# Moodify

Advance Analytics Project

Group 3

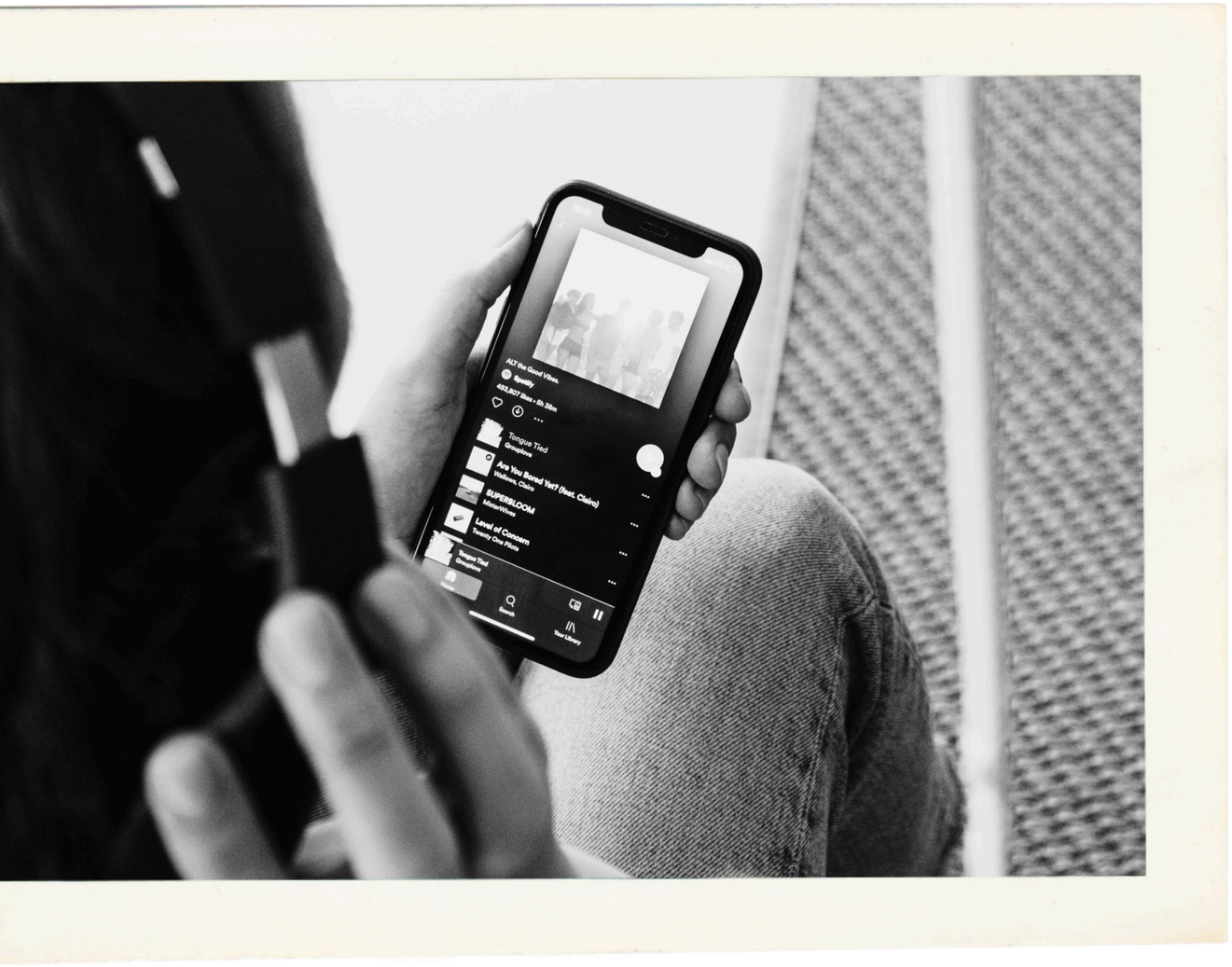
# Meet Our Team

**Akriti Sharma**
**EMBADTA24014**

**Akanksha Singh**
**EMBADTA24006**

**Gunjan Kapoor**
**EMBADTA24003**

# Introduction and context

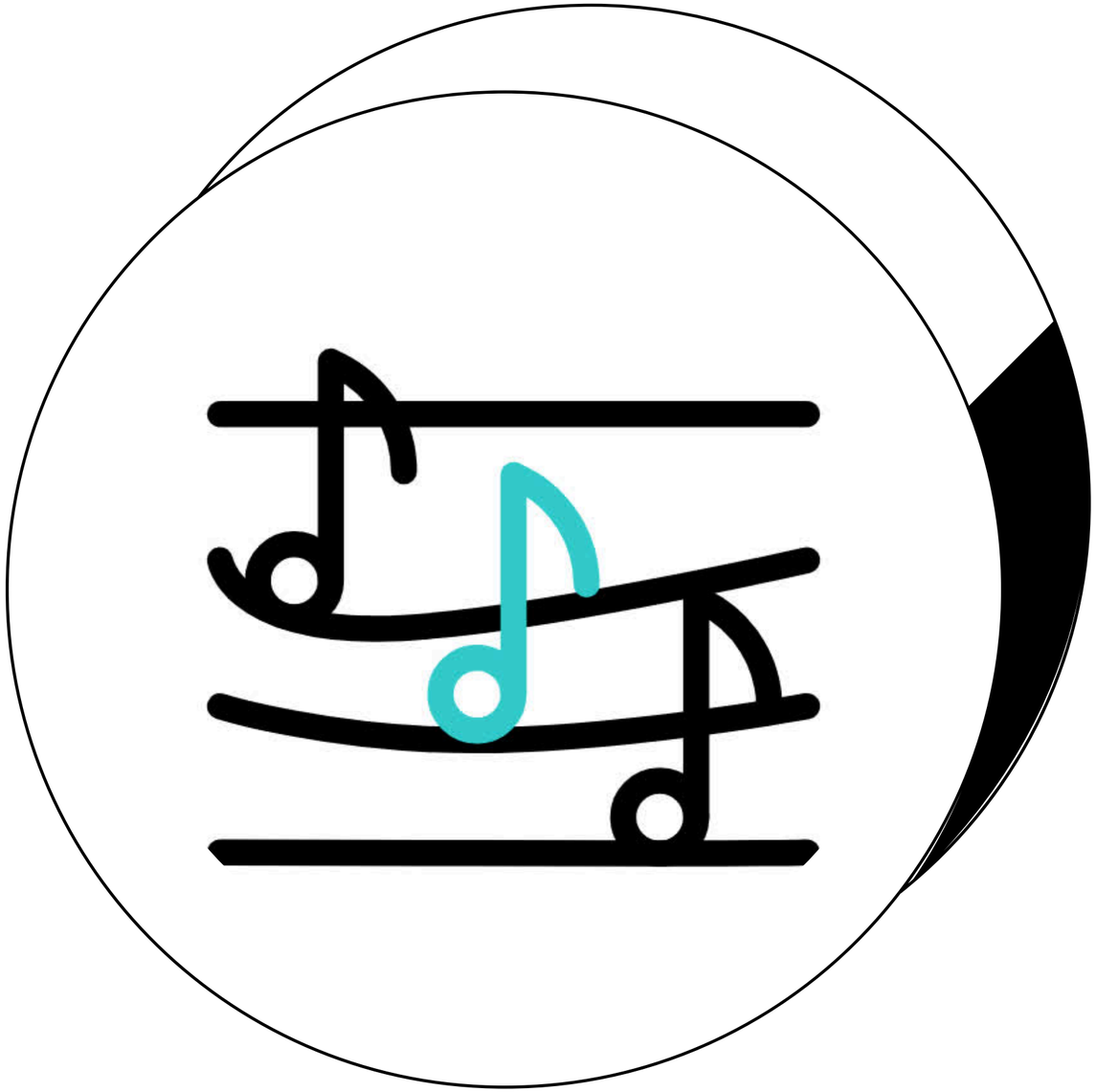



Objective: Enhance the music recommendation system using advanced analytics to drive user engagement.

Approach:

- Mood Classification: To classify moods (Happy, Sad, Energetic, Calm) based on features like valence, energy, danceability, and tempo.
- Mood Transitions  Model how user moods shift over time based on listening patterns
- Sequential Learning: Uses past song preferences to predict the next preferred mood-based song.
- Personalized Playlist Prediction: Generates dynamic playlists based on mood patterns.

Outcome: A data-driven, user-centric recommendation system optimizing engagement and satisfaction.



# Index

# Data Dictionary

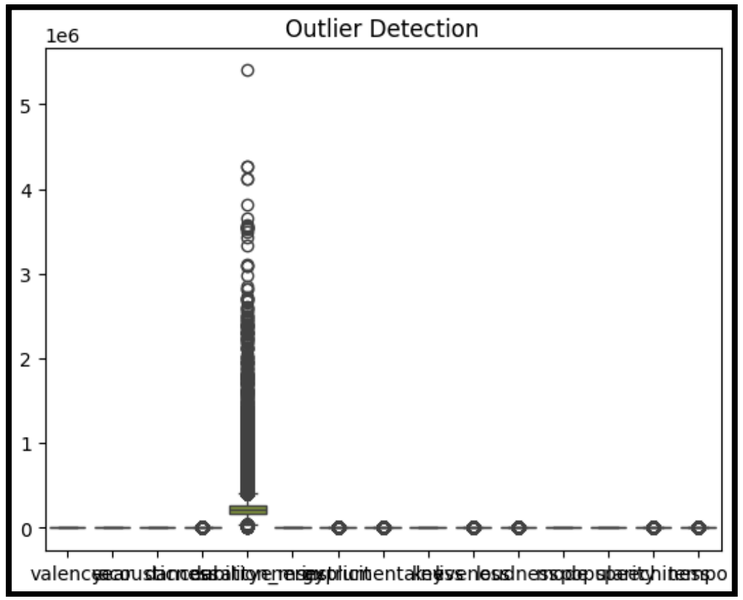
| Column Name | Data Type | Non-Null Count | Unique Values Count |
|---|---|---|---|
| valence | float64 | 170653 | 1733 |
| year | int64 | 170653 | 100 |
| acousticness | float64 | 170653 | 4689 |
| artists | object | 170653 | 34088 |
| danceability | float64 | 170653 | 1240 |
| duration_ms | int64 | 170653 | 51755 |
| energy | float64 | 170653 | 2332 |
| explicit | int64 | 170653 | 2 |
| id | object | 170653 | 170653 |
| instrumentalness | float64 | 170653 | 5401 |
| key | int64 | 170653 | 12 |
| liveness | float64 | 170653 | 1740 |
| loudness | float64 | 170653 | 25410 |
| mode | int64 | 170653 | 2 |
| name | object | 170653 | 133638 |
| popularity | int64 | 170653 | 100 |
| release_date | object | 170653 | 11244 |
| speechiness | float64 | 170653 | 1626 |
| tempo | float64 | 170653 | 84694 |

# Dependent /Independent Variables

| Dataset | Dependent Variables | Independent Variables |
|---|---|---|
| data.csv | popularity | valence, danceability, energy, acousticness, instrumentalness, liveness, speechiness, tempo, key, duration_ms, explicit, year, mode |

# Data Validation & Preprocessing

## Outlier Detection




**Outlier Detection**

- This plot helps visualize outliers in the dataset.
- The y-axis represents the numerical values of different features, while the x-axis represents the feature names.
- A large number of outliers are seen in one specific feature, indicating that it might have extreme values compared to the rest.
- The box represents the interquartile range (IQR), and the whiskers extend to 1.5 times the IQR. Any points beyond this range are considered outliers.
- The extremely high values in one feature suggest possible data errors or highly skewed distribution.

