Data Mining and Machine Learning (F21DL)

Title of Project: *Beat The Charts*

Group Name: *Cool Crew*

Github Repository Link : *https://github.com/dmml-heriot-watt-2024-25/group-coursework-cool-crew*

Group Members:

Abdullahi Salihu Audu - H00462758

Abhishek Basavaraju - H00462156

Omana Prabhakar - H00459314

Reet Sharma - H00462451

Sandeep Srinivasan - H00410219 (Group Leader)

Introduction:

"Beat the Charts" is a research project which aims to predict which songs will become chart-topping hits by analysing audio features of the song. The team decided to work on a song dataset as most of the team members are active Spotify listeners and chose to investigate and uncover insights about the features of the song. The research conducted by (Harriman Samuel Saragih, 2023) is limited to Indonesian streaming users whereas in this study we have used the Spotify dataset as Spotify stands out as one of the leading platforms with a vast amount of user data (Trivedi *et al.*, 2024) which is publicly accessible on Kaggle which consists of 30,000 rows and 14 columns which we will be using to explore the audio features of the songs. It also contains songs from various countries having various musical features (Pinarbaşi, 2019; Suh, 2019) which adds diversity to our study, allowing us to explore musical patterns across regions and is interesting in conducting this research. Our objectives are to identify key attributes of popular songs and develop a model for predicting song popularity; Specifically, we aim to answer the question such as "Which audio feature contributes more to the popularity of the song?". We carry out the analysis using Python for data cleaning, EDA, and visualisations. Using multiple model comparisons for Regression, KNN, and Decision Trees, this study identifies KNN and Regression as the three predictive approaches with the highest accuracy. Preliminary studies by (Harriman Samuel Saragih, 2023) have shown correlations between certain audio features and popularity with an emphasis on Indonesia based on consumer culture theory but our research expands on this by using a large dataset with more varied geographical representation. We also discussed how these findings could benefit end-users; so, finally, this research could benefit end-users such as music producers and artists, providing them with predictive insights into which songs are more likely to become hits. By identifying patterns in popular songs, we hope to contribute actionable recommendations for those in the music industry, as well as a deeper understanding of musical trends."

Research Questions:

RQ1: Can machine learning models accurately predict a song's likelihood of becoming a hit by analysing its audio features?

RQ2: Which ML algorithm predicts the best hit song?

RQ3: Which audio feature contributes more to the popularity of the song?

Related work:

"Hit song science," according to some researchers, is the study of how a song's audio features contribute to its popularity (Middlebrook and Sheik, 2019). Raza & Nanath, 2020 are the earliest scholars who thoroughly discussed how audio features in an audio track could predict its popularity in the market (Harriman Samuel Saragih, 2023). With over 381 million monthly active users, Spotify provides a rich dataset for understanding music listening habits. (Trivedi *et al.*, 2024).

(Trivedi *et al.*, 2024) study contributes to the scientific literature on hit songs by examining the influence of audio features on a song's popularity using both classification and regression machine learning methods, with an emphasis on Indonesia based on consumer culture theory. Based on consumer culture theory the paper explains the different causes for music not to be popular even though they have the potential to be.

As a product, music offers several genres targeted toward different consumer segments (Pérez-Gálvez et al., 2017). Due to the different social and demographic characteristics, some music genres in a particular region might not be as popular in demand as others (Kotarba, 2013). This concept explains why jazz is fragmented and displays exclusivity, but pop and easy listening are more acceptable to music consumers.

As mentioned in (Harriman Samuel Saragih, 2023) paper that the study by Nijkamp (2018) carried out a separate study of a specific key but indicated that it is a "high base key. This finding is in line with the results from Indonesia that included 1000 songs using Spotify data. His findings implied relatively weak explanatory power (R2=20.2%) regarding the relationship of audio features to popularity. However, some audio features are not significant at the 95% level (Harriman Samuel Saragih, 2023). Febirautami et al. (2018) are the first to examine the relationship between audio features and popularity in the Indonesian music market. Their work utilised 200 Spotify data, primarily from local artists, and a Decision Tree categorization method. The target variable is composed of the terms "popular" and "unpopular." The overall accuracy resulted in a rather high score of 72.8%, with five significant features for the classification, namely acousticness, liveness, energy, valence, and key (they did not specify a specific key) which again highlight that different countries and regions have different audio characteristics that lead to popularity– even if the sample is relatively small (Febirautami et al., 2018; Pinarbaşi, 2019).

Middlebrook and Sheik (2019) obtained data from approximately 16,000 songs globally from 1985 to 2018. Using several machine learning algorithms, random forest methods yielded the highest accuracy. Their results indicated that based on hit and non-hit songs, audio features predict 88% of popularity; however, their results did not report feature importance. Of all the studies that suggested Spotify's audio features to be popular, several authors also suggested no significant relationship between the two (Amsterdam, 2019; Raza & Nanath, 2020).

(Amsterdam, 2019; Raza & Nanath, 2020) study reveals that music is an art form that has been around for centuries. Constantly in flux, it has proved to be a very diverse field. With revenue numbers in millions of dollars each year, the music industry is growing rapidly.

