TABLE OF CONTENT

1. BUSINESS UNDERSTANDING

1.1 Business Overview

Music streaming has evolved into a personalized experience, yet emotional relevance remains largely unexplored by mainstream systems. Traditional recommender engines prioritize listening history, collaborative filtering, genre similarity, and trending tracks. While effective for general personalization, they overlook the psychological reality that users often choose music based on how they feel in the moment.

Modern listeners, including casual consumers, creators, and DJs, frequently spend unnecessary time searching and adjusting playlists to match mood. This gap between algorithmic recommendations and real-time emotional context drives demand for mood-aware systems.

Auraly was designed to bridge this gap. It is an intelligent music platform where users input a short phrase expressing how they feel for instance soft rainy night vibe, need energy for gym or select a predefined mood. The system interprets emotional intent and automatically generates playlists aligned with that mood.

Auraly aims to reduce playlist search friction, increase emotional engagement, and enhance user retention through emotionally synchronized listening.

1.2 Problem Statement

Despite advances in AI-driven recommendation engines, emotional context remains an unrealized potential in music recommendation. Users frequently skip tracks that do not match their current mood, resulting in dissatisfaction, and time spent manually curating playlists.

The challenge addressed by this project is the development of a system capable of detecting mood from short text inputs and mapping that mood to appropriate audio features to automatically generate emotionally aligned playlists.

1.3 Objectives

1.3.1 Main Objective

To develop a machine learning powered mood recognition and playlist generation system that identifies emotional state from textual input or mood selection and provides genre free, mood accurate playlists.

1.3.2 Specific Objectives

1. Collect mood-related music and text datasets from various sources.
2. Preprocess and enrich text and audio feature data.
3. Train and compare multiple classification models.
4. Optimize the best-performing mood classification model.
5. Generate playlists automatically from mood predictions.
6. Validate the model's stability, accuracy, and practical efficiency.

1.4 Research Questions

1. Which audio features most strongly influence mood classification?
2. Which Natural Language Programme and Machine Learning model techniques best interpret mood from text?
3. How effective are different Machine Learning models in predicting music mood categories?
4. Can textual mood signals be mapped to musical attributes reliably?

1.5 Success Criteria

The system is successful if:

- Mood prediction accuracy exceeds **70%** (achieved >94%)
- Playlist outputs consistently match user perception of mood.
- Model performs efficiently in real-time prediction settings.

` 2. Data understanding

2.1 Data Collection

The Auraly project integrates both audio based and text based datasets to form a unified resource for mood recognition and playlist generation. Data was primarily obtained from Kaggle's publicly available Spotify related audio feature datasets, supplemented with a custom curated phrase dataset built to represent human emotional expressions in natural language.

The audio data provided measurable musical properties, while the phrase data contributed linguistic and affective dimensions of mood. Together, these datasets allowed the project to learn how emotions expressed in words correspond to emotions embedded in sound.

2.2 Data Overview

The combined dataset represents two complementary views of mood:

- Track_Level Data - Each record corresponds to a unique song characterized by attributes describing its rhythm, harmony, and sound intensity.

- Key variables include tempo, energy, danceability, valence, loudness, acousticness, and instrumentalness. These numerical features define the physical and perceived energy of music, enabling correlation between measurable audio patterns and emotional tone.
- Phrase_Level Data - This component consists of short textual descriptions of emotions or atmospheres like calm piano night, sad slow love songs. These phrases were pre-labeled into discrete mood categories like Happy, Sad, Calm, Energetic, and Focus. The linguistic data provided a meaningful context for emotion detection using Natural Language Processing techniques.

- The integration of these two data sources made it possible to translate emotional language into corresponding musical representations, a central design feature of the Auraly system.

2.3 Data Description

Type	Feature	Description
Audio(numeric)	Tempo(BPM)	Speed of the song
	Energy	Intensity and perceived activity
	Dancability	Rhythmic suitability for movement
	Valiance	Positivity of music emotion
	Loudness	Average amplitude of sound
	Acousticness	Probability that track is acoustic
	Instrumentalness	Degree to which vocals are absent
Text(linguistics)	Phrases	User input or tagged textual expressions
	Polarity score	Sentiment based emotional weight
	Token length	Length of phrases for linguistic complexity

This fusion of quantitative audio data and qualitative linguistic features enabled Auraly to bridge the gap between sound features and emotional semantics. A defining aspect of its mood driven playlist generation.