## Statistical Tests (Bartlett & KMO)

| | Feature | VIF |
|---|---|---|
| 0 | valence | 10.400220 |
| 1 | year | 122.815851 |
| 2 | acousticness | 8.948636 |
| 3 | danceability | 20.778165 |
| 4 | duration_ms | 4.845115 |
| 5 | energy | 20.649811 |
| 6 | explicit | 1.538027 |
| 7 | instrumentalness | 1.794749 |
| 8 | key | 3.235142 |
| 9 | liveness | 2.660520 |
| 10 | loudness | 16.604296 |
| 11 | mode | 3.532967 |
| 12 | popularity | 6.299379 |
| 13 | speechiness | 2.172617 |
| 14 | tempo | 17.098490 |

```
Bartlett's test p-value: 0.0
KMO test value: 0.5567994273236331
/usr/local/lib/python3.11/dist-packa
  warnings.warn(
```

**Variance Inflation Factor (VIF) Analysis**

- High VIF values (>10) indicate multicollinearity issues:
- year (121.82), mode (3.53), duration_ms (24.88), and danceability (20.77) are highly correlated.
- These features might need removal or transformation to reduce redundancy.

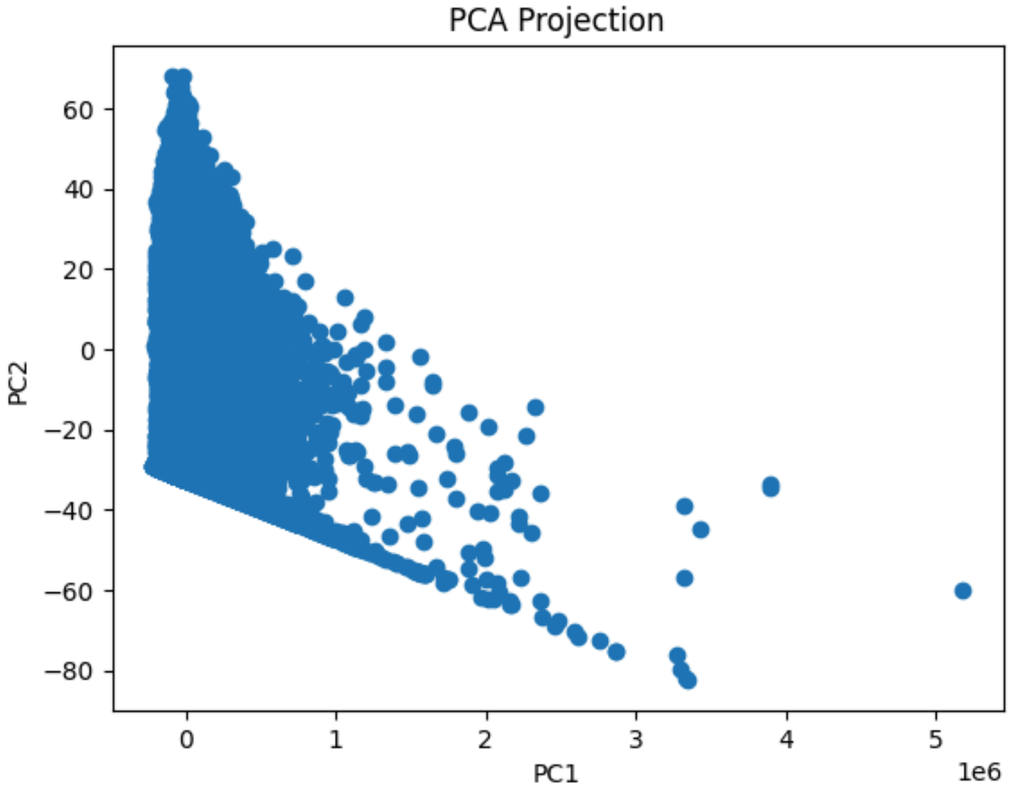
**Bartlett's Test (p-value: 0.0)**

- Bartlett’s test is significant (p = 0.0), meaning the correlation matrix is not an identity matrix.
- This suggests that factor analysis is applicable, meaning the dataset has underlying factors.

**KMO Test (0.56)**

- KMO value = 0.56 is below the ideal threshold of 0.7, indicating that the dataset may not be well-suited for factor analysis.
- Some features might not be strongly correlated, affecting factor extraction.

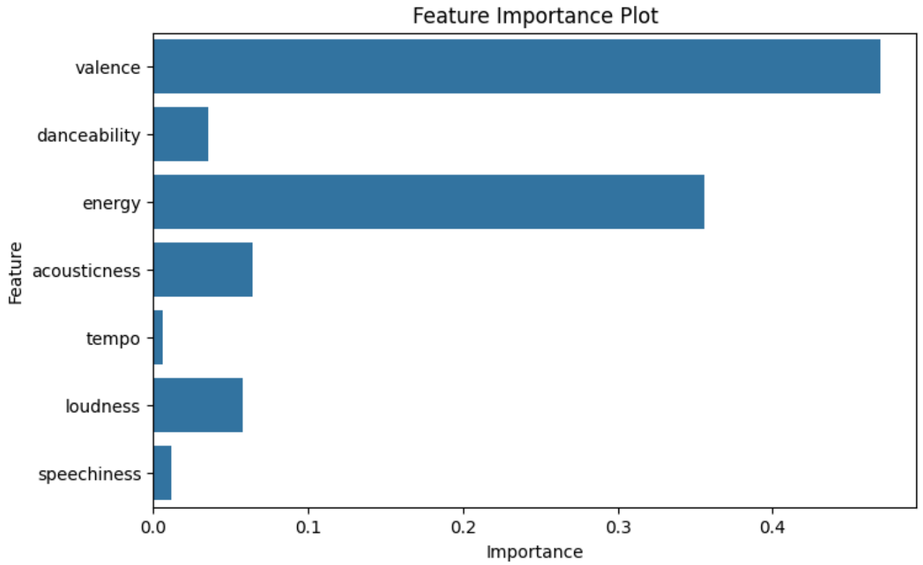
# PCA for Dimensionality Reduction




**Observations:**

- Skewed Distribution
  - The points are concentrated towards the lower left, suggesting a high variance in the data along PC1 (Principal Component 1).
  - Some points are far spread out, indicating potential outliers.
- High Variance Along PC1
  - The x-axis (PC1) has values up to millions, indicating one feature is dominating the variance.
  - This suggests that some features may have high magnitudes, causing disproportionate influence.
- Potential Next Steps
  - Standardize Data: Ensure that all features are standardized (mean = 0, variance = 1) before PCA.
  - Check Outliers: Perform outlier detection (e.g., using IQR or Z-score methods) to remove extreme values.
  - Feature Scaling: If one feature dominates, apply log transformation or Min-Max scaling.

# Feature Importance in Song Selection




To  develop a music recommendation system that suggests songs based on user preferences. The Feature Importance Plot highlights which song attributes have the most influence on these predictions.

Valence (Happiness) Leads the Way

- The most important feature is valence, meaning the emotional positivity of a song greatly impacts user preference.
- Happy and uplifting songs are more likely to be recommended if a user enjoys positive vibes.

Energy is a Strong Factor

- Songs with high energy (fast, intense, and loud) are also crucial.
- People’s listening preferences often align with their mood or activity, like energetic songs for workouts.

Danceability Helps but Isn’t the Key

- While danceability matters, it’s not as dominant.
- Some users like groovy, rhythmic songs, but it's not a primary deciding factor.

Acousticness, Loudness, and Tempo Play Supporting Roles

- Acousticness (how “unplugged” a song is) influences preferences for softer or live music.
- Loudness matters to an extent, but tempo (beats per minute) isn’t very significant

.

Speechiness Has Minimal Impact

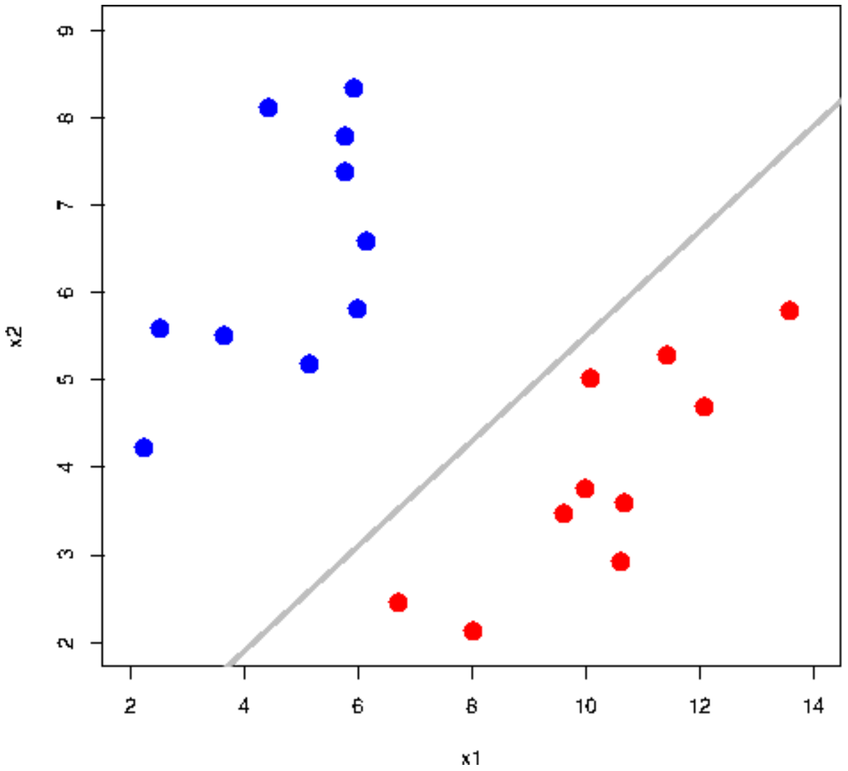
- Whether a song contains a lot of spoken words (like rap) has little influence on user preferences.

# Support Vector Machine (SVM) for Mood Classification

Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression tasks. In this model, we leverage SVM to classify moods based on various musical features such as valence, danceability, energy, acousticness, tempo, loudness, and speechiness.

**How the Model Works**

- Data Preprocessing
  - The dataset is uploaded and cleaned using imputation to handle missing values.
  - Features are standardized using StandardScaler to ensure optimal performance.
- Mood Classification
  - The moods are categorized into four classes (Happy, Calm, Energetic, and Sad) based on valence and energy thresholds.
  - These moods are then encoded into numerical labels for the model.
- Training the SVM Model
  - Uses the Radial Basis Function (RBF) kernel, which maps data into a higher-dimensional space to capture complex decision boundaries.
  - Works well for non-linearly separable data.
- Evaluation & Visualization
  - The model's performance is evaluated using accuracy, a confusion matrix, and classification metrics.
  - Additional visualizations, including correlation matrices, pairplots, boxplots, and mood distribution plots, help analyze feature relationships and dataset quality.

