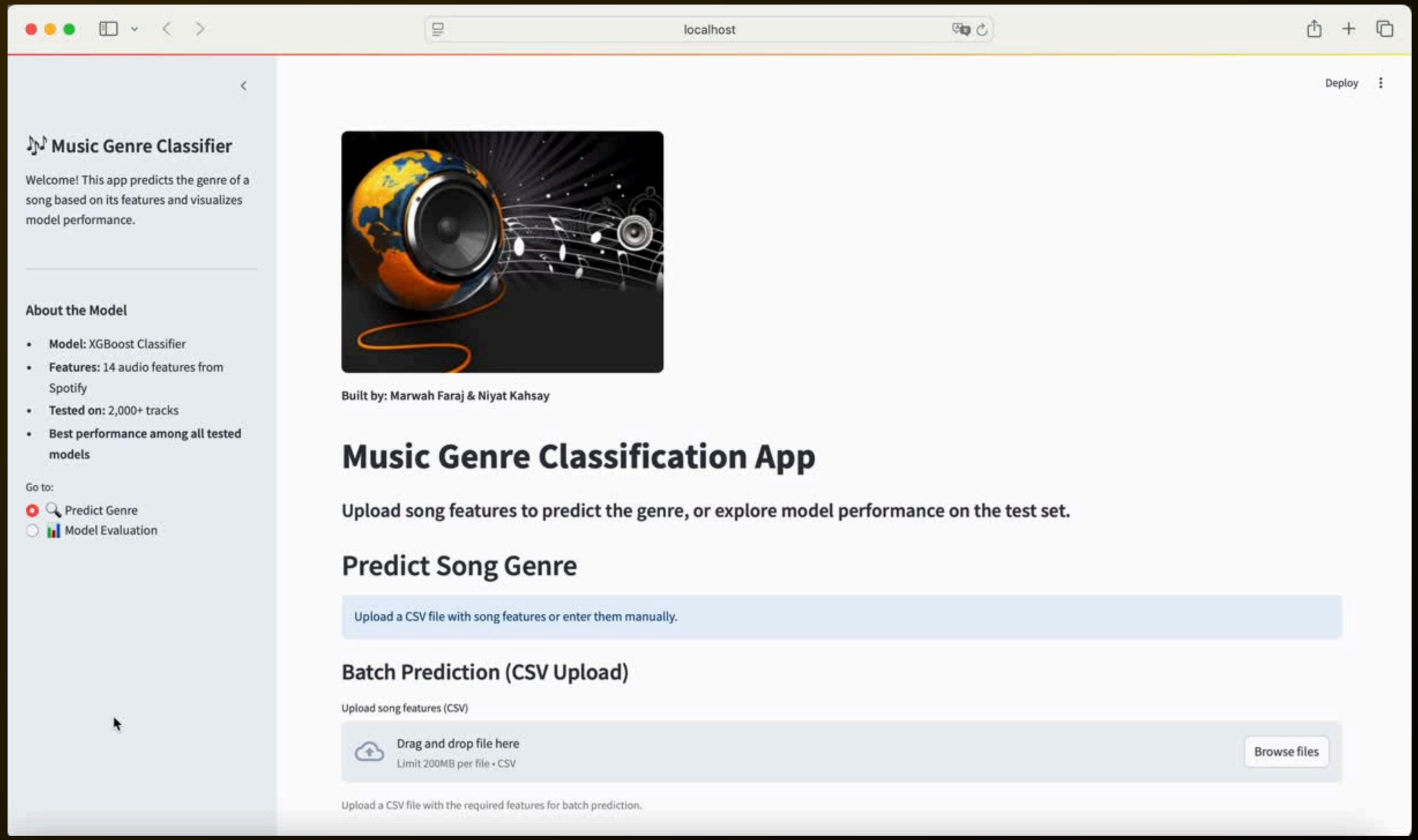
- Overlap between similar genres (Pop vs. Latin) reduced precision.
- PCA dimensionality reduction dropped accuracy to ~95% (loss of variance).
- Imbalanced feature influence (e.g., tempo skew).

Insights:

- Nonlinear tree-based models outperform linear SVM for this dataset.
- Feature engineering could further enhance minority class recall.

