**Spotify Songs’ Genre Segmentation**

**Project Report**

**About Dataset**

The Spotify dataset used in this project is rich in audio-related features and metadata, making it ideal for genre classification and clustering tasks.

* **File Information:** The dataset was sourced from a CSV file named spotify.csv and loaded into the project for analysis.
* **Dataset Size:** It includes hundreds of rows representing individual songs and their corresponding features.
* **Features:** Columns such as danceability, energy, tempo, and valence serve as the primary input features for analysis, while the genre column is the target for classification tasks.
* **Initial Checks:** The dataset was thoroughly checked for:
  + **Null Values:** Missing values were detected and handled using appropriate imputation or removal strategies.
  + **Data Types:** Each column's data type was validated to ensure compatibility with machine learning algorithms.  
    This dataset provides an excellent foundation for both exploratory and predictive analysis.

**Structure and Metadata**

The dataset’s structure and metadata were analysed to gain insights into its organization and key properties:

* **Feature Columns:** Key numerical features such as danceability, energy, tempo, and loudness quantify the audio characteristics of each song.
* **Metadata Columns:** These include genre (target label), track\_id, and popularity. These metadata columns add context to the dataset and help in grouping and labelling.
* **Preprocessing Steps:**
  + **Standardization:** To normalize the range of numerical features, StandardScaler was applied, ensuring that all features contribute equally to the models.
  + **Train-Test Split:** The data was divided into 80% for training and 20% for testing, ensuring robust evaluation of model performance.  
    By understanding the dataset's structure and metadata, we ensured a streamlined workflow for further analysis.

**Project Workflow**

**Data Loading and Exploration**

The dataset was loaded using pandas, and its key properties were explored to identify trends, distributions, and potential outliers.

* **Initial Exploration:** Commands like data.head() and data.info() were used to preview the dataset and summarize its structure.
* **Null Values:** A detailed analysis of missing data was conducted, ensuring that the models work with a clean and complete dataset.
* **EDA:** Exploratory Data Analysis revealed patterns and correlations between features, providing insights into how audio characteristics influence song genres.

**Preprocessing**

* **Feature Scaling:** All numerical features were scaled to a standardized range using StandardScaler to improve model convergence.
* **Encoding:** The genre column, being categorical, was encoded into numerical format, making it compatible with machine learning algorithms.

This workflow ensured the data was clean, structured, and ready for modeling.

**Model Architecture**

**KMeans Clustering**

KMeans was implemented as an unsupervised learning approach to segment the songs into clusters based on their features.

* **Cluster Selection:** The optimal number of clusters was identified using the elbow method, ensuring meaningful grouping of songs.
* **Insights:** Clustering revealed trends in audio features, such as grouping songs with similar energy levels or tempos.

**Logistic Regression**

A Logistic Regression model was employed for supervised classification tasks, predicting the genre of each song.

* **Tuning:** Hyperparameters were optimized to achieve the best trade-off between accuracy and generalization.
* **Advantages:** Its simplicity and interpretability made it an ideal choice for this task.

**K-Nearest Neighbours (KNN)**

KNN was evaluated as a baseline model for genre classification.

* **Working:** It classified songs based on the similarity of their audio features to neighbouring songs in the feature space.
* **Limitations:** Performance depended on the choice of k (number of neighbours) and was sensitive to feature scaling, which was addressed through preprocessing.

**Model Compilation**

For the classification tasks, models were compiled with appropriate loss functions and optimizers.

* **Loss Function:** Categorical cross-entropy was used for Logistic Regression, as it is well-suited for multi-class classification problems.
* **Metrics:** Accuracy served as the primary metric to measure model performance, complemented by other metrics such as precision and recall during evaluation.

**Training and Validation**

The models were trained and validated using an 80-20 train-test split, ensuring sufficient data for both training and testing.

* **Training Parameters:** Multiple epochs were used to allow the models to learn feature patterns adequately.
* **Validation Strategy:** During training, performance on the validation set was monitored to prevent overfitting and ensure generalization to unseen data.

**Evaluation**

**Clustering Evaluation**

* KMeans clusters were evaluated for cohesion and separation using metrics like silhouette score.
* Clusters revealed meaningful groupings, such as high-energy or low-tempo songs.

**Classification Evaluation**

* **Logistic Regression:** Achieved a high accuracy rate in predicting genres, demonstrating its effectiveness for this dataset.
* **KNN:** Served as a baseline and performed adequately, but slightly lagged behind Logistic Regression in accuracy.

Performance metrics such as confusion matrices, precision, recall, and F1-scores provided a detailed evaluation of each model’s strengths and weaknesses.

**Key Results**

* Logistic Regression emerged as the best model for genre classification, with high accuracy and robust predictions.
* KMeans clustering provided valuable insights into audio feature-based segmentation, helping to uncover trends and patterns within the dataset.

**Saving the Model**

The final Logistic Regression model was saved for future use, enabling easy deployment in real-world applications like personalized playlist recommendations or music analytics tools.