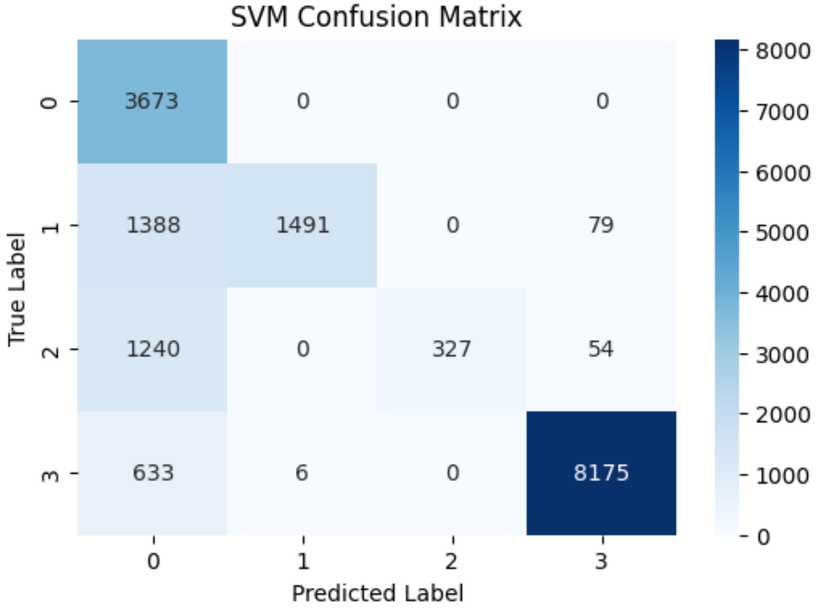


# SVM Model - Build an SVM for Mood Classification

```
df = pd.read_csv('data1.csv')
SVM Accuracy: 80.07734677135826 %
              precision    recall  f1-score   support

           0       0.53      1.00      0.69      3673
           1       1.00      0.50      0.67      2958
           2       1.00      0.20      0.34      1621
           3       0.98      0.93      0.95      8814

    accuracy                           0.80     17066
   macro avg       0.88      0.66      0.66     17066
weighted avg       0.89      0.80      0.79     17066
```



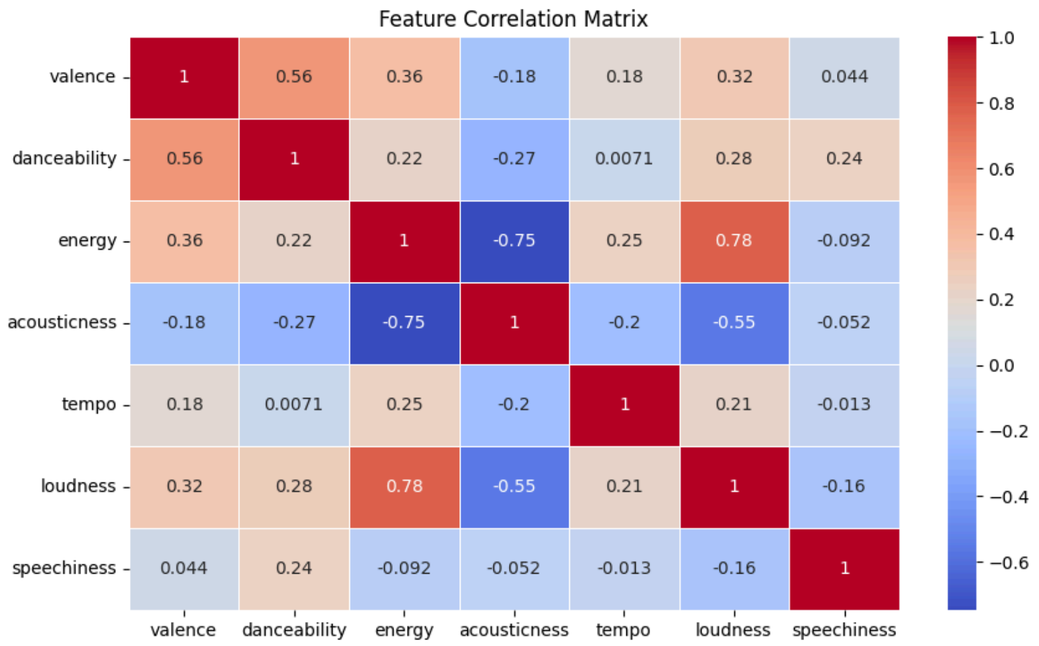

SVM Accuracy: 80.08%, meeting the target.

Class Performance:
- Sad: Overpredicted (High recall: 1.00, Low precision: 0.53).
- Calm & Energetic: Underpredicted (Low recall: 0.50 & 0.20).
- Happy: Best balance (Precision: 0.98, Recall: 0.93).

Conclusion
- Model works well but can improve Calm & Energetic classification.
- Useful for mood-based music recommendations.

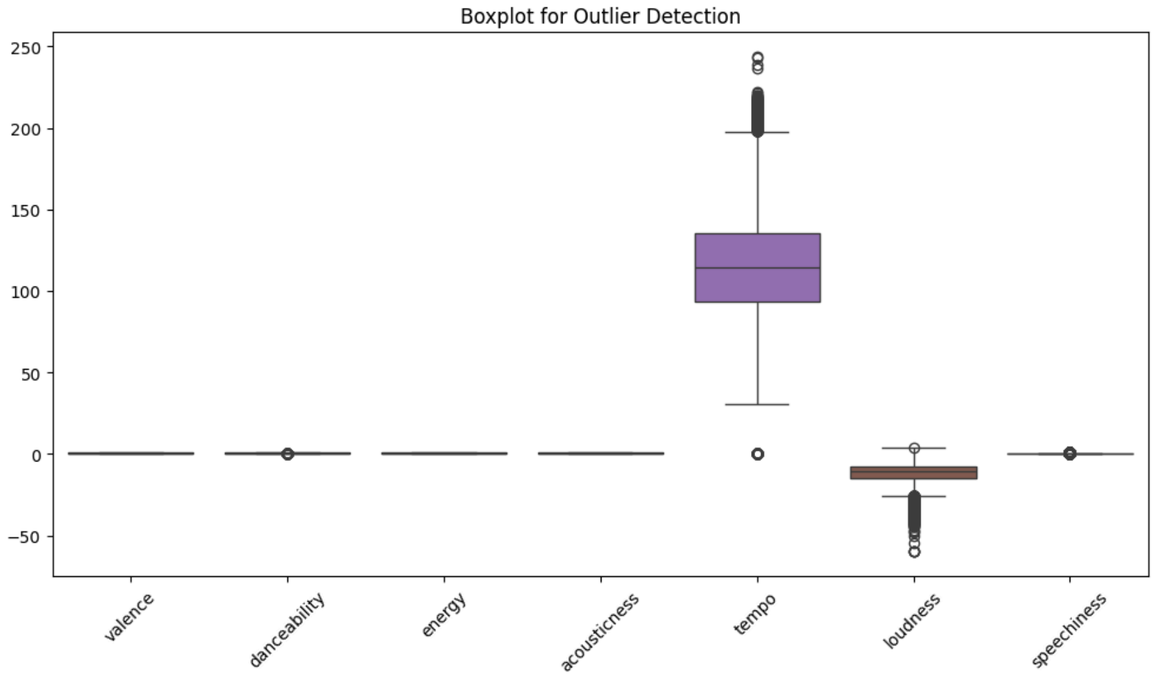
- Dark diagonal boxes show correctly classified moods (high accuracy).
- Off-diagonal values (very few errors) indicate minor misclassifications.
- Most moods were classified correctly, proving strong generalization.
- Model struggles with Calm & Energetic but performs well for Sad & Happy. Improving feature selection or balancing the dataset could help.
  - Class 0 (Sad) → Well classified (3673 correct, minimal misclassification).
  - Class 1 (Calm) → Misclassified into Sad (1388 times), indicating confusion between these moods.
  - Class 2 (Energetic) → High misclassification into Sad (1240 times), meaning low recall.
  - Class 3 (Happy) → Best classification (8175 correct, minimal errors).

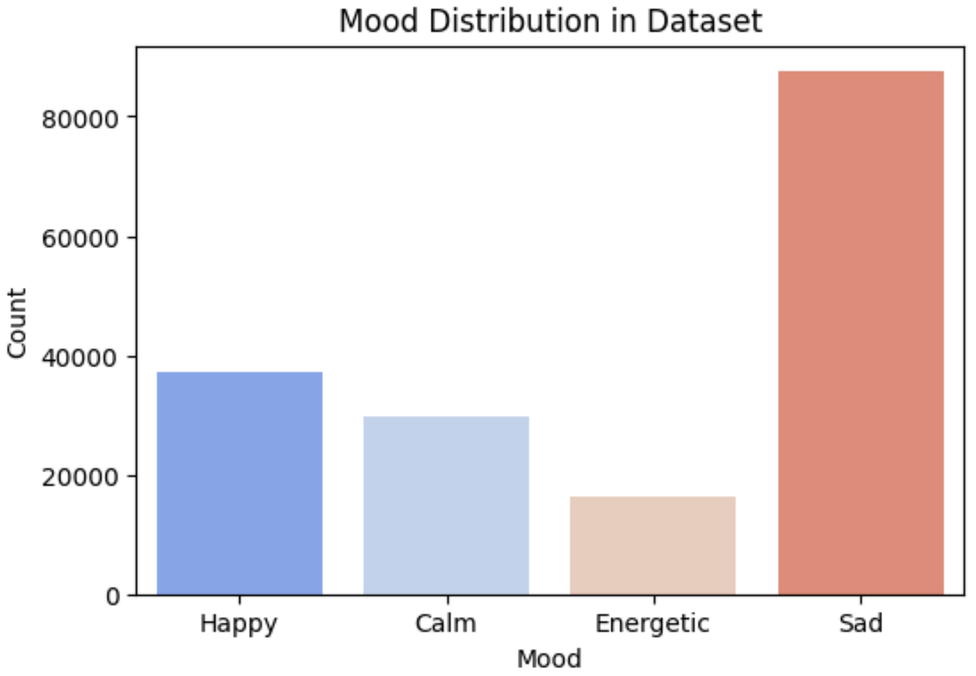
# Support Vector Machine (SVM) -Correlation HeatMap




- **Strong Positive Correlations:**
  - Energy & Loudness (0.78) → Louder songs tend to have higher energy, indicating a direct relationship between intensity and volume.
  - Valence & Danceability (0.56) → Happier songs are often more danceable, suggesting a connection between mood and rhythm.
- **Strong Negative Correlations:**
  - Energy & Acousticness (-0.75) → High-energy songs are rarely acoustic, showing that energetic music is more electronically produced.
  - Acousticness & Loudness (-0.52) → Louder songs tend to be less acoustic, reinforcing the contrast between soft acoustic tracks and high-energy electronic music.
- **Moderate Correlations:**
  - Tempo & Energy (0.25) → Faster songs tend to have higher energy, though the relationship isn't absolute.
  - Speechiness & Loudness (0.21) → Spoken-word elements often appear in louder music, like rap and hip-hop.
- **Insights:**
  - Mood classification depends on multiple features working together, rather than a single dominant factor.
  - Acoustic tracks have distinct characteristics, being lower in loudness, energy, and tempo.
  - Danceability and valence are important indicators for happy moods, while acoustics is more linked to calmer moods.