Dataset Description and Analysis:

Dataset overview

The dataset includes 8,361 Spotify song entries that provide metadata and audio attributes. These entries were most likely obtained from Spotify's API. The dataset used is publicly available in Kaggle and does not contain any personal identifiers. This dataset is ethically cleared and publicly accessible for research purposes. We have chosen this dataset because it contains a variety of auditory characteristics that can affect a song's attractiveness, including danceability, energy, and valence. These qualities are necessary for popularity prediction since they are closely related to listener enjoyment and engagement.

Columns & Features : The dataset consists of 17 columns consisting

Metadata : a. Track_name -The title of the song

b. Track_artist - The artist of song or Song Performer

c. Track_popularity - Song Popularity score on spotify rating between 0 to 100

d. Playlist_genre - Main genre category of song like pop,rock etc

e. Playlist_subgenre - A Sub genre classification of the song

Link to the dataset : https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs

2. Audio Features :

Feature	Data Type	Description
Danceability	*Float*	Danceability describes how suitable a track is for dancing based on a combination of musical elements. A value of 0.0 is least danceable, and 1.0 is most danceable.
Energy	*Float*	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.
Key	*Int*	Predicts whether a track contains no vocals
Loudness	*Float*	The overall loudness of a track in decibels (dB). Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
Mode	*Float*	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
Speechiness	*Float*	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording, the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words.
Acousticness	Float	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
Instrumentaln ess	Float	Predicts whether a track contains no vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.
Liveness	Float	Detects the presence of an audience in the recording.
Valence	Float	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.

Figure 1: Description of the audio features explained by (Raza and Krishnadas, 2020)

Dataset Analysis: Data Preprocessing is a critical step in machine learning to ensure the data quality and consistency. We used excel and python to pre-process the data, aiming to optimise model performance. The preprocessing stages included data cleansing, handling missing values, feature selection, and scaling. First we loaded the necessary libraries to read the data and did the sanity check using "df.shape" and "df.info()" to understand the data. We have used the "df.isnull().sum()" function to find the missing values provided by the pandas library. Data cleaning included finding the missing values, duplicates and garbage values. After cleaning the data we used the "describe()" function to obtain descriptive statistics for numerical variables and we have used scatterplot, histogram etc,. using matplotlib library to understand the distribution of the data.

Experimental Setup:

Clustering

In this section we will describe the algorithmic Choices that we have used to perform the prediction starting with Clustering task algorithms like K-means Clustering as it was implemented for unsupervised dataset based on audio features. DBSCAN was taken to detect any natural clustering within song popularity data with parameter tuning to handle noise and varying compactness.

Classification

Ahead for the Classification task we have primarily used K-Nearest Neighbours (KNN) . It was chosen because it makes it easy to distinguish between hits and non-hits while taking feature value similarity into account. Decision trees(DT)used to improve interpretability and provide insight into the significance of features. Logistic Regression(LR) algorithm was selected to efficiently anticipate binary events, namely to determine hit or non-hit potential and Neural Networks (CNN and MLP) Used to capture complex patterns, where CNN handles structured feature correlations and MLP addresses non-linear relationships.

Architectural and Hyperparameter Choices

Every model was fine tuned to increase its capacity for prediction.For the purpose of to further optimise the DT, LR, and KNN models for the dataset, Grid Search was used to optimise their hyperparameters.To prevent overfitting, CNN and MLP models utilised manual layer, node, and dropout rate adjustment.

To minimise overfitting and increase the effectiveness of MLP and CNN training, the Adam optimiser was chosen.

Evaluation Metrics

Accuracy, precision, recall, and F1 score were used to evaluate classification models' performance in song hit prediction. In order to identify convergence and possible overfitting, neural networks were further assessed using validation accuracy and loss metrics.This made it possible to test various algorithms with improved architectures and hyperparameters to successfully answer the research question, each of which helped to determine hit potential based on pre-release characteristics.

Algorithm Choice for Prediction

We have chosen algorithms like KNN, Logistic Regression, Decision Trees and Neural Networks to predict the hit songs. Each algorithm was used differently based on its strength among which was handled with the different audio features. As output we have got the hit songs with audio features which will satisfy our RQ1.

Algorithm Performance Analysis

As we had a comparison among the Algorithms and after applying evaluation metrics in which KNN gave us the most accurate hit song. Which will give an answer to RQ2.

Feature Selection and Analysis

For RQ3,Using correlation analysis to identify key audio features impacting Popularity we found out that Instrumentalness and Speachness directly impact on the song popularity as they are main features that will interact and engage with the people.

Results:

The findings from the experiments and analyses performed, as well as the outcomes of various machine learning models tested to answer the research questions are given below:

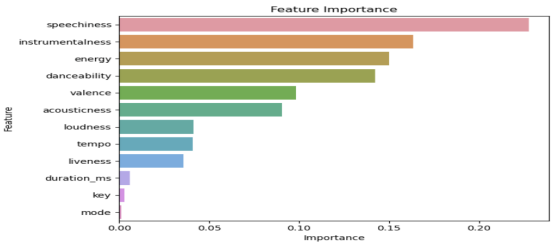
1. Correlation Analysis Insights:

The Highly Correlated Feature Pair**s** are "**Energy and Loudness** (0.704)".

