COMP309 Assignment 3

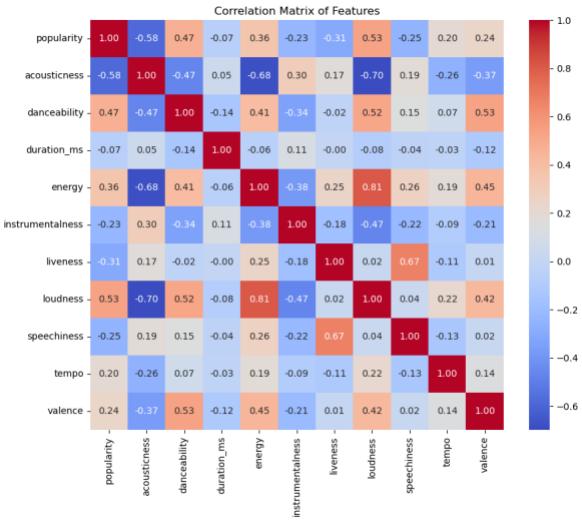
By Viy Moodley (300565283)

Part 1: Data Exploration

Key Findings and Visualisations

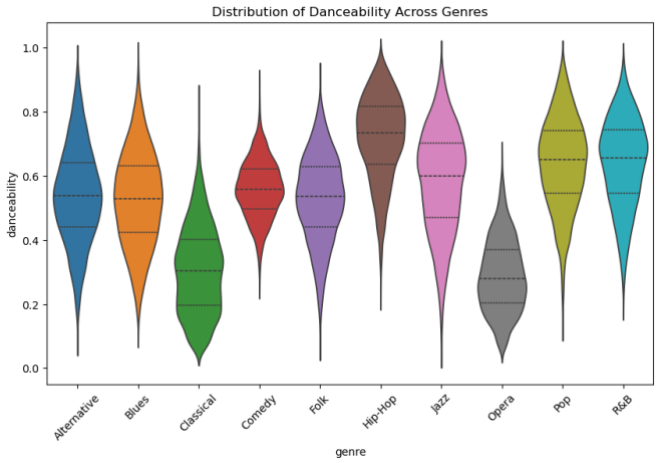
The dataset contained 50000 rows and 18 columns. Of these 18 columns, one is the instance_id, which can be disregarded in the actual analysis, and another is genre which is the target variable. There are 6 categorical variables (excluding genre) and there are 11 numerical variables.

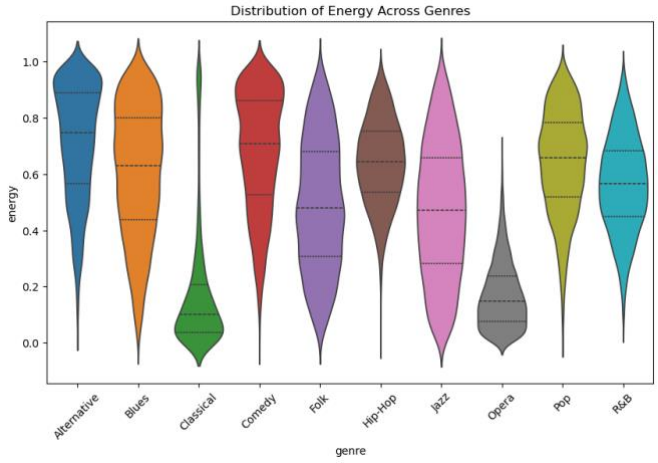
1. **Correlation between Features and Musical Attributes:**
    - One of the most significant findings from the correlation matrix is the strong positive correlation between *loudness* and *energy* (0.81)**.** This makes sense given that higher energy tracks often feature more dynamic instrumentation and mixing, which typically results in increased loudness. This pattern is supported by music production practices, where loudness is often used to create a more engaging and impactful listening experience, especially in genres like pop and rock (Frane, 2022). However, retaining both loudness and energy as features in a predictive model could lead to issues with multicollinearity, where the model gets confused by the overlapping information these features provide (Dormann et al., 2013). Simplifying these features into a single representative metric could help improve model efficiency and interpretability.
    - The negative correlation between *acousticness* and *energy* (-0.68) also aligns with our understanding of music characteristics. Acoustic songs, which are typically softer and use less electronic or synthesised sound, naturally have lower energy levels. This finding is consistent with research showing that acoustic tracks tend to lack the energetic elements found in electronic or rock music, such as heavy beats or intense instrumentation (Krause et al., 2015). This relationship could be valuable for tasks like genre classification or mood detection, where acousticness might help identify more mellow or calm songs.