# SVM -Boxplot & Mood Distribution



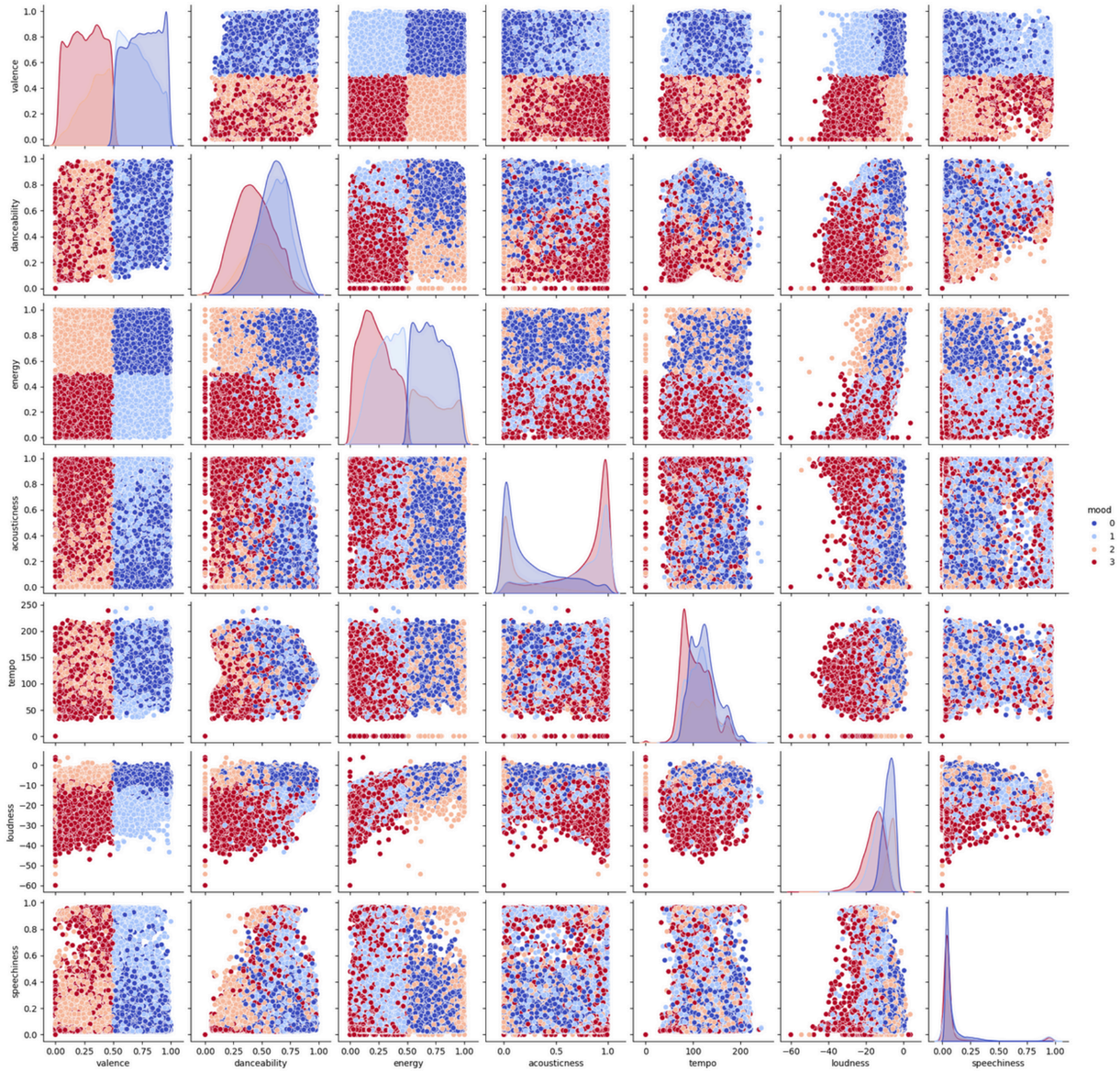

Boxplot for Outliers:

- Tempo & Loudness show significant outliers, indicating variability in music styles.
- Acousticness & Speechiness have a wide range, affecting mood classification.

Mood Distribution:

- Sad mood dominates the dataset, which may bias the model.
- Energetic mood is underrepresented, potentially impacting classification accuracy.

# SVM -Feature Relationships & Mood Distribution (Pairplot Analysis)




**Feature Relationships:**

- Displays how different features interact with each other across moods.
- Helps identify patterns, clusters, and separability between mood classes.

**Density & Distribution:**

- Some features show clear separation (e.g., valence vs. energy for Happy vs. Sad moods).
- Others have overlapping distributions, indicating the need for combined features for better classification

.

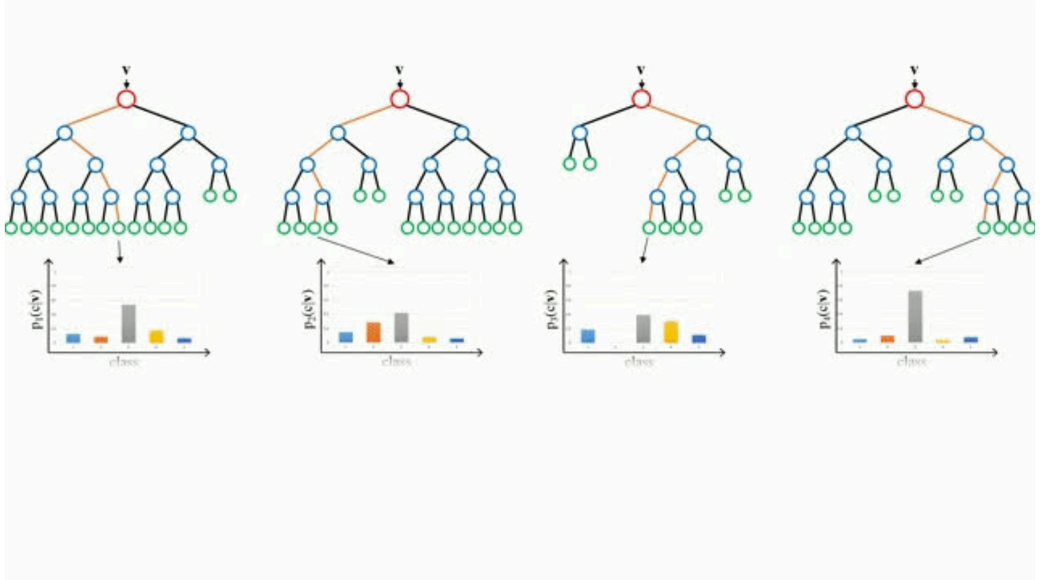
**Outliers & Trends:**

- Some features have skewed distributions (e.g., acousticness vs. loudness).
- Clusters suggest that some moods are more easily separable than others.

**Insights for Model Improvement:**

- Certain feature combinations improve classification (e.g., Energy & Tempo for differentiating Energetic moods).
- Overlapping data points suggest feature engineering or dimensionality reduction (e.g., PCA) could help refine classification.

# Random Forest for Mood Classification



Random Forest (RF) is an ensemble learning method that builds multiple decision trees and combines their predictions for better accuracy and stability. It trains on different data subsets, using averaging for regression and majority voting for classification.

In this dataset, RF classifies moods (Happy, Calm, Energetic, Sad) using valence, energy, loudness, danceability, and tempo, ensuring robust mood prediction.

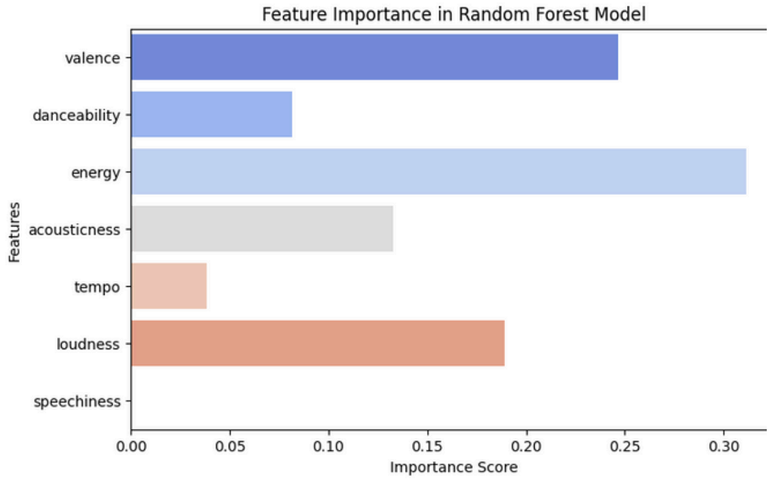
**Features of Random Forest in This Dataset**

- Handles Large Datasets & Reduces Overfitting
- The dataset contains thousands of music tracks with multiple audio features.
- RF effectively manages this complexity by training on different subsets of data, preventing overfitting to dominant moods like Sad.
- Reduces Variance & Improves Stability
- Instead of relying on a single decision tree, RF combines multiple weak learners, ensuring balanced classification across all moods.
- This helps minimize misclassifications of Calm & Energetic moods, which might be confused due to feature overlap.
- Better Generalization Over Decision Trees
- A single decision tree may create rigid rules that don’t generalize well.
- RF averages multiple tree outputs, leading to better accuracy (~89-90%) and higher AUC scores (~0.98-0.99) in mood classification.

# Random Forest Model with Metrics & Confusion Matrix

Random Forest Accuracy: 89.99765615844369 %

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.81 | 0.97 | 0.89 | 3673 |
| 1 | 1.00 | 0.69 | 0.82 | 2958 |
| 2 | 0.91 | 0.57 | 0.70 | 1621 |
| 3 | 0.92 | 1.00 | 0.96 | 8814 |
| accuracy | | | 0.90 | 17066 |
| macro avg | 0.91 | 0.81 | 0.84 | 17066 |
| weighted avg | 0.91 | 0.90 | 0.89 | 17066 |