2. Decision Tree Classifier (Refer to Appendix, fig 5 and fig 6):

Performance Metrics: Accuracy, Precision, Recall, F1 Score: All were at 0.882, indicating consistent performance across metrics.

Top 5 Feature Importances: Instrumentalness (0.212), Speechiness (0.186), Energy (0.160), Danceability (0.122), Valence (0.092).These results suggest that instrumental quality, speechiness, and energy contribute significantly to popularity predictions. (Refer to Appendix, fig 8 or the image attached below):



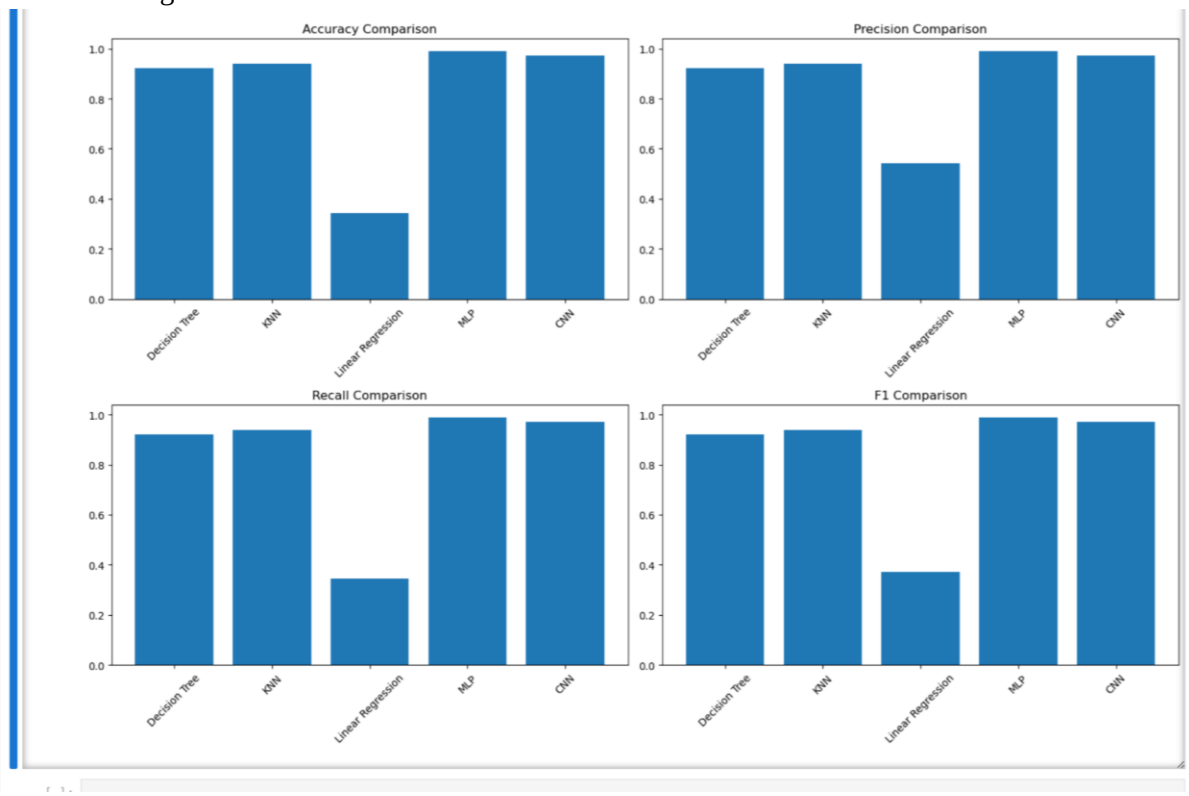
False Positive Rate

The other algorithms and their performance metrics are given below :-

Algorithms	Performance Metrics
K-Nearest Neighbors (KNN)	Accuracy: 0.913, Precision: 0.916, Recall: 0.913, F1 Score: 0.913.
Linear Regression	Accuracy: 0.325, Precision: 0.539, Recall: 0.325, F1 Score: 0.352.
Multilayer Perceptron (MLP)	Accuracy, Precision, Recall, and F1 Score were all 0.990
Convolutional Neural Network (CNN)	Accuracy, Precision, Recall, and F1 Score were each at 0.953.

(Refer to Appendix, fig 9: Detailed Analysis and Fig 10: Comprehensive Analysis Summary)

The **Best-Performing Model** is **MLP** achieved the highest accuracy (0.990), followed by CNN (0.953) and KNN (0.913). This result indicates MLP's suitability for this prediction task. You can see the figure below that aligns with this result section:



Discussions

Interpretation of Findings:

The key insights of this project are 1. The machine learning model , mainly MLP, can effectively predict the song popularity. 2. The feature importance where instrumentalness affects the most influential feature for predicting song popularity.

Connection to Research Questions:

- Research Question 1: The high accuracy of MLP and CNN shows that machine learning models can be helpful in predicting song popularity with considerable accuracy.
- Research Question 2: Among all the tested models, the MLP provided the best results that gives the answer about which ML algorithm is most effective for predicting hit songs.

- Research Question 3: Instrumentalness, speechiness, and energy emerged as key features, supporting the prediction that certain audio features can play an essential role in predicting popularity of a song.