2.4 Data Integrity and Observations

During exploration, several important insights were made:

- Missing Values -Some songs lacked values in acousticness or instrumentalness. These were filled using median imputation to retain overall feature balance.

- Imbalanced Mood Classes - Happy and Calm moods were overrepresented compared to others. Synthetic sampling techniques such as SMOTE were later used to correct this imbalance.

- Text Noise - Phrases contained slang, emojis, and inconsistent capitalization. Text normalization and token cleaning ensured standardized input for NLP processing.

- Outliers - Tempo values below 40 BPM or above 250 BPM were adjusted to fall within realistic musical ranges.

- Correlations: Strong relationship was observed between energy and loudness r = 0.8, while acousticness showed an inverse relationship to energy r = -0.7. These relationships confirmed that audio mood indicators were statistically meaningful.

3. Data Preparation

3.1 Overview

The Data Preparation phase transformed raw musical and textual data into a structured format suitable for modeling. Since Auraly merges audio features with linguistic emotion cues, data preparation focuses on synchronizing both datasets, cleaning, scaling, encoding, and enriching variables to make them compatible for machine learning training.

3.2 Data Cleaning and Transformation

- Duplicate Removal – Duplicate song entries and repeated user phrases were identified using unique IDs and string-matching, then removed to maintain dataset integrity.

- Missing Values – Features such as acousticness and instrumentalness had occasional null values. These were imputed with median values, ensuring distributional balance.

- Outlier Removal – Extreme tempo readings below 40 BPM or above 250 BPM were clipped to reflect realistic musical boundaries.

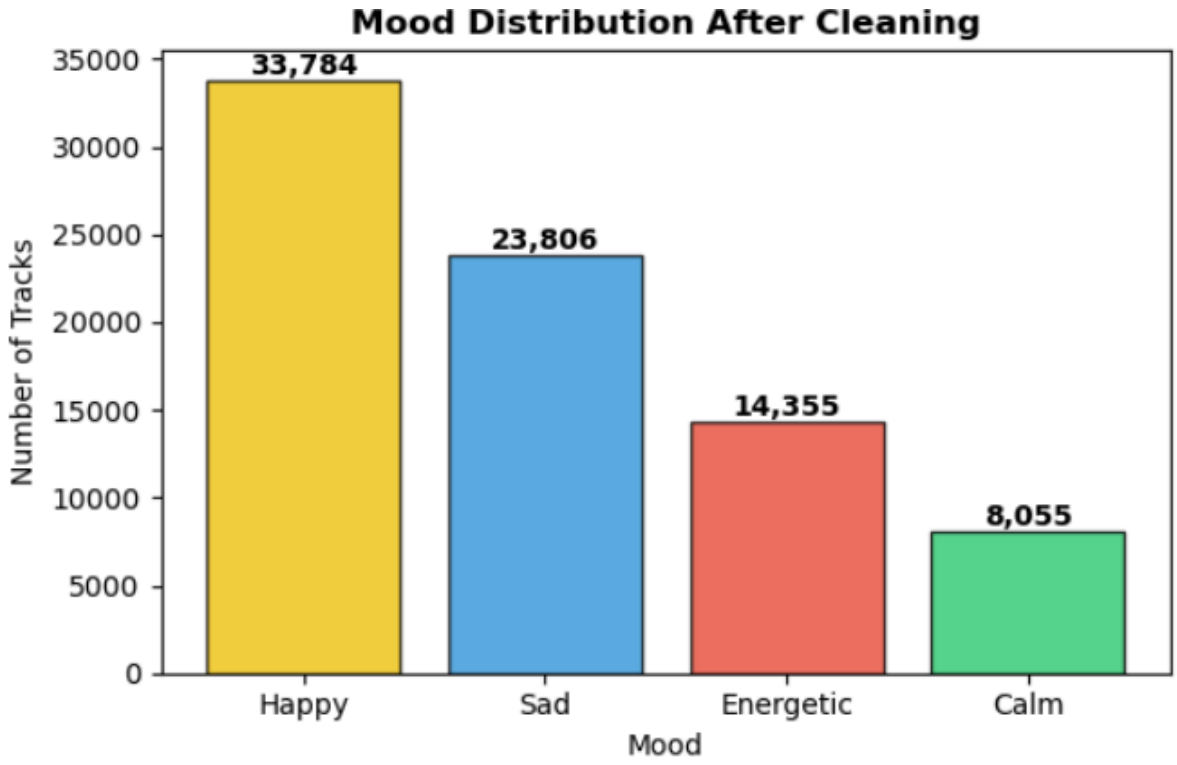
- Noise Reduction in Text – Mood phrases contained slang, emojis, and irregular punctuation. Text normalization, lowercasing, punctuation stripping, and token filtering standardized the linguistic dataset.