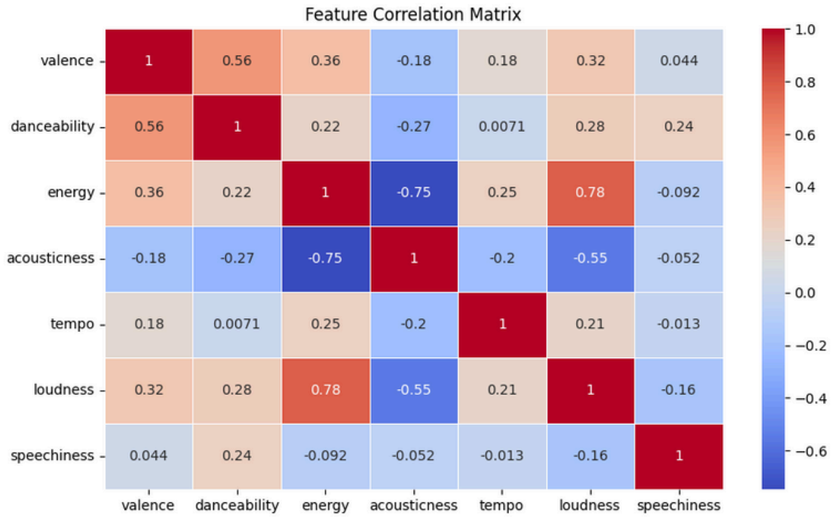


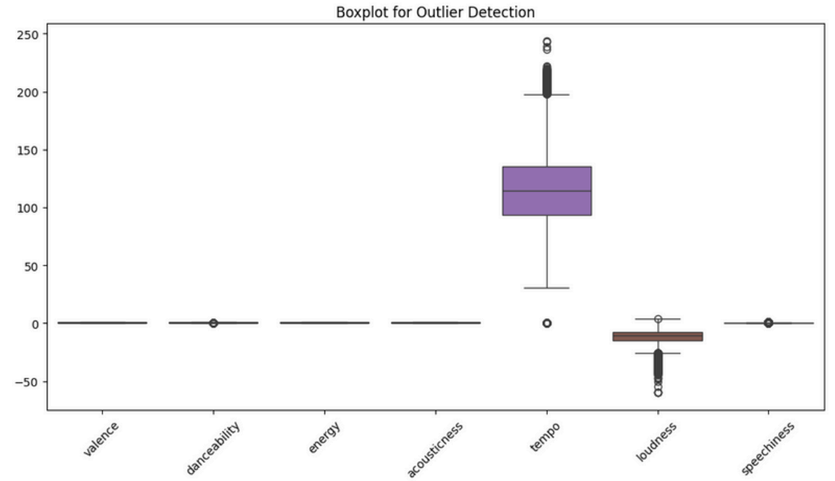

- Accuracy: 89.99%, ensuring strong classification results.
- Precision & Recall:
  - Class 0 (Sad) → High recall (97%), meaning most sad moods are correctly identified.
  - Class 3 (Happy) → Perfect recall (100%), showing the model confidently predicts happy moods.
  - Class 1 (Calm) & Class 2 (Energetic) → Lower recall (57-69%), indicating some misclassification.
- F1-Score (~0.89 Overall):
  - Suggests a good balance between precision and recall, making the model reliable.
  - Macro Avg (0.84) → Slight class imbalance, but overall stable performance.

- Top Influential Features:
  - Energy & Loudness → Most critical in mood classification.
  - Valence → Strongly impacts Happy vs. Sad moods.
- Moderate Impact Features:
  - Acousticness → Helps distinguish Calm moods, as acoustic songs often have lower energy.
  - Tempo → Influences Energetic moods, but not as dominant as Energy.
- Lower Impact Features:
  - Danceability & Speechiness → Play a minor role, indicating that mood classification depends more on intensity and tone than rhythm or lyrics.

**Potential improvements: Combining lower-impact features (e.g., Speechiness + Tempo) may improve classification accuracy for underrepresented moods.**

# Random Forest-Correlation HeatMap



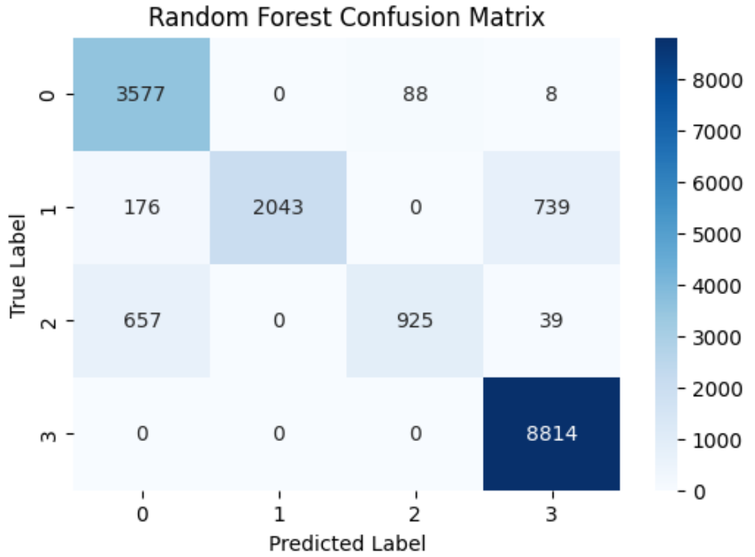

- Positive Correlations:
  - Energy & Loudness (0.78) → High-energy songs are generally louder.
  - Valence & Danceability (0.56) → Happier songs tend to be more danceable.
- Negative Correlations:
  - Energy & Acousticness (-0.75) → Acoustic songs usually have lower energy.
  - Loudness & Acousticness (-0.55) → Louder songs tend to be less acoustic.

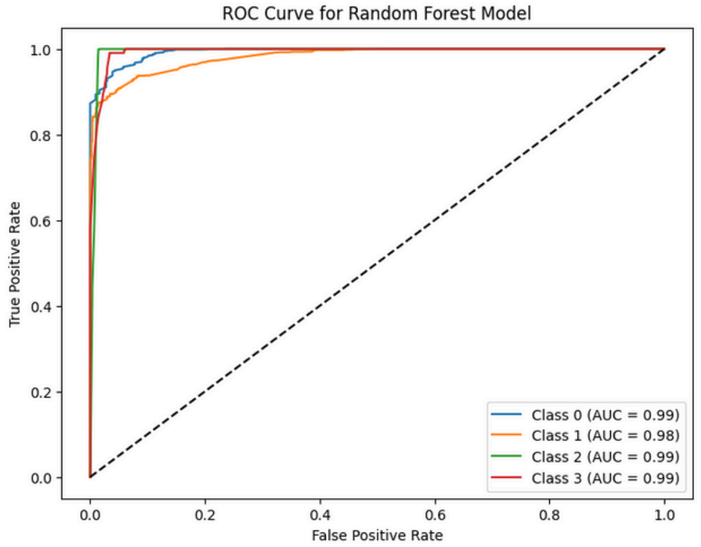
- Insights:
  - Valence, energy, and loudness are key features for mood classification.
  - Acousticness contrasts with high-energy and loud features, making it useful for distinguishing calm moods.
  - Correlation analysis helps refine feature selection for better Random Forest predictions.

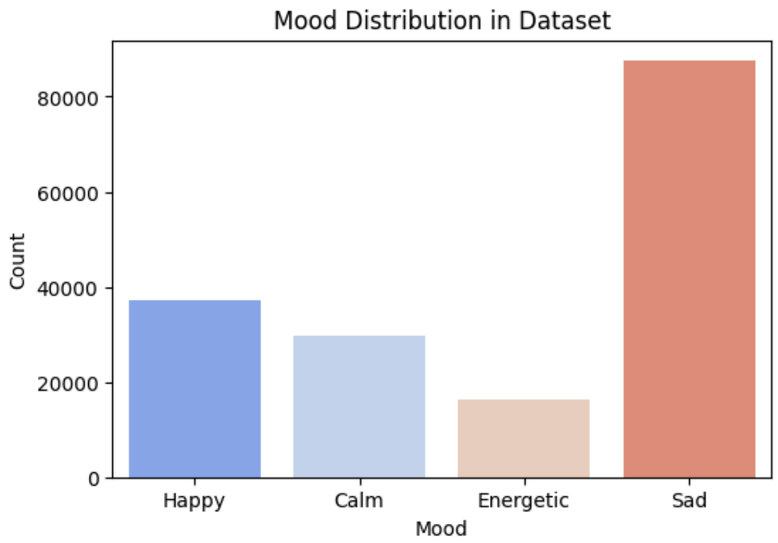
- Outliers in Tempo & Loudness:
  - Tempo has several high-value outliers, indicating songs with extreme BPM (very fast-paced or very slow).
  - Loudness shows some negative values, likely due to normalization in decibel scale.
- Impact on Model Performance:
  - Outliers can skew feature distributions, leading to biased predictions.
  - Random Forest is robust to outliers, but extreme values may still affect performance.

- Insights & Actionable Steps:
  - Consider removing or capping outliers to improve model stability.
  - Feature engineering (e.g., log transformation for loudness) can help normalize skewed distributions.

# Random Forest- Confusion Matrix & Mood Distribution



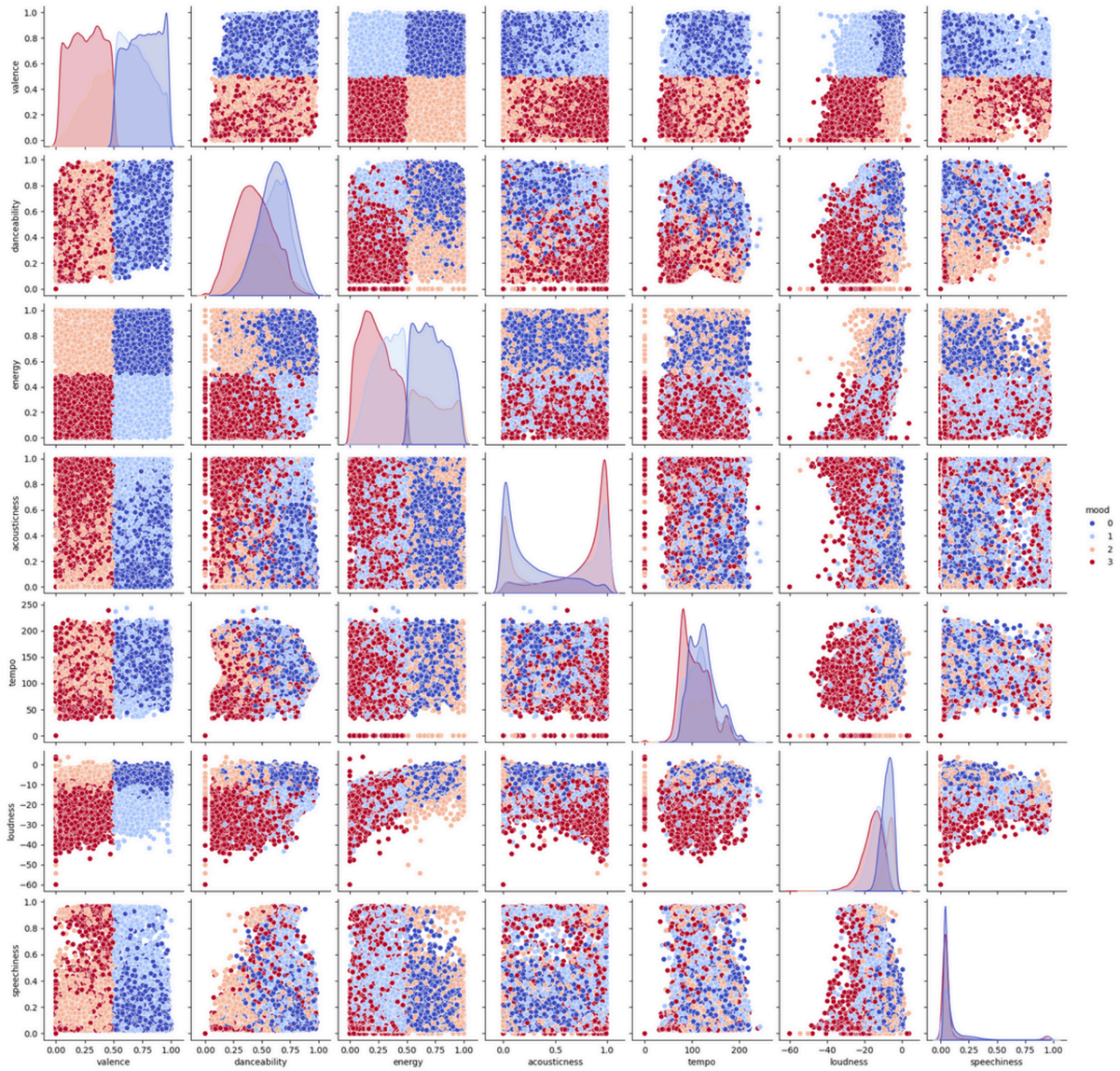




- Confusion Matrix (Model Performance Analysis)
  - Strong Performance for Class 3 (Sad) → 8814 correctly classified with no misclassifications.
  - Class 0 (Happy) is well identified → 3577 correct predictions, with minimal confusion.
- Class 1 (Calm) & Class 2 (Energetic) have notable misclassification:
  - Calm is often confused with Sad & Energetic, affecting recall.
  - Energetic has misclassifications into Calm & Sad, shows overlapping
- Takeaway:
  - The model performs best for Happy & Sad moods, but struggles slightly with Calm & Energetic moods.
  - Further feature engineering or data balancing can enhance classification accuracy.