Limitations and future work :

1. Only audio features were considered in this study, but there may be many different influential factors like artist popularity or marketing reach, that affect song popularity.
2. The dataset size might limit how well the model can work on new, unseen data in future.

In future, testing with larger amount of real-world dataset could validate the models further. We can also use different Machine learning algorithms to predict hit song more accurately and effectively. For further studies, we can look into some other features that considered to be more likely to predict the song popularity.

Conclusion:

With this whole analysis we come to the conclusion that machine learning model, mainly MLP, can effectively predict the song popularity. The key insights of this project also includes the feature importance where instrumentalness affects the most influential feature for predicting song popularity. While the models will perform differently and more successfully with a larger dataset, they can be extremely useful and promising in predicting song popularity in the music industry.

References

1. Harriman Samuel Saragih (2023) 'Predicting song popularity based on spotify's audio features: insights from the Indonesian streaming users', *Journal of Management Analytics*, 10(4), pp. 693–709. Available at: https://doi.org/10.1080/23270012.2023.2239824.
2. Middlebrook, K. and Sheik, K. (2019) *Song Hit Prediction: Predicting Billboard Hits Using Spotify Data*, *arXiv.org*. Available at: https://arxiv.org/abs/1908.08609.
3. Trivedi, D. *et al.* (2024) 'Harmonizing Insights: Python-Based Data Analysis of Spotify's Musical Tapestry', pp. 28–44. Available at: https://doi.org/10.1007/978-3-031-48888-7_3.
4. Raza, A.H. and Krishnadas, K. (2020) *Predicting a Hit Song with Machine Learning: Is there an apriori secret formula?* IEEE.