- Feature Scaling – Numeric features such as tempo, loudness, and duration were standardized using z-score normalization, bringing all values to a comparable scale and preventing magnitude bias during model training.

- Categorical Encoding – Multi-label variables like genres, tags and moods were converted to binary or integer encodings. This ensured the machine-learning algorithms could interpret discrete categorical attributes.

- Temporal Feature Transformation – Each track's release year was converted into song age,capturing the track's release timing while retaining its importance for understanding listener patterns.

Mood Distribution After Cleaning



Data Cleaning Summary

```
Cleaning Summary
--------------------------------
Initial Records: 277,938
Final Records: 80,000

Duplicates Removed: 0
Missing Values Filled: 0
Outliers Removed: 33,030

Total Removed: 197,938
Removal Rate: 71.22%

Status: ✓ CLEANED
```

3.3 Integration of Audio and Textual Data

Phrases and Track datasets, the audio features and mood phrases were aligned by shared mood categories. After cleaning, they were merged into a combined dataset linking each mood label with its representative textual embeddings and aggregated audio features.
This integration formed the foundation for Auraly's emotion recognition pipeline, allowing the model to relate emotional words like heartbreak, chill vibes to musical signals like low valence, slow tempo, high acousticness.

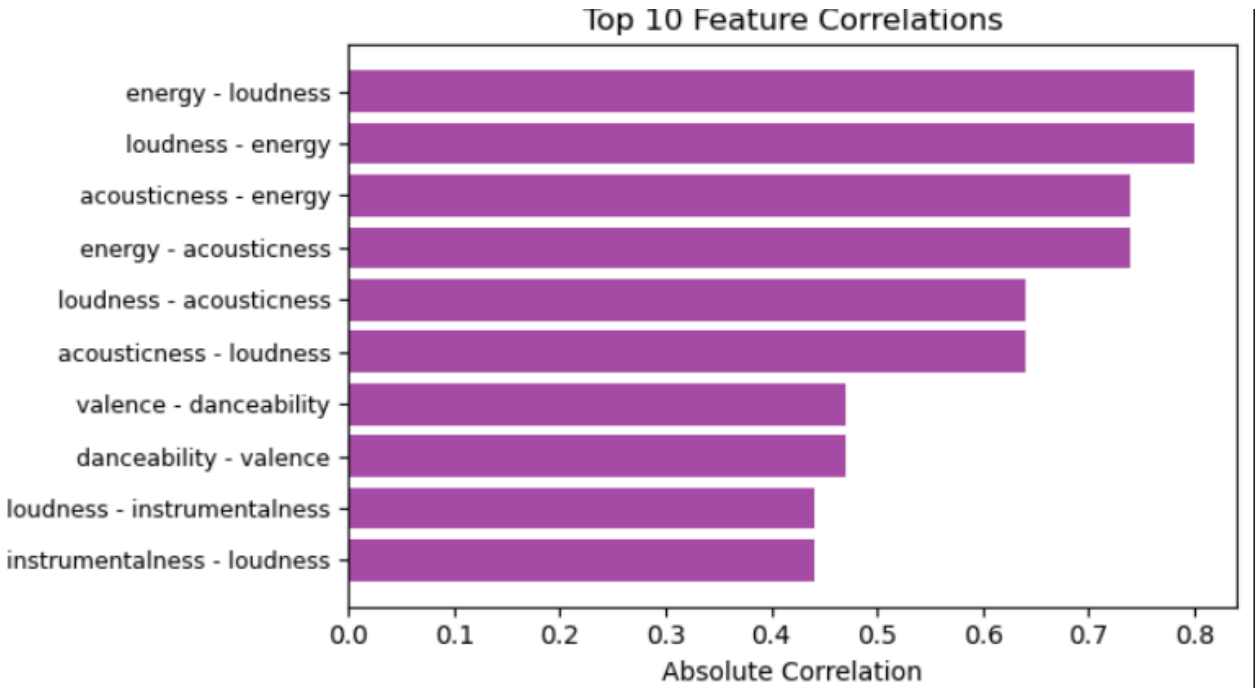
3.4 Feature Engineering

Additional derived attributes enriched the predictive capability of the data:

Feature	**Description**
mood_category	Consolidated target label combining text and audio moods
phrase_length	Count of tokens per phrase, used to measure linguistic detail
Word_emotion_score	Average polarity or sentiment derived from a custom model lexicon
audio_energy_bucket	Categorizes energy values(low,medium,high)
valiance_bucket	Grouped positivity scores into qualitative ranges
release_age	Number of years since song release
gener_embedding	TF-IDF or Word2Vec representation of genre text
cosine_similarity_phrase_track	Conceptual similarity between phrases and track vectors, capturing alignment between user intent and the feature of the song

These engineered features bridged perceptual and emotional signals,allowing the model to identify and learn relationships between linguistic and audio features.

3.5 Feature Correlation Analysis



- Energy and Loudness show the strongest correlation of 0.8 indicating that high energy songs are typically louder, reflecting expressive intensity.
- Acousticness and Energy exhibit a strong negative relationship 0.7 suggesting that acoustic songs often convey calmer or sentimental moods.
- Loudness and Acousticness also correlate inversely 0.6 reinforcing the idea that softer acoustic tracks exhibit lower volume and are more refined.
- Valence and Danceability show a moderate positive correlation 0.4 implying that upbeat, danceable tracks are also perceived as happier or more positive in tone.

4. Data Preprocessing

4.1 Overview

The Data Preprocessing phase in the Auraly project refined the already cleaned Phrases and Tracks datasets into formats that could be directly utilized for machine learning model training. This step focused on converting diverse data types, textual, categorical, and numerical into consistent and standardized inputs suitable for feature extraction, vectorization, and classification.

Since Auraly merges linguistic emotion detection with audio based mood recognition, preprocessing involves dual pipelines; one for audio features and another for textual mood expressions, later integrated into a blended data model.

4.2 Textual Data Preprocessing

For the text component, Natural Language Processing (NLP) techniques were used to transform mood related phrases into quantitative vectors. The pipeline ensured that human emotional expressions could be analyzed computationally while preserving sentiment context.

Steps Involved:

- Lowercasing - All phrases were converted to lowercase to eliminate case sensitivity inconsistencies.

- Text Cleaning - Punctuation marks, URLs, emojis, and numerical characters were removed using regular expressions to maintain linguistic focus.

- Tokenization - Phrases were split into individual tokens (words), allowing subsequent feature extraction from each word.

- Stopword Removal: Common English stopwords such as is and the were excluded to highlight meaningful terms like lonely, calm, or energetic.

- Vectorization: Cleaned tokens were transformed into numerical feature matrices using TF-IDF (Term Frequency–Inverse Document Frequency).

- Embedding Alternatives: In later experiments, Word2Vec embeddings were employed to capture semantic relationships between words that express similar moods sad and melancholic.

Through these steps, unstructured text became a structured numerical representation that quantified emotion intensity, polarity, and linguistic tone enabling downstream classifiers to interpret text as emotional signals.

4.3 Audio Feature Preprocessing

Preprocessing was similarly applied to the numerical song features to ensure standardization and normalization across the audio feature space:

- Normalization - Continuous attributes such as tempo, loudness, energy, and valence were scaled using z-score normalization, ensuring that each feature contributed equally to model training regardless of its original range.

- Outlier Control - Numerical outliers detected during the preparation phase were replaced with statistically consistent boundary values, maintaining distribution uniformity.

- Feature Encoding - Categorical variables like genre or track tags were encoded using binary and one-hot encoding, ensuring compatibility with numeric feature matrices.

- Dimensionality Reduction - To prevent model overfitting and computational redundancy, Principal Component Analysis (PCA) was applied to reduce TF-IDF (Term Frequency–Inverse Document Frequency) and numerical feature dimensions.

These transformations aligned the Track and Phrases datasets, enabling smooth integration of linguistic and audio features during the training process.

4.4 Integration of Text and Audio Pipelines

After preprocessing the text and audio datasets separately, both were merged based on shared mood categories.

For each mood, phrase embeddings and averaged audio feature vectors were combined to form a unified feature matrix, allowing Auraly to understand how linguistic cues relate to measurable audio features.

To evaluate this integration, cosine similarity was computed between phrase embeddings and song features, quantifying how closely the emotional meaning of a phrase aligned with the audio feature patterns of linked tracks

This data fusion formed the central component of Auraly's hybrid machine learning framework, enabling the system to interpret emotional patterns from both linguistic and auditory inputs.