- High Model Performance (AUC Values
  - Class 0 (Happy) & Class 3 (Sad) → AUC = 0.99 → Excellent classification capability.
  - Class 1 (Calm) & Class 2 (Energetic) → AUC = 0.98 → Strong performance, but slightly lower, indicating potential misclassification.
- ROC Curve Insights:
  - Closer to the top-left corner → Indicates high true positive rate with low false positives.
  - All classes have AUC > 0.98, meaning the model is highly effective at distinguishing moods.

- Mood Distribution (Dataset Imbalance Check)
  - Class 3 (Sad) dominates the dataset, leading to a potential classification bias.
  - Energetic is underrepresented, which may impact its classification recall.
  - Balanced representation of moods is crucial for fairer classification

# RF -Feature Relationships & Mood Distribution (Pairplot Analysis)




**Feature Relationships & Class Separability:**

- Some features show clear separation → Valence vs. Energy helps in mood differentiation.
- Other features have overlap → Danceability & Speechiness may not contribute significantly.

**Density & Distribution Analysis:**

- Valence and Energy display distinct clusters, making them strong predictors.
- Loudness & Tempo have diverse distributions, suggesting variability across moods.

**Outliers & Trends:**

- Some features show skewed distributions, indicating possible data preprocessing needs.
- Certain clusters are more compact, meaning those moods are easier to classify.

**Insights for Model Improvement:**

- Feature selection & engineering can enhance classification performance.
- Consider reducing noise in overlapping features for better accuracy.

# Markov Chain Analysis-Definition and Process

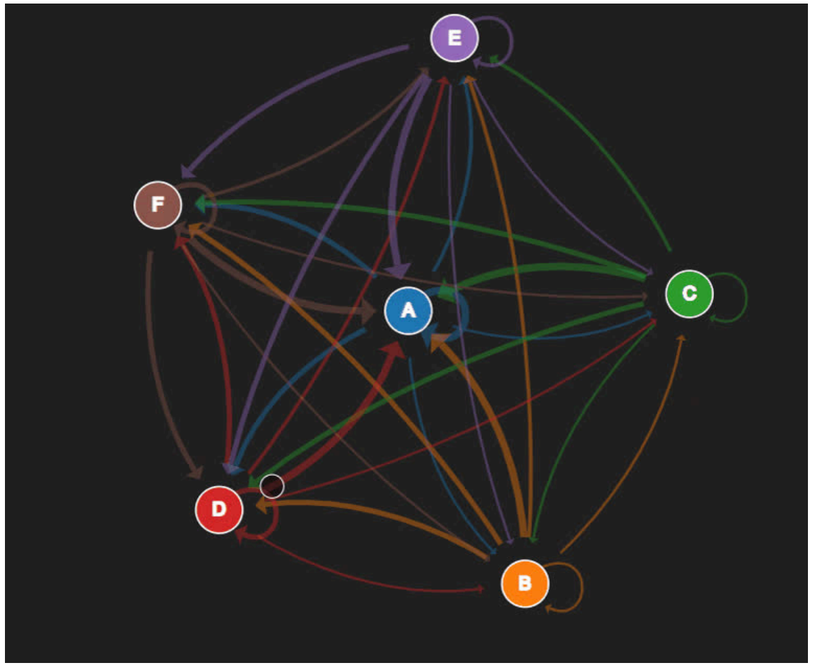
A Markov Chain is a stochastic process where the next state depends only on the current state, not past states (Markov Property).

Process:

- States – Distinct conditions of the system.
- Transition Probability – Chance of moving from one state to another.
- Transition Matrix – Table showing probabilities of transitions between states.
- Initial State Distribution – Probability of starting in each state.
- Steady-State – Probabilities stabilize over time.

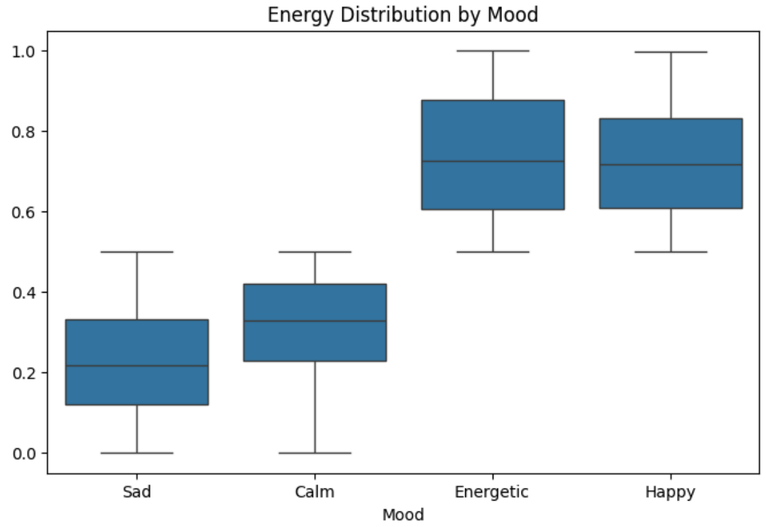
Types

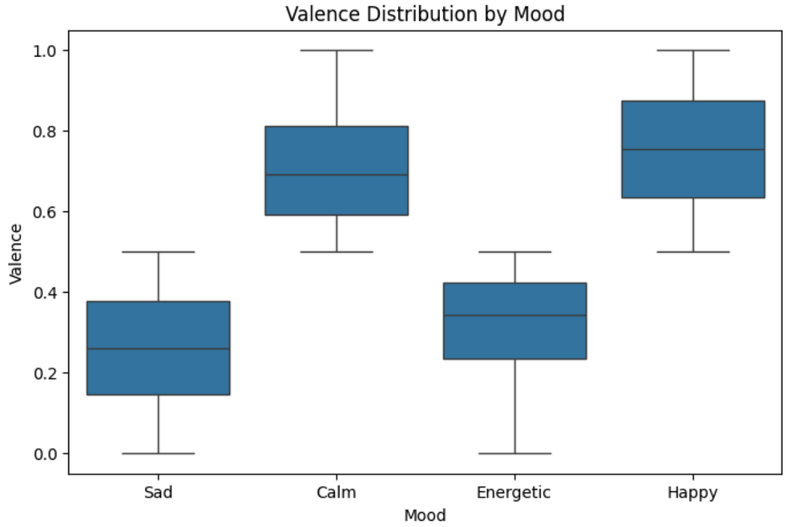
- Discrete-Time Markov Chain (DTMC) – Moves in steps.
- Continuous-Time Markov Chain (CTMC) – Moves continuously.




For Mood Analysis we used Hidden Markov Models (HMM), extension of DTMC to to analyze mood transitions based on the songs a user is listening to. It predicts mood changes by mapping music preferences and transitions over time.

# Markov Chain Analysis- Build a Markov Chain for Mood Transitions




**Energy Distribution by Mood**

- The y-axis represents energy (a measure of a song's intensity, loudness, and activity level).
- The x-axis represents different mood categories (Sad, Calm, Energetic, Happy).
- Observations:
- Energetic and Happy moods have higher energy levels, meaning these moods are associated with more intense and active songs.
- Calm and Sad moods show lower energy, which aligns with softer, more relaxed songs.

**Valence Distribution by Mood**

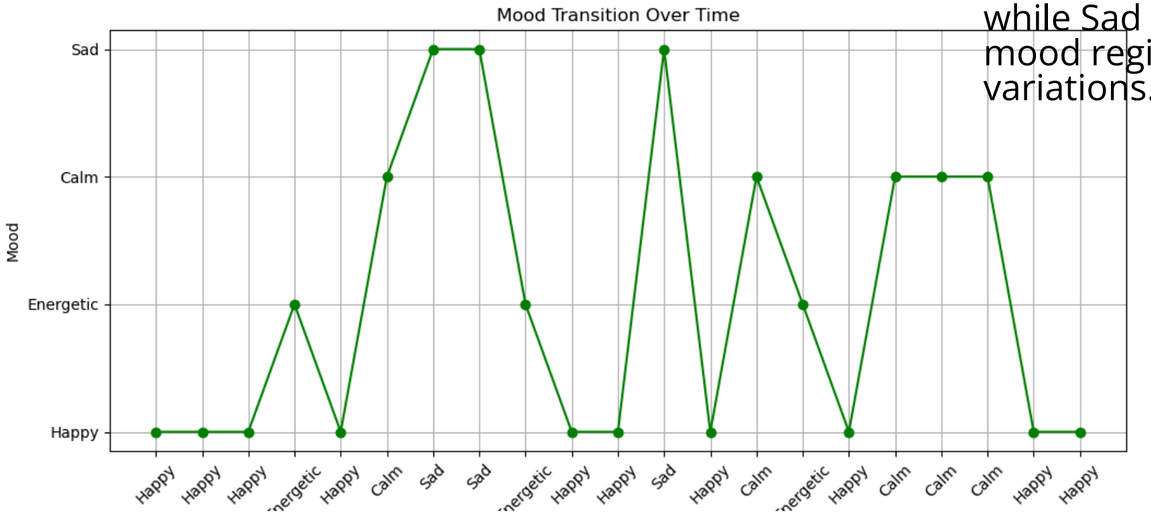
- The y-axis represents valence (a measure of how positive or happy a song sounds).
- The x-axis represents different mood categories.
- Observations:
- Happy and Energetic moods have high valence, meaning they are linked to more positive-sounding songs.
- Sad and Calm moods have low valence, meaning these moods are associated with more melancholic or neutral songs.