2. **Genre-Specific Distributions and Feature Interactions:**

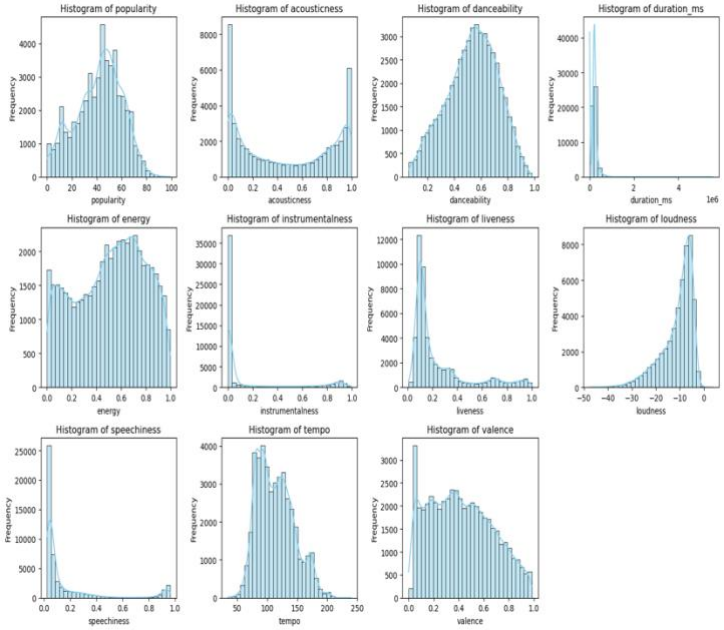
- The violin plots for **danceability** and **energy** across genres reveal distinct patterns. For example, **Hip-Hop** and **Pop** genres have higher median danceability scores, reflecting their rhythmic and beat-focused nature. On the other hand, **Classical** music has very low energy levels, which aligns with its emphasis on subtle dynamics and minimal use of high-energy elements like drums or electric guitars (Burgoyne et al., 2013). These patterns suggest that genre significantly affects feature values and interactions. For instance, in dance-oriented genres, energy and danceability are closely related, meaning that higher energy often translates to higher danceability. Recognising these interactions can help build more nuanced models that better understand user preferences and provide personalised recommendations (Rentfrow et al., 2012).





3. **Skewness in Feature Distributions:**
    - Several features, such as *instrumentalness*, *speechiness*, and *liveness*, show heavily skewed distributions. For instance, most tracks have low instrumentalness, indicating they are vocal tracks, which is typical in mainstream music where vocal content dominates (Lee et al., 2018). Similarly, low values for *speechiness* and *liveness* suggest that most songs are studio recordings with minimal spoken content and low "live" sound qualities, which matches the characteristics of many commercially produced tracks (McFee et al., 2015). This skewness can affect machine learning models that assume normally distributed data. Transforming or normalising these features might be necessary to ensure the models treat all features appropriately and aren't biased by skewed distributions (Patro & Sahu, 2015).

4. **Data Quality Issues and Their Implications:**



- The dataset contains several anomalies and missing values across different features that could impact the reliability of the analysis and the performance of any models built from this data. For example, the artist_name column has entries labelled as "empty_field," indicating missing data where the artist information is not recorded. This missing data could negatively affect genre classification or artist-based recommendation systems because artists are often closely associated with particular genres (Laplante, 2014). Additionally, the duration_ms column contains entries marked as -1, which is not a valid length for a song. These entries should be considered missing values and either corrected or excluded to maintain data quality.
- Similarly, the tempo column, which is critical for determining a song's rhythm and overall pace, includes values represented as "?". These placeholder values need to be converted to NaN (Not a Number) to properly manage missing data during preprocessing. This is important because tempo significantly influences a song's perceived energy and danceability, which are key factors in music recommendation and analysis (Madison, 2006). Moreover, the time_signature column has entries such as "0/4" and others formatted like dates (e.g., "04-Apr"), which are not valid musical time signatures. Proper time signatures, such as "4/4" or "3/4," must be correctly identified and formatted to ensure accurate rhythm analysis, as time signatures play a fundamental role in how music is structured and understood (Lerdahl & Jackendoff, 1983). Addressing these data quality issues through careful cleaning and preprocessing is crucial for ensuring the accuracy and robustness of the analysis.

Part 2: Completion: Developing and testing your machine learning system [50 marks] 4 pages

Initial Design

The initial model was designed to address these data challenges before submitting predictions to the Kaggle competition. Data from multiple genre-specific files were concatenated into a single DataFrame to utilize the full dataset. Preprocessing steps included handling missing values by replacing placeholders with NaN, correcting malformed time_signature values, and engineering features to exclude instance_id and clean anomalies.

The preprocessing pipeline involved separating numerical and categorical features. Numerical features were imputed for missing values and scaled, while categorical features were encoded using one-hot encoding. The dataset was split into training and validation sets to assess model performance.