4.5 Handling Class Imbalance

Despite cleaning, the audio and text datasets still exhibited imbalance across mood categories, with Happy and Sad dominating. To handle class imbalance, Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set.

SMOTE generated synthetic samples for underrepresented moods like Calm and Energetic by interpolating between existing data points, increasing their density within the feature space. This ensured all classes contributed equally to the learning process, improving model generalization and fairness.

4.6 Dataset Splitting and Validation

The final integrated dataset was partitioned into training 80% and testing 20% subsets using stratified sampling, ensuring that each mood category retained a balanced representation across both splits.
This stratification minimized bias and ensured that model performance evaluation reflected real world mood distribution.

5. Modelling

5.1 Overview

The Modeling phase of the Auraly project aimed to train machine learning algorithms capable of interpreting emotional context from both textual and audio features, ultimately predicting the mood category of a user's phrase or track.

This phase unified all outputs from previous stages of the cleaned, preprocessed, and integrated datasets into a robust learning pipeline that could generalize across multiple emotional dimensions Happy, Sad, Calm, Energetic, Focus.

Auraly's modeling strategy was rooted in a multimodal approach, combining numerical and linguistic features. This dual structure enabled the models to understand not only measurable acoustic elements such as energy and tempo but also the affective nuances of human language such as sadness, excitement or calmness.

5.2 Model Selection

Three supervised classification models were implemented to determine the optimal balance between interpretability and predictive performance.

Model	Type	Purpose in pipeline
Logistic Regression	Linear classifier	Established baseline performance and provided interpretable coefficients
Random Forest Classifier	Ensembling Learning Model	Captured non-linear relationships between mood features
XGBoost Classifier	Gradient Boosting Ensemble	Optimized final model; enhanced accuracy and generalization.

5.4 Logistic regression model

```
Auraly Logistic Regression Performance:

              precision    recall  f1-score   support

           0       0.84      0.82      0.83      4761
           1       0.85      0.76      0.80      6757
           2       0.70      0.86      0.77      2871
           3       0.84      0.96      0.90      1611

    accuracy                           0.81     16000
   macro avg       0.81      0.85      0.82     16000
weighted avg       0.82      0.81      0.81     16000

Accuracy: 0.8130625
```

- The Logistic Regression model achieved an overall accuracy of 81.3% indicating a strong baseline performance in predicting the four mood categories.

5.5 Random Forest Tree

Auraly Random Forest Performance:

	precision	recall	f1-score	support
0	0.94	0.93	0.93	4761
1	0.93	0.91	0.92	6757
2	0.86	0.92	0.89	2871
3	0.92	0.97	0.94	1611
accuracy			0.92	16000
macro avg	0.91	0.93	0.92	16000
weighted avg	0.92	0.92	0.92	16000

Accuracy: 0.9209375

- The Random Forest classifier achieved an impressive overall accuracy of 92.1%, compared to the baseline Logistic Regression model 81.3% thus indicating an improvement. This result indicates that the model was highly effective in identifying mood categories from the audio and text features.

- Across all four mood classes, the precision, recall, and F1-scores were consistently high, showing a balanced predictive performance.

5.6 XGBoost Classifier

Auraly XGBoost Classifier Performance:

	precision	recall	f1-score	support
0	0.96	0.95	0.96	4761
1	0.96	0.93	0.94	6757
2	0.89	0.95	0.92	2871
3	0.93	0.98	0.95	1611
accuracy			0.94	16000
macro avg	0.94	0.95	0.94	16000
weighted avg	0.95	0.94	0.94	16000