Appendix

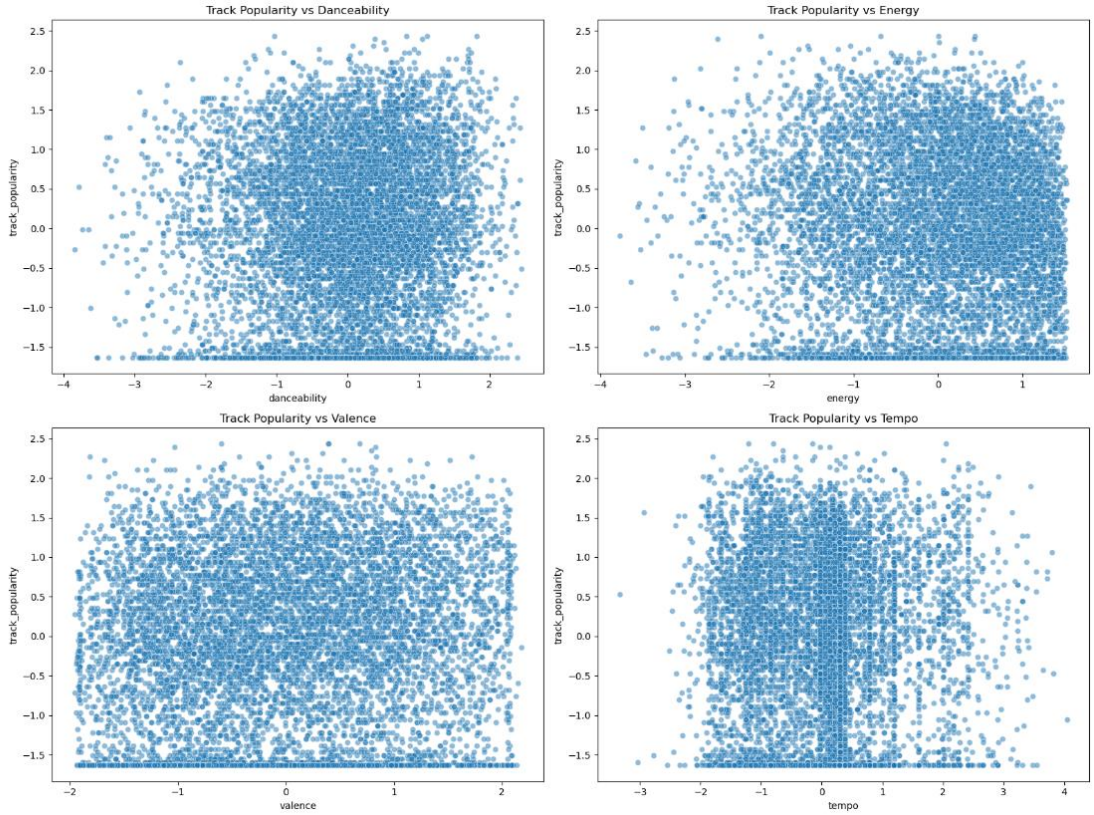


Fig 1 :EDA

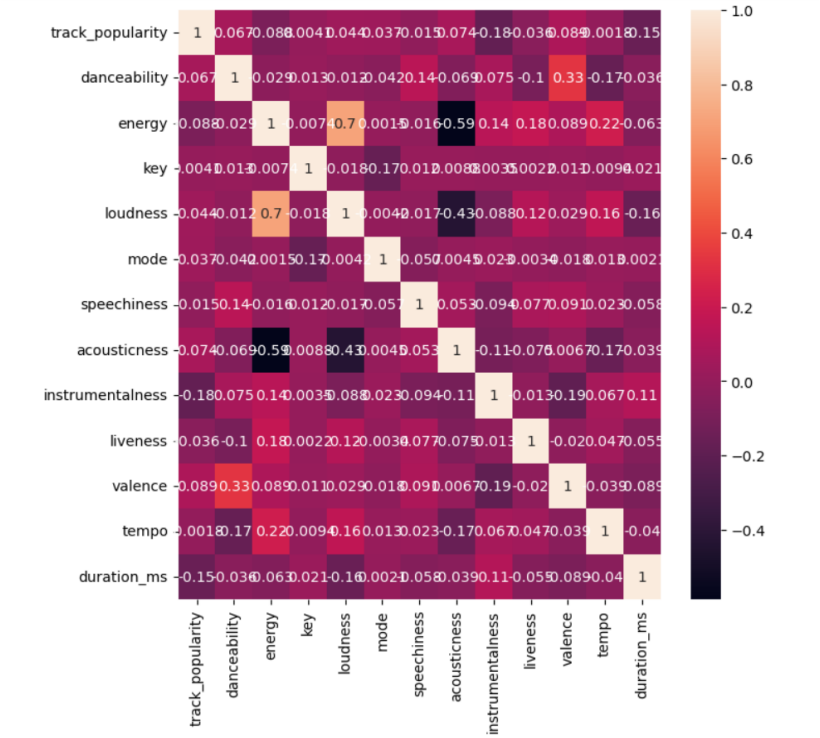


Fig 2: EDA

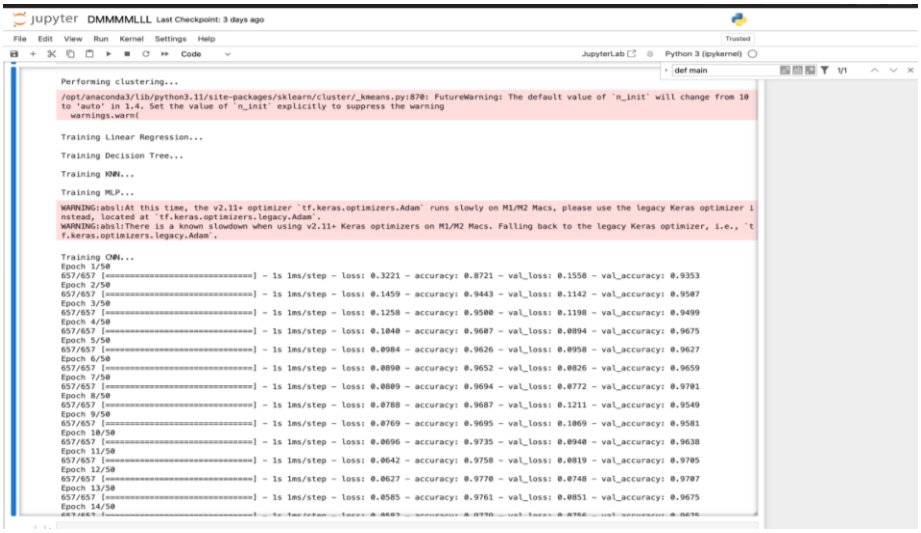


Fig 3: Performing clustering and training the model