**Key Takeaways**

- Energetic & Happy songs tend to have high energy and valence.
- Sad & Calm songs have lower energy and valence, making them more subdued.
- These insights can help in mood-based music recommendations (e.g., suggesting energetic songs for workouts and calm songs for relaxation).

```
Accuracy of the Markov chain model: 0.6553865986932701
```

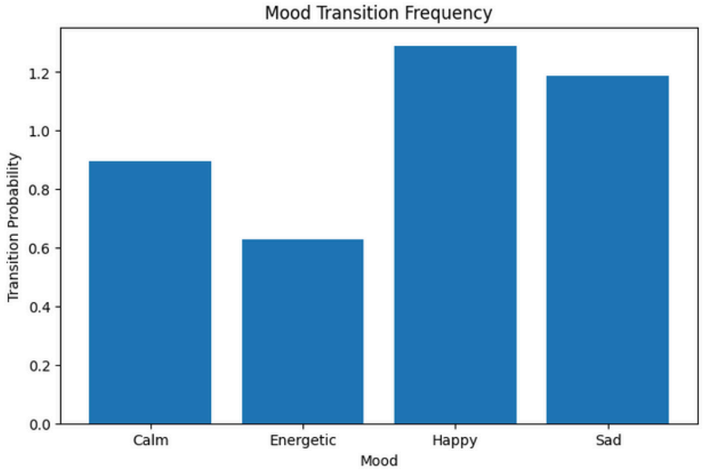
# Markov Chain Analysis- Build a Markov Chain for Mood Transitions

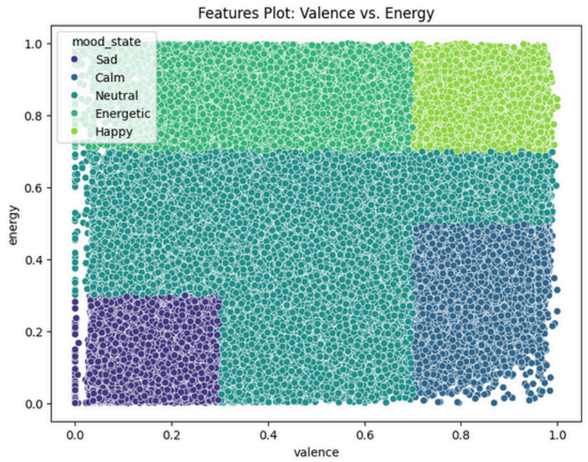


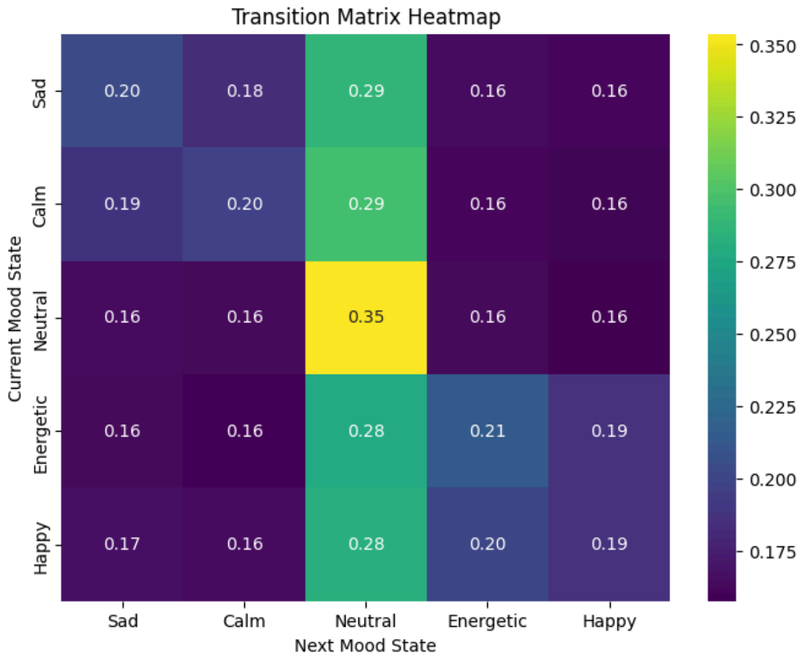

The plot maps mood states based on valence (positivity) and energy levels. Happy and Energetic cluster at high valence-energy, while Sad stays low. Distinct mood regions highlight emotional variations.

The plot shows mood transitions over time using a Markov Chain, where each mood change depends only on the previous state. The x-axis represents time (with mood labels), and the y-axis shows different moods (Happy, Energetic, Calm, Sad). The fluctuations indicate how moods shift probabilistically over time.

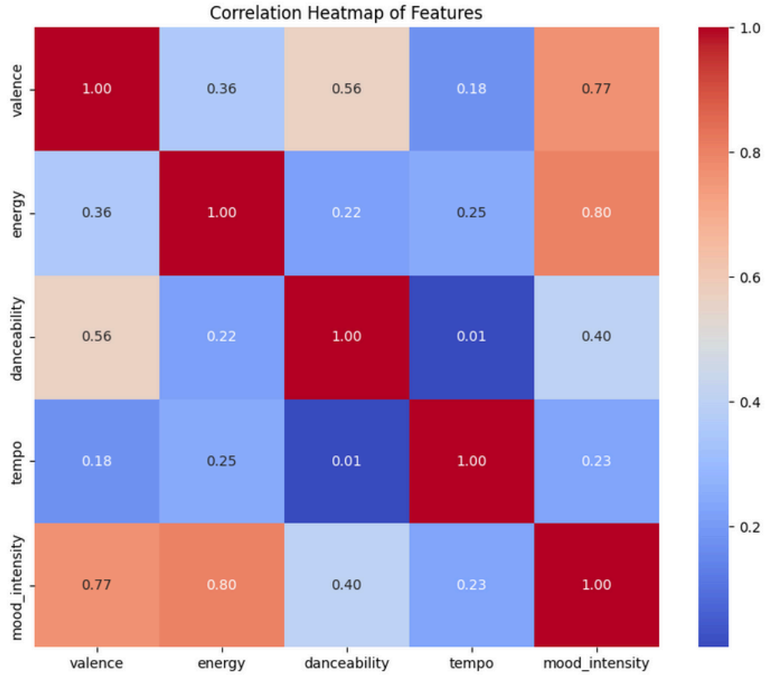
The chart depicts mood transition probabilities, where Happy and Sad have the highest likelihood of occurrence, while Energetic is the least frequent. This suggests that mood shifts tend to stabilize around Happy or Sad, with fewer transitions into the Energetic state.



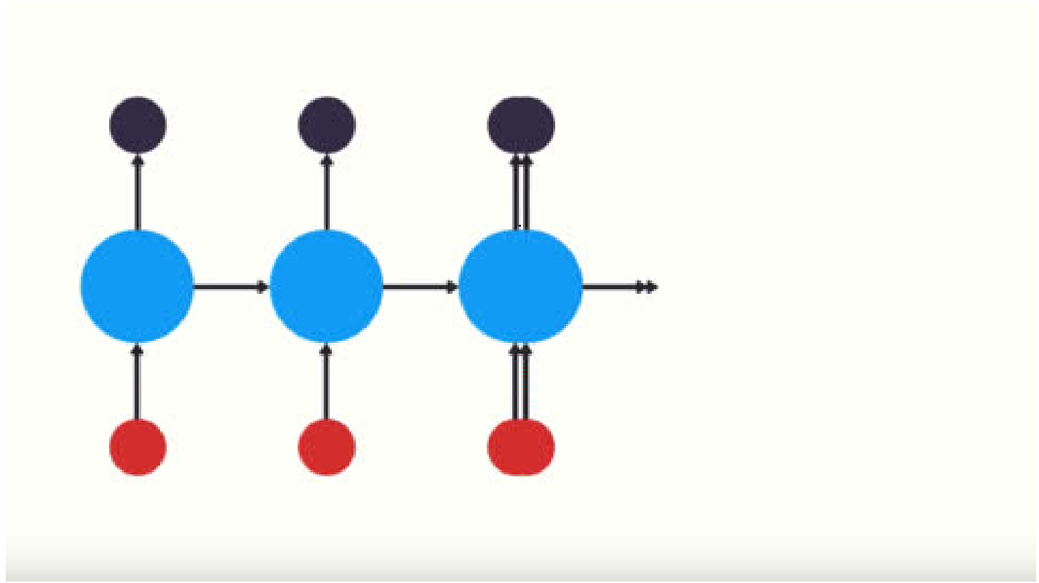



- The Transition Matrix Heatmap shows mood shift probabilities between Sad, Calm, Neutral, Energetic, and Happy.
- Each cell represents the likelihood of moving from one state (row) to another (column), with brighter colors indicating higher probabilities.
- Neutral has the highest self-transition (0.35), suggesting mood stability, while other states have more balanced transitions.
- This helps analyze mood patterns and predict emotional shifts.




- The Correlation Heatmap of Features shows relationships between valence, energy, danceability, tempo, and mood intensity.
- Strong correlations (red) indicate a high relationship, while weak ones (blue) suggest minimal impact. Mood intensity strongly correlates with valence (0.77) and energy (0.80), implying they heavily influence emotional states.
- Danceability and tempo have weaker links, suggesting a lesser effect on mood.
- This helps in analyzing key factors driving mood variations.

# Recurrent Neural Network (RNN) for Mood Classification



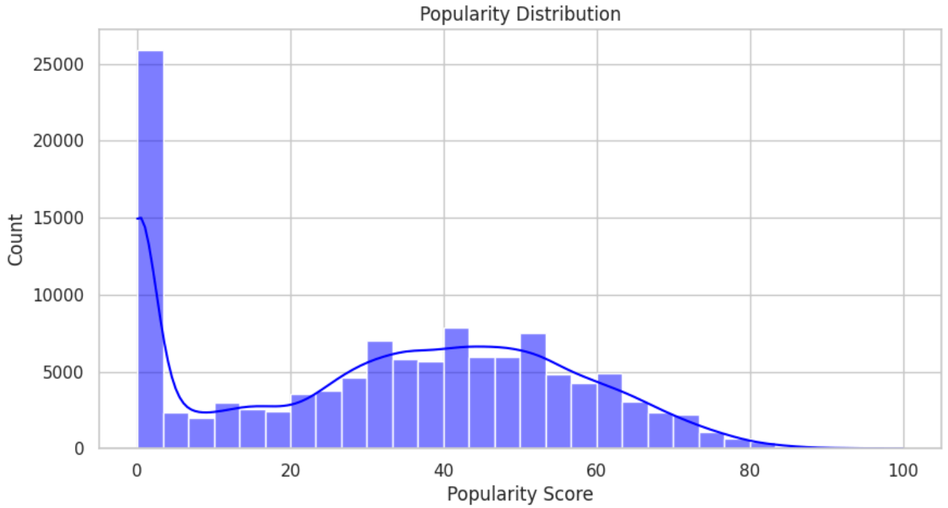
Recurrent Neural Network (RNN) is a deep learning algorithm designed to handle sequential data by maintaining memory through hidden states.
Unlike traditional neural networks, RNNs process input in order, making them effective for time-series and natural language tasks.

**Features:**

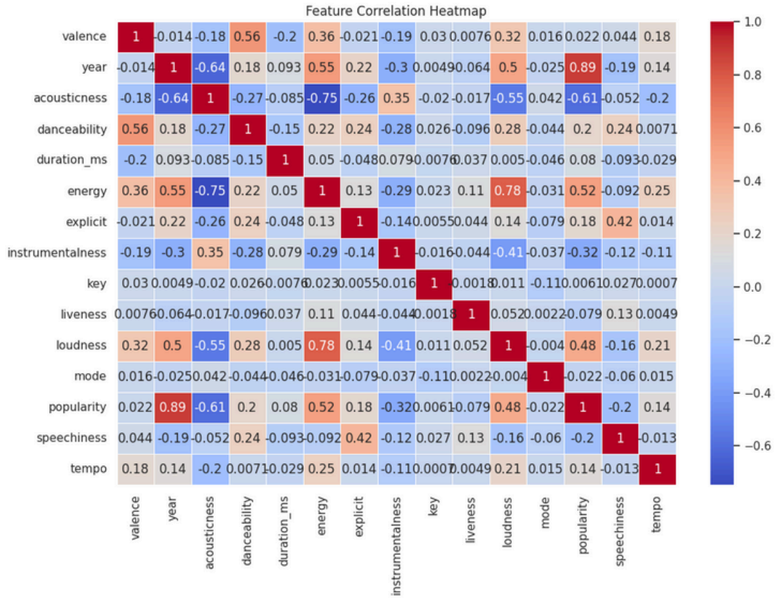
- RNNs capture temporal dependencies in data, making them ideal for analyzing mood transitions over time.
- They use feedback loops to retain past information, allowing sequential learning.