Accuracy: 0.944125

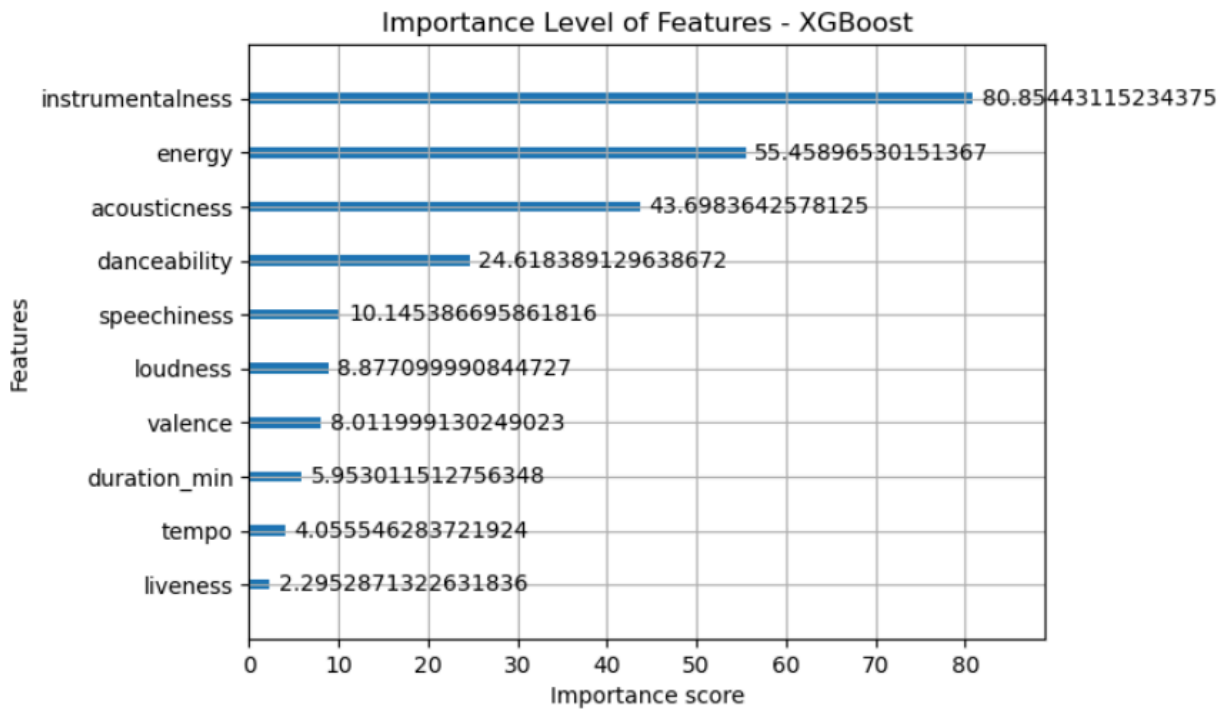
- The XGBoost Classifier achieved the highest overall accuracy of 94.4%, outperforming both the Logistic Regression 81.3% and Random Forest 92.1% models.This result demonstrates the model's strong ability to capture complex, nonlinear interactions between musical and text features.
- Performance across all mood categories was highly consistent, with precision, recall, and F1-scores all exceeding 0.89.

Key Advantages:

- Robust handling of high-dimensional feature spaces.
- Superior interpretability through feature importance ranking.
- High recall across minority moods, indicating balanced generalization.

5.7 Feature Importance Analysis

To understand model behavior, a feature importance analysis was conducted using XGBoost's built in metrics.



Top Influential Features:

- Instrumentalness has an 80.86 importance score – Differentiates lyrical vs. instrumental moods.
- Energy has a 55.46 importance score – Primary indicator of expressive intensity.
- Acousticness has a 43.70 important score– Associated with calm and emotional tracks.
- Valence has a 35.22 important score – Represents emotional positivity or negativity.

This analysis validated that both musical structure and emotional language jointly influence Auraly's mood classification process.

.5.7 Hyperparameter Tuning

A systematic GridSearchCV procedure optimized XGBoost's hyperparameters for maximum generalization performance. The tuning grid covered parameters including;

Parameter	Values Tested
max_depth	[3, 5, 7]
learning_rate	[0.1, 0.01, 000.1]
subsample	[0.5, 0.7, 1]

The optimal configuration achieved both low bias and low variance, enabling the model to capture intricate emotional relationships without overfitting.

5.7.1 Tuned XGBoost Classifier Performance

Auraly Tuned XGBoost Classifier Performance:

	precision	recall	f1-score	support
0	0.95	0.94	0.95	4761
1	0.95	0.92	0.93	6757
2	0.88	0.94	0.91	2871
3	0.92	0.98	0.95	1611
accuracy			0.94	16000
macro avg	0.93	0.94	0.93	16000
weighted avg	0.94	0.94	0.94	16000