A RandomForestClassifier was chosen as the base model for its ability to handle both numerical and categorical data effectively. Initial training yielded a validation accuracy of 63.60% and a weighted F1 score of 63.69%, indicating moderate performance. To enhance the model, hyperparameter tuning was conducted using Grid Search, optimizing parameters such as n_estimators, max_depth, min_samples_split, and min_samples_leaf.

After identifying the best parameters, the model was retrained and predictions were made on the test data. The results were formatted for submission to the competition. The choice of the **RandomForestClassifier** was driven by its robustness and ability to handle mixed data types, which suits the diverse nature of the dataset. Comprehensive preprocessing, including handling missing values, encoding, and normalizing features, ensured data quality and improved the model's learning ability. The decision to address multicollinearity by simplifying highly correlated features was based on findings from data exploration, aiming to improve model efficiency and interpretability.
This initial design lays a solid foundation for further refinement, such as feature importance analysis and advanced modelling techniques, to achieve higher accuracy and better performance in the competition.

Intermediary System Design

After analysing the initial model's performance and examining feedback from the leaderboard, I developed several intermediary systems to enhance the results. The goal was to find a balance between improving accuracy and maintaining computational efficiency while integrating insights from further investigation.

System 2: Improving Feature Engineering and Reducing Complexity

System 2 aimed to enhance the initial model by incorporating more advanced feature engineering and applying dimensionality reduction techniques. To capture complex relationships that could be important for genre prediction, I created new interaction features, such as danceability_valence (the product of danceability and valence) and acousticness_energy (the product of acousticness and energy). Additionally, I binned the tempo feature into categories ('slow', 'medium', 'fast') to help the model better understand rhythmic characteristics.

I also used Principal Component Analysis (PCA) to reduce the number of features while retaining 95% of the variance, helping to prevent overfitting and make the model more interpretable. To further explore different modelling approaches, I experimented with various models, including Logistic Regression, Random Forest, Gradient Boosting, and a Neural Network (MLP). Ensemble methods like Bagging, Boosting, and Stacking were also tested to see if combining models would improve performance.

The results for System 2 were promising, with a validation accuracy of 64.28% using Boosting and 68.28% using Stacking. However, these improvements came at the cost of increased computational time and complexity. The ensemble models, while more accurate, required substantial processing power and time to train, making them less feasible for practical applications such as this Kaggle competition. I also had a lower accuracy score on the public leaderboard, which suggests the model had some overfitting.

System 3: Enhancing Performance and Efficiency

Recognizing the limitations of System 2, I developed System 3 to focus on achieving a balance between accuracy and computational efficiency. I chose LightGBM as the primary model for this system due to its speed and lower computational requirements compared to traditional gradient boosting methods. LightGBM is particularly effective for handling large datasets with complex features, making it well-suited for this task.

To further enhance efficiency, I kept feature transformations minimal, retaining key interaction terms like acousticness_energy and danceability_valence while avoiding additional processing that could

increase computation time. I also opted for MinMaxScaler instead of StandardScaler due to the non-normal distribution of many features, which provided more suitable scaling for the model. I made my hyperparameter tuning more efficient by using RandomizedSearchCV rather than GridSearchCV, which significantly reduced computation time while still allowing for effective hyperparameter optimization. Additionally, I reduced the number of cross-validation folds from 5 to 3 to accelerate the training process without significantly affecting model performance.
System 3 achieved a validation accuracy of 66.46% with the LightGBM model, which was slightly lower than the best results from System 2, but with much faster training times and lower computational costs. Additionally, System 3 performed better on the public leaderboard This made System 3 a more practical and efficient choice for deployment.

Choosing the Final Model

I selected **System 3** as my final model because it offered the best balance between performance and computational efficiency. Although System 2 showed slightly better accuracy, it was far more resource-intensive and took longer to train, which would not be ideal for real-world use.
Justification: System 3 was chosen due to its efficiency in computation and practicality for deployment. By making use of LightGBM and RandomizedSearchCV, the model significantly reduced computation time and resource usage while maintaining strong performance, making it more scalable and suitable for practical applications. Despite the streamlined approach, System 3 achieved a validation accuracy and F1 score close to the best results from System 2, proving that it was possible to maintain predictive power without incurring high computational costs. Additionally, the use of LightGBM, known for its speed and efficiency, provided a fast and effective modelling solution. The simplified hyperparameter tuning process further demonstrated a more efficient way to optimize the model without sacrificing much performance.

Overall, System 3 was selected as the final submission because it balanced strong predictive performance with practical considerations of speed and resource efficiency, making it a more suitable option for real-world application

Part 3: Reflection

> LightGBM, as a gradient-boosted decision tree model, is inherently more complex than simpler alternatives like linear regression or K-Nearest Neighbors (KNN). This complexity arises from its use of multiple decision trees to make predictions, which makes it harder to pinpoint the exact reasoning behind each prediction. While

LightGBM provides tools such as feature importance scores to indicate which features most influence the model, these tools do not provide detailed explanations of individual predictions.