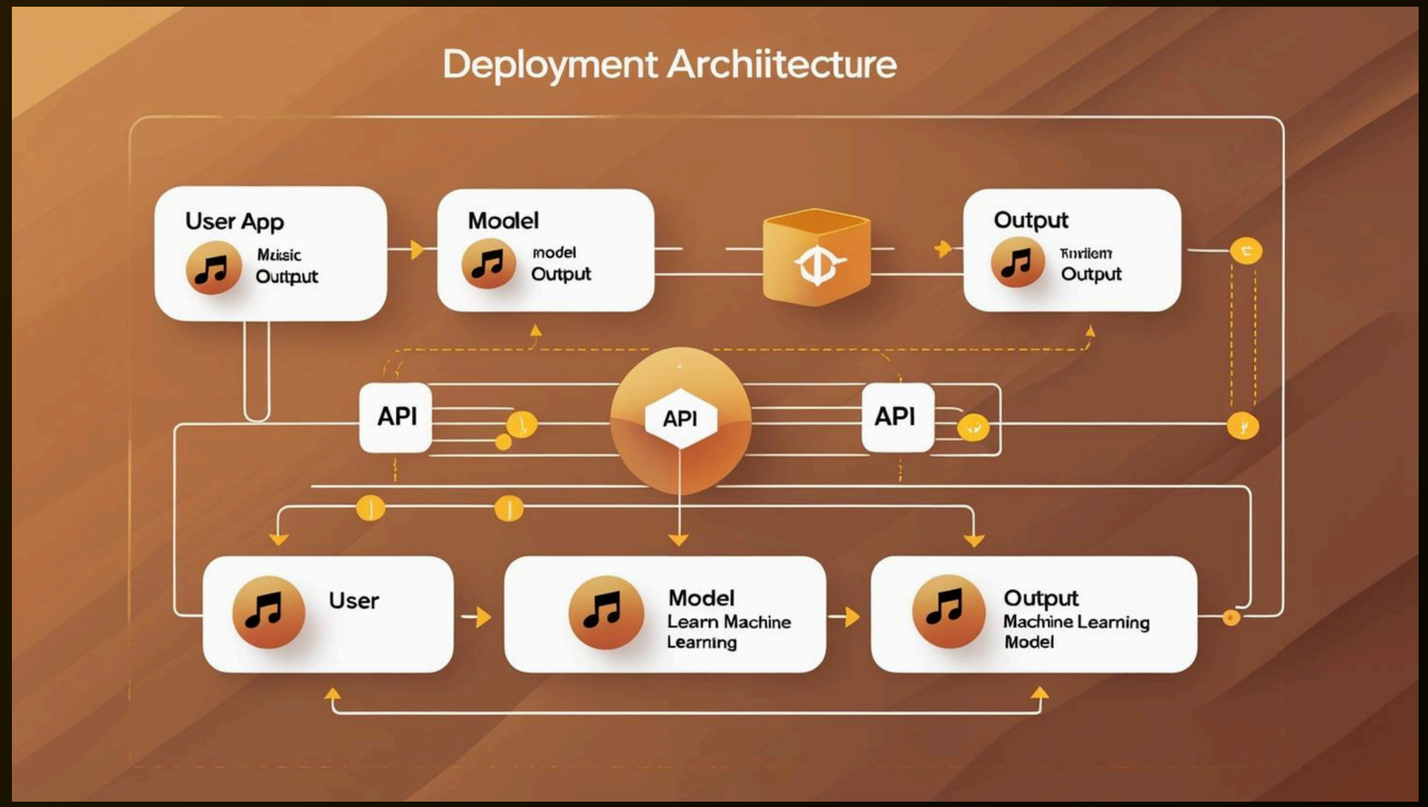


Model Deployment



Conclusion & Future Work

- XGBoost achieved 58% accuracy, outperforming all other models.
- Feature importance confirmed rhythmic & energetic attributes as strong predictors.
- Future improvements:
- Use spectrogram-based CNNs for raw audio input.
- Explore data augmentation for minority genres.
- Integrate deployment into a public-facing API.



Niat & Marwah

THANK YOU