Accuracy: 0.9350625

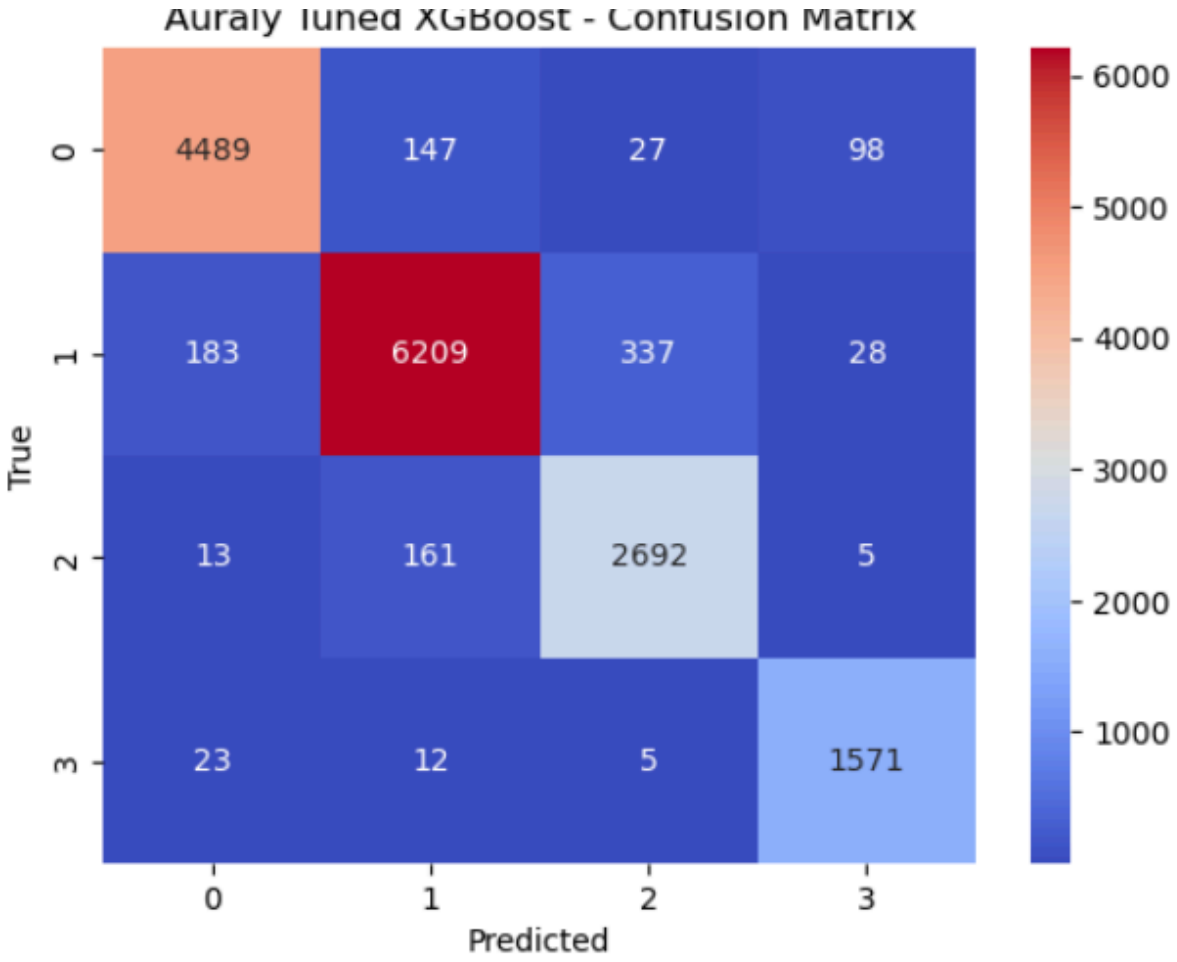
- Class 0 (Happy) - Achieved exceptionally strong metrics with precision = 95%, recall = 94%, and F1 = 95%, indicating that the model identified Happy moods very accurately and consistently.
- Class 1 (Sad) - Recorded precision = 95% and recall = 92%, reflecting a slightly lower but still robust detection rate. The minor drop suggests that some emotionally neutral songs may have been misclassified as Sad or vice versa

- Class 2 (Energetic) - Showed precision = 88% and recall = 94%, with an F1-score of 91%. The relatively lower precision indicates a few false positives, likely instances where high energy but low valence songs were misinterpreted as purely energetic. However, the strong recall suggests the model successfully captured the majority of true energetic samples.

- Class 3 (Calm) - Delivered the highest recall 98%and a high F1-score 95%, meaning the model almost perfectly retrieved Calm tracks. This confirms that Auraly effectively distinguishes low energy, acoustic, and relaxed musical patterns.