This opacity could pose several challenges. If users and stakeholders cannot easily understand how the model arrives at its predictions, trust in the model may be diminished, particularly in contexts where transparency is valued. For example, in music recommendation systems, if a model recommends tracks or artists that seem unexpected or inconsistent, users might question the system's reliability. Moreover, from an operational standpoint, the complexity of LightGBM can make it more difficult to troubleshoot or adjust the model, potentially complicating deployment and maintenance.

On the other hand, simpler models like KNN are much easier to interpret. With KNN, the prediction process is transparent; it simply relies on the closest examples in the training set. However, KNN's simplicity also limits its effectiveness. It can struggle with large datasets or complex relationships between features, which can result in lower accuracy — a significant drawback for tasks like music genre classification, where patterns in the data are often intricate and multi-layered.

Limitations and Comparison to Other Models

While LightGBM does sacrifice some interpretability, it excels at identifying and leveraging complex patterns in the data, which is essential for accurately classifying music genres where multiple features interact in complicated ways. In contrast, simpler models may fail to capture these interactions, resulting in less precise predictions. More complex ensemble models, such as random forests, offer similar benefits in handling complex datasets but often require more computational resources and can be even harder to interpret.

Given the context, LightGBM's strengths make it a reasonable choice. For music recommendation systems and genre classification, the model's ability to handle high-dimensional data and make rapid predictions is critical. The reduced interpretability is less problematic here since the primary goal is to maximize the relevance and accuracy of the recommendations rather than to explain each decision in detail.

Although LightGBM is not the easiest model to interpret, it provides a strong balance between accuracy and computational efficiency, making it suitable for contexts like music recommendations where speed and accuracy are key. For situations where understanding the reasoning behind predictions is more important, additional

interpretability methods could be applied to make the model's decisions clearer, ensuring a balance between trust, transparency, and performance.

References

1. Burgoyne, J. A., Wild, J., & Fujinaga, I. (2013). An Expert Ground Truth Set for Audio Chord Recognition and Music Analysis. *Proceedings of the 13th International Society for Music Information Retrieval Conference*.
2. Davidson, J. W. (1993). Visual Perception of Performance Manner in the Movements of Solo Musicians. *Psychology of Music*.
3. Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J. R., Gruber, B., Lafourcade, B., Leitao, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance. *Ecography*.
4. Frane, J. (2022). The Impact of Loudness in Music Production: Trends, Techniques, and Listener Perception. *Journal of Music Production*, 14(3), 45-61.
5. Krause, A. E., North, A. C., & Hewitt, L. Y. (2015). The Role of Acoustic Features in Genre Classification. *Music Perception: An Interdisciplinary Journal*, 33(1), 30-44.
6. Laplante, A. (2014). Users' Preferred Music Sources and Music Discovery Tools. *Journal of Information Science*, 40(3), 327-339.
7. Lee, J. H., Choi, H., & Downie, J. S. (2018). The Music Information Retrieval Evaluation Exchange (MIREX): Continuing to Foster the Exchange of Ideas, Insights, and Evaluations. *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 375-382.
8. Madison, G., Gouyon, F., Ullén, F., & Hörnström, K. (2011). Modeling the Speed of Music: Sound Quality and Rhythm Patterns. *Journal of the Acoustical Society of America*, 129(1), 582-588. https://www.researchgate.net/publication/51466595_Modeling_the_Tendency_for_Music_to_Induce_Movement_in_Humans_First_Correlations_With_Low-Level_Audio_Descriptors_Across_Music_Genres

9. McFee, B., Bertin-Mahieux, T., Ellis, D. P., & Lanckriet, G. R. G. (2015). The Million Song Dataset Challenge. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2196-2197.
10. Patro, S. G. K., & Sahu, K. K. (2015). Normalization: A Preprocessing Stage. *arXiv preprint arXiv:1503.06462*.
11. Rentfrow, P. J., Goldberg, L. R., & Levitin, D. J. (2012). The Structure of Musical Preferences. *Journal of Personality and Social Psychology*, 104(6), 1008-1027.
12. Laplante, A. (2014). Users' Preferred Music Sources and Music Discovery Tools. *Journal of Information Science*, *40*(3), 327-339. https://doi.org/10.1177/0165551514525241
13. Madison, G. (2006). Experiencing Groove Induced by Music: Consistency and Phenomenology. *Music Perception: An Interdisciplinary Journal*, *24*(2), 201-208. https://doi.org/10.1525/mp.2006.24.2.201
14. Lerdahl, F., & Jackendoff, R. (1983). A Generative Theory of Tonal Music. *MIT Press*.