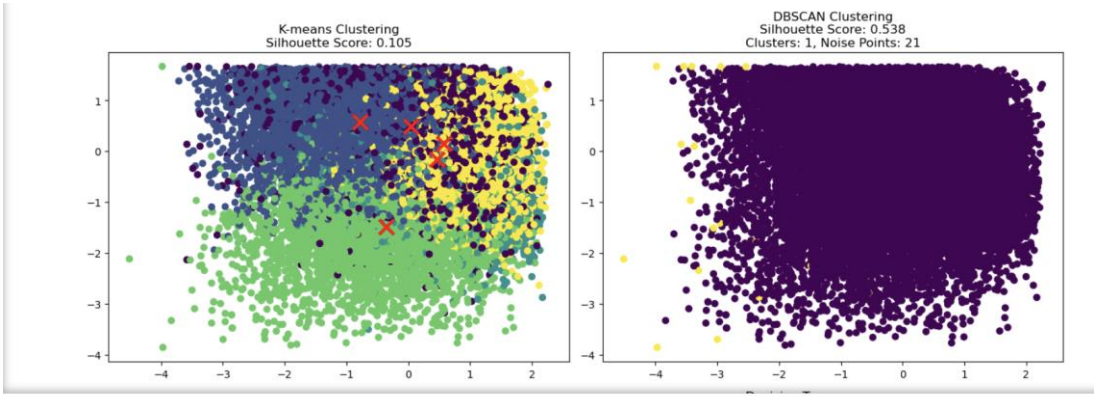


Fig 4: K- means clustering and DBSCAN Clustering

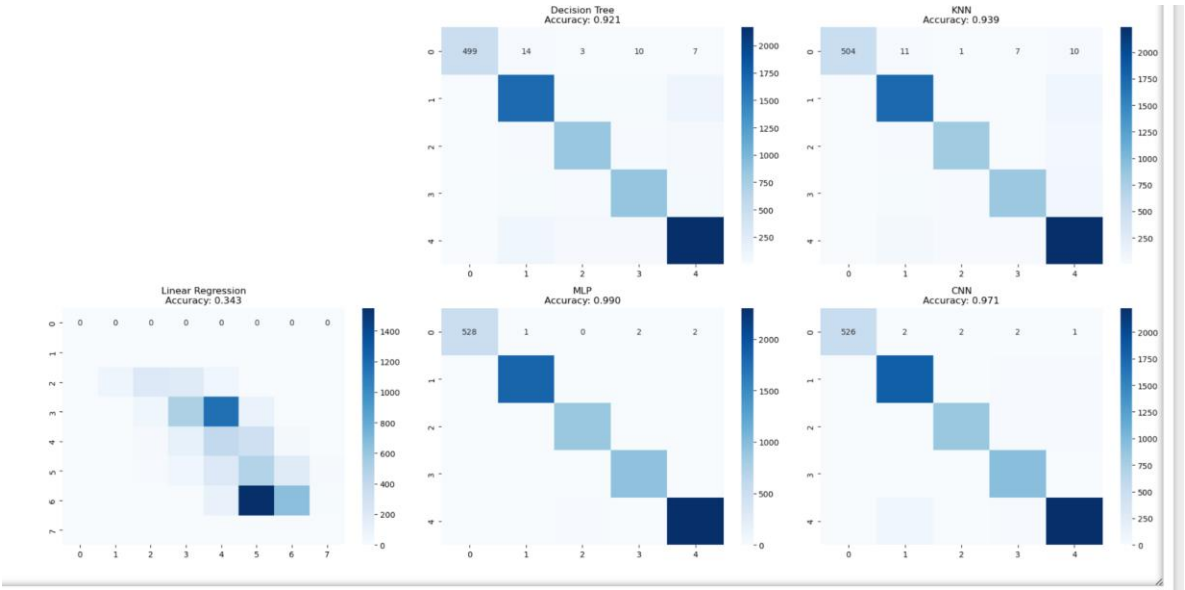


Fig 5:Performance metrics

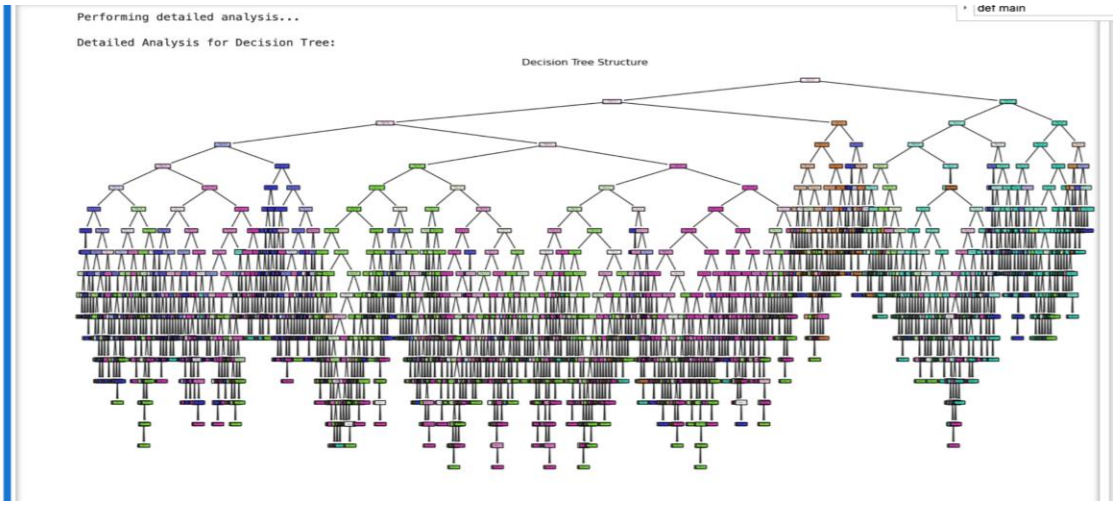