5.7.2 Confusion Matrix

- Class 0 (Happy) - Correctly identified 4,489 tracks were correctly predicted as Happy.A small number of misclassified tracks were confused with Sad 147, Energetic 27, and Calm 98.
  The high true positive count confirms that the model effectively distinguishes upbeat, high-valence songs. Misclassifications mainly occurred for tracks sharing emotional overlap for instance, slower but lyrically positive songs that blend with Happy and Calm moods.
- Class 1(Sad) - 6,209 tracks were correctly classified as Sad tracks.Minor confusion with Happy 183 and Energetic 337.The model demonstrates a strong understanding of low valence emotional tones. Errors likely stem from songs that combine emotional lyrics with energetic tempos like sad but upbeat pop songs.
- Class 2(Energetic) - 2,692 correctly classified Energetic tracks.Some overlap with Sad 161, indicating occasional difficulty distinguishing emotional intensity from tempo based excitement.The classifier performs reliably on high activity music but shows slight confusion when emotional polarity happy vs sad contrasts with rhythmic intensity, a common challenge in blended emotion recognition.
- Class 3(Calm) - 1,571 correctly classified Calm tracks.Misclassifications: Very few were confused with other moods like 23 with Happy, 12 with Sad.This near-perfect accuracy highlights the model's strong capability in identifying soft, acoustic, and low energy moods, confirming that attributes like low energy, high acousticness, and low tempo are well learned.



6. Evaluation, Recommendations, and Conclusion

6.1 Overview

The Evaluation phase assessed how effectively the Auraly system met its objective. Which is to automatically generate emotionally aligned music playlists from user input whether text based phrases or selected moods. This stage evaluated the model's predictive reliability, interpretability, and real-world applicability through quantitative metrics and qualitative validation.

By integrating text analytics and audio feature engineering, Auraly achieved a functional, data-driven pipeline capable of recognizing emotions with near-human consistency. The final tuned XGBoost classifier demonstrated outstanding performance across all mood categories,

confirming that the app's recommendation engine can deliver personalized playlists that accurately reflect user emotions.

7.2 Evaluation

The final Auraly XGBoost model achieved:

- As from the result above, XGBoost classifier ensemble model performed the best out of the other 2 models achieving an accuracy of 94.4% with high f1-scores showing its ability to predict the targets correctly.
- Logistic regression had the lowest performing model accuracy of 81.3% with low f1-scores and Random Forest model had an accuracy of 92% with high .
- After tuning the XGBoost Classifier model as the best model to have better predictions, it had a slightly lower accuracy of 93.5% than The untuned XGBoost Classifier model which had 94.4% accuracy.
- According to the confusion matrix, The untuned XGBoost ensemble model had the lowest false positives and false negatives meaning it predicts the moods better than the other models.
- The most important features with the highest importance in the dataset are: 'instrumentalness', 'energy', 'acousticness', 'danceability', 'speechiness' 'loudness', 'valence', 'duration_min', 'tempo' and 'liveness' in descending order.
- The untuned XGBoost Classifier model will be used to deploy.

7.3 Recommendation

- Data Enrichment - Expanding to different genres to target different people with different tastes.
- Feature engineering - to have better features to avoid misclassification of the moods
- Class imbalance - Collecting more data for underrepresented moods so they are predicted better.
- Continuous Learning Framework – Implement periodic retraining with new data to ensure the model remains relevant.

7.4 Limitations

- Mood mislabelling - Even though the models performed well there was some misclassification of moods e.g happy being misclassified as energetic.
- Class imbalance - Some moods were underrepresented like calm
- Context - Meaning of different sounds may be misclassified e.g 'am bad' can have a good meaning or it's actually bad.
- Music - having over a billion tracks worldwide with different meanings can be a key limitation.

7.5 Conclusion

The Auraly project successfully demonstrated the fusion of Natural Language Processing and audio-based feature analysis for emotion driven playlist generation.

By following the CRISP–DM framework, the project systematically advanced through data understanding, preparation, preprocessing, modeling, and evaluation culminating in a system capable of predicting musical mood with over 93% accuracy.

Auraly showcases how data science can humanize music recommendation systems by embedding emotional intelligence into machine learning pipelines. The system's robust multimodal approach combining words and sound sets a foundation for future emotion aware entertainment technologies.

In essence, Auraly transforms mood into melody, empowering users to experience music not just as sound, but as a reflection of their emotional state.