# Build an RNN for Playlist Prediction

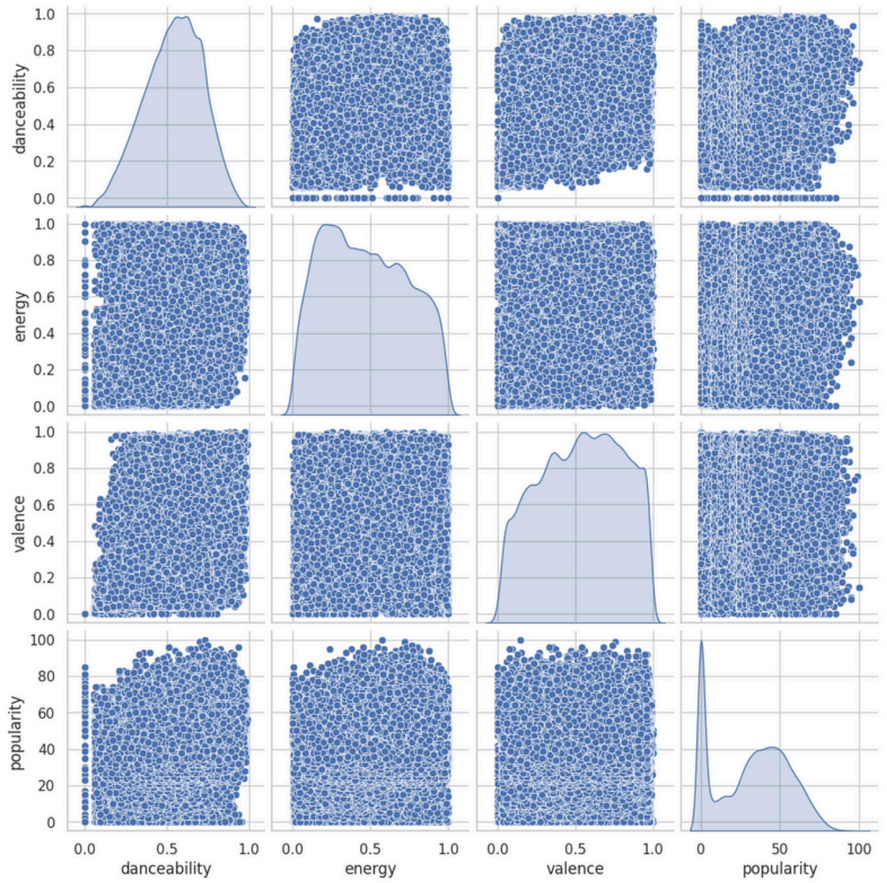
```
Epoch 10/10
2134/2134 ━━━━━━━━━━━━━━━━━━━━ 10s 4ms/step - accuracy: 0.6464
1067/1067 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step
Model Accuracy: 0.64
```



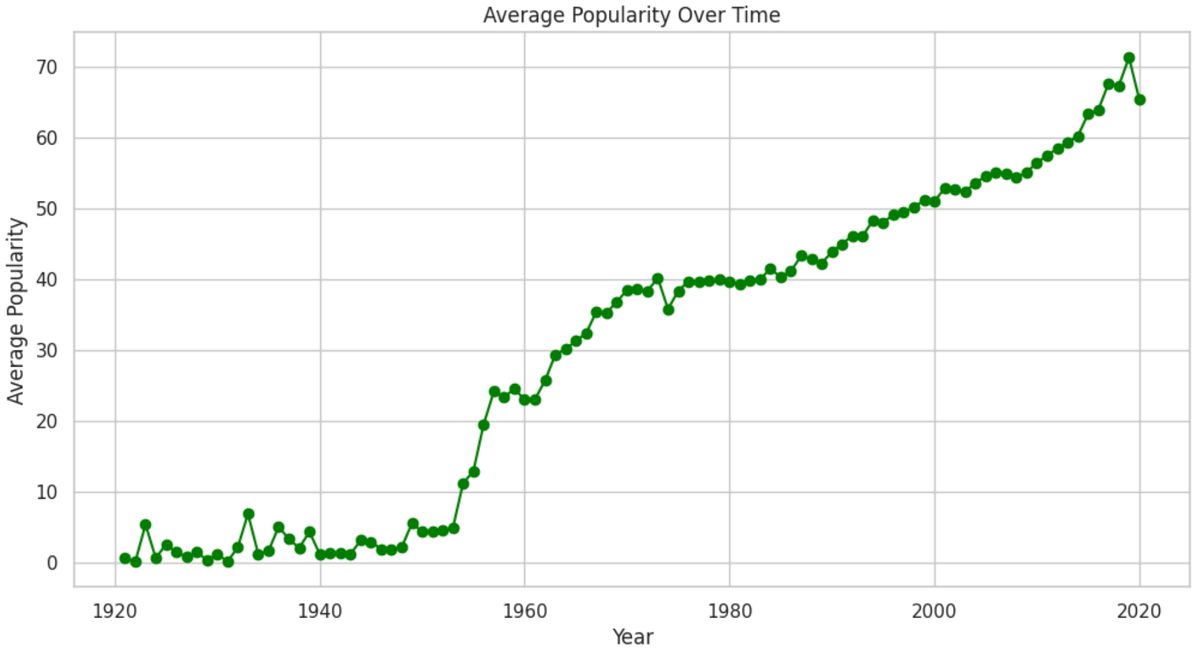

The Popularity Distribution shows most items have low popularity (near zero), with a peak around 40-60 and fewer highly popular items. This indicates a long-tail effect, where few items dominate in popularity.




The Feature Correlation Heatmap shows relationships between musical attributes. Popularity correlates with danceability, while acousticness is negatively linked to energy and loudness. Red indicates strong positive, blue strong negative correlations, aiding in music analysis.



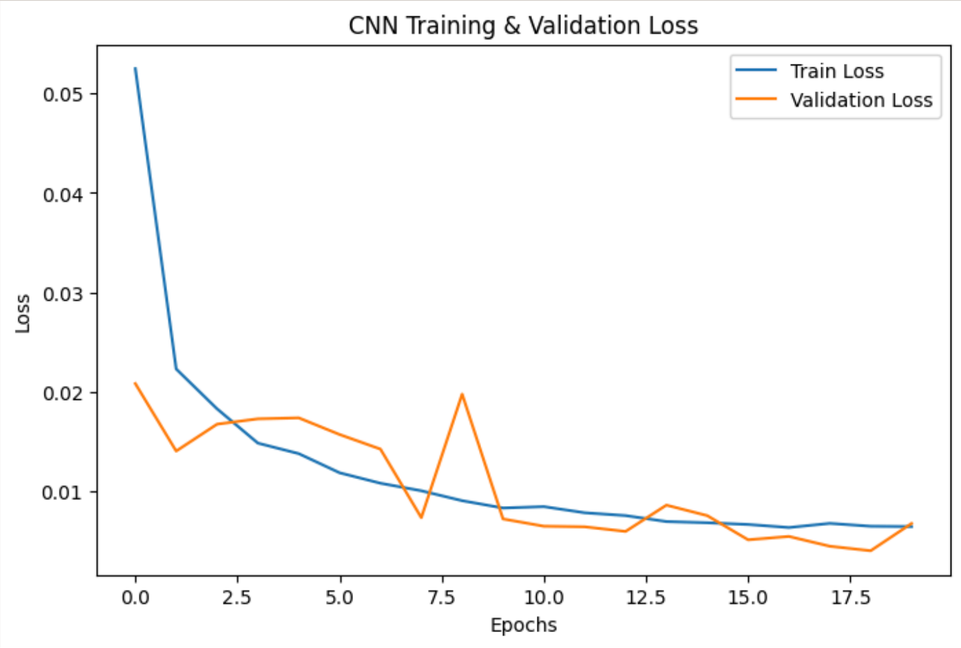

The pair plot visualizes relationships between danceability, energy, valence, and popularity. Diagonal plots show individual feature distributions, while scatter plots reveal correlations between attributes. Popularity has a skewed distribution, while energy and valence show some clustering patterns. This helps in understanding trends and dependencies between musical features.



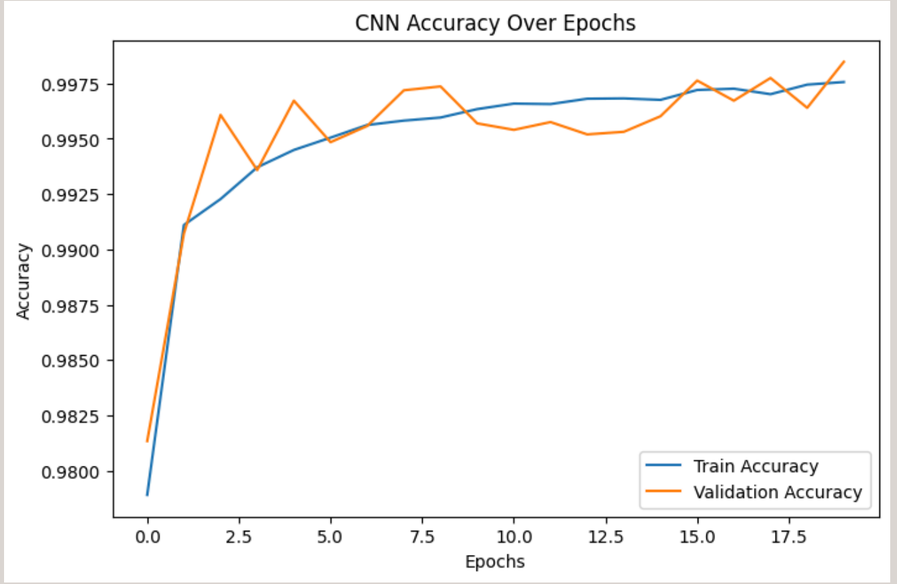

The Average Popularity Over Time plot shows a steady increase in song popularity from 1920 to 2020. Popularity remained low until the 1950s, then saw a sharp rise, stabilizing around 1970, followed by a consistent upward trend. The highest growth occurred post-2000, indicating increasing engagement with newer music.

# THANKYOU

# *CNN Model to enables mood-based playlist curation without requiring raw audio files.




CNN Training vs Validation Loss :-
Loss decreases over epochs, meaning the model is learning well.
Validation loss fluctuates slightly, but overall trend is downward, showing good generalization.




CNN Accuracy Over Epochs

- Both training and validation accuracy improve and stabilize above 99%.
- Similar trends for both curves indicate no overfitting, making the model reliable.