Fig 6: Decision Tree

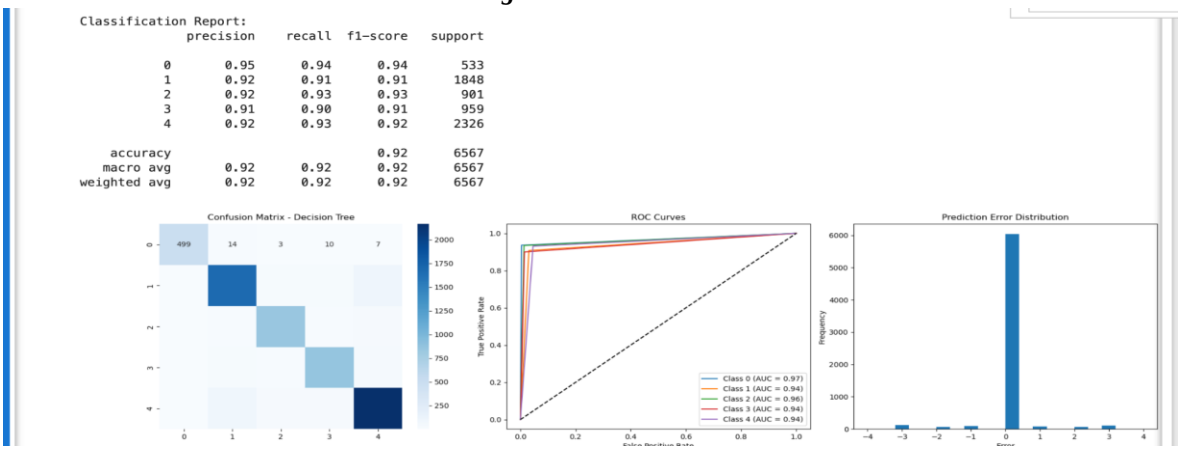


Fig 7: Classification report

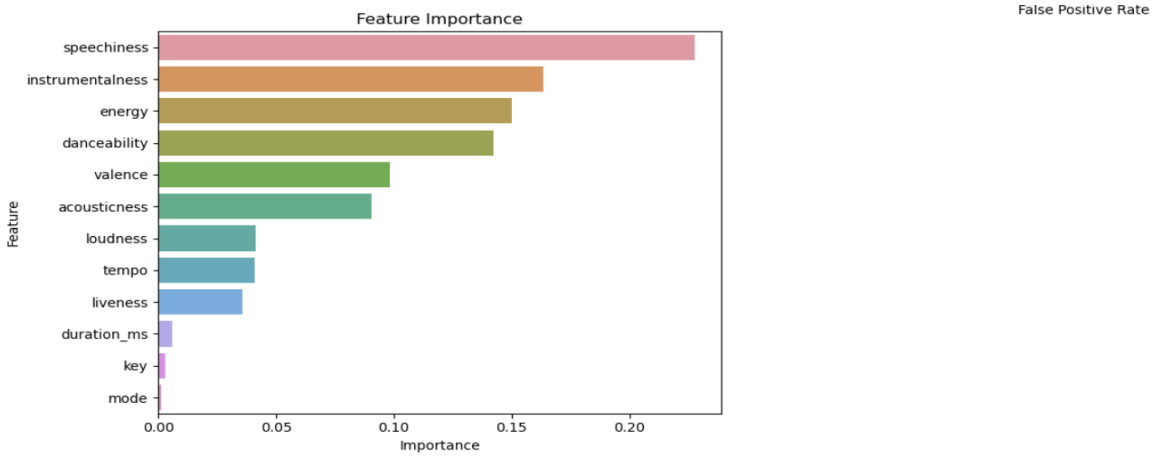
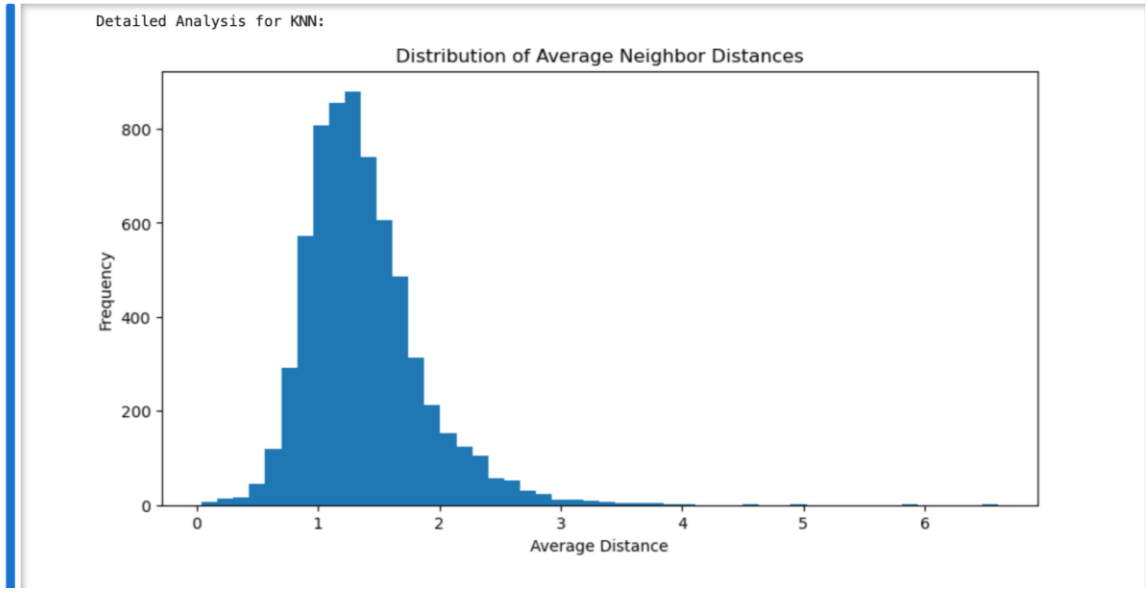


Fig 8: Feature Importance



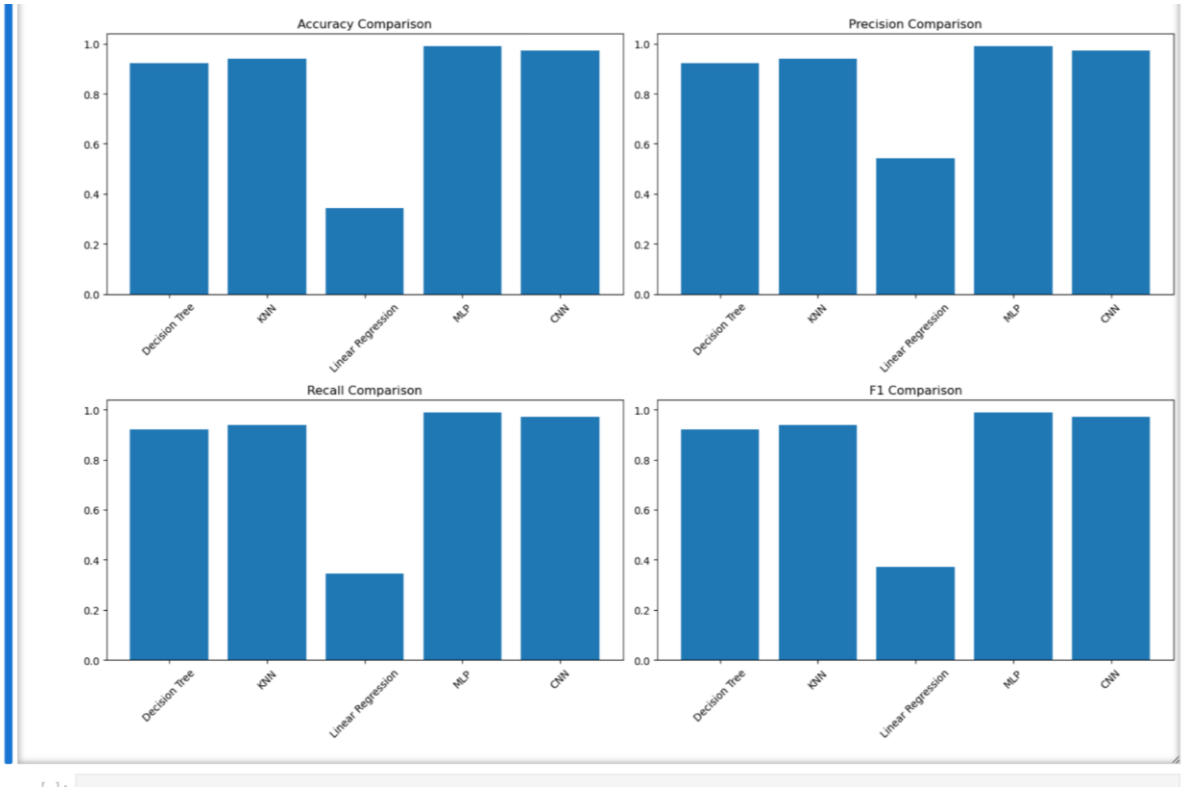


Fig:9 : Detailed analysis

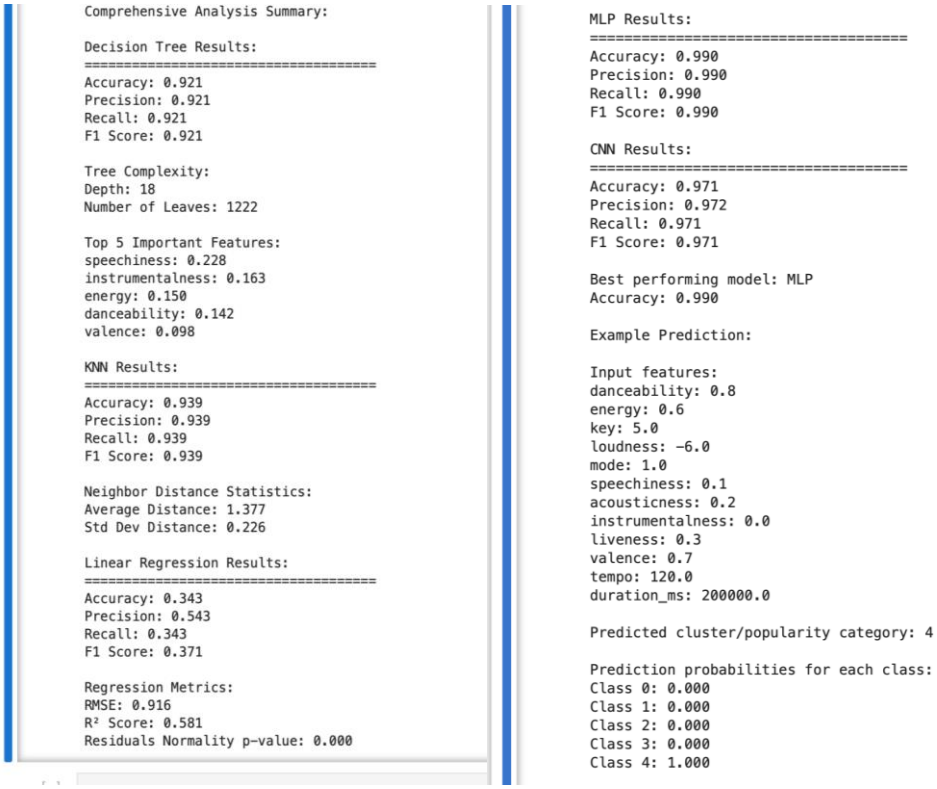


Fig 10: Comprehensive Analysis Summary