ABSTRACT

- The Spotify Songs' Genre Segmentation project aims to enhance the music recommendation system provided by Spotify, a popular music streaming platform. The project leverages data analysis techniques and clustering algorithms to group songs based on their auditory properties, ultimately enabling more accurate and personalized music recommendations. The project utilizes a dataset containing information about songs, including their features, popularity, and genre.
- The project begins with data preprocessing operations to handle missing values and convert categorical variables. Exploratory data analysis techniques are then applied to gain insights into the distribution of track popularity, relationships between features, and overall patterns within the dataset. Visualizations, such as histograms and scatter plots, are used to present these findings.
- Clustering analysis is performed using the K-means algorithm to identify distinct groups of songs based on their danceability, energy, loudness, and tempo. The resulting clusters provide valuable information about the characteristics and preferences of different genres.

- A recommendation system is built based on clustering analysis, where a user can input a song ID, and the system identifies the cluster label of that song. Similar songs from the same cluster are then recommended to the user based on their relevance. The recommendations are ranked using a suitable algorithm or similarity metric.
- The project concludes with the evaluation of the recommendation system's performance and suggestions for improvement. Feedback mechanisms and alternative algorithms are discussed as potential enhancements to enhance the accuracy and user satisfaction of the recommendation system.
- Overall, the Spotify Songs' Genre Segmentation project offers a comprehensive approach to understanding and segmenting music genres, resulting in an improved music recommendation system that enhances user experience and engagement on the Spotify platform.

INTRODUCTION

- The Spotify Songs' Genre Segmentation project focuses on enhancing the music recommendation system provided by Spotify, one of the leading music streaming platforms. The goal of the project is to improve the accuracy and relevance of music recommendations by leveraging data analysis techniques and clustering algorithms to segment songs based on their auditory properties.
- Music recommendation systems play a crucial role in providing personalized music experiences to users. These systems analyze user preferences, historical listening patterns, and song features to suggest songs and playlists that align with the user's taste. One key aspect in developing effective recommendation systems is understanding the underlying genres of songs, as genre preferences greatly influence user satisfaction.
- The project utilizes a dataset containing a diverse collection of songs, along with their associated features, such as danceability, energy, loudness, tempo, and more. These features are extracted using advanced audio analysis techniques and provide valuable insights into the acoustic characteristics of each song.
- To begin, the dataset undergoes comprehensive data preprocessing operations to handle missing values and convert categorical variables into

numerical representations. This ensures that the data is in a suitable format for further analysis.

- Exploratory data analysis techniques are then applied to gain a deeper understanding of the dataset. Various visualizations, including histograms, scatter plots, and correlation matrices, are used to explore the distribution of track popularity, identify relationships between different song features, and uncover any underlying patterns or trends.
- The project proceeds with clustering analysis using the popular K-means algorithm. This algorithm groups songs into clusters based on their similarities in terms of danceability, energy, loudness, and tempo. By identifying these distinct clusters, the project aims to uncover genres inherent within the dataset.
- The resulting clusters are then used to build a recommendation system. When a user inputs a song ID, the system identifies the cluster label of that song. Similar songs from the same cluster are then recommended to the user, providing a personalized and genre-specific set of recommendations. The recommendations are ranked using suitable algorithms or similarity metrics to ensure the most relevant suggestions are provided.
- The performance of the recommendation system is evaluated, considering factors such as recommendation accuracy, user satisfaction, and diversity of recommendations. Feedback mechanisms are incorporated to gather user input and improve the system over time. Additionally, potential methods for

further enhancing the recommendation system, such as incorporating user feedback or exploring advanced algorithms, are discussed.

- In conclusion, the Spotify Songs' Genre Segmentation project aims to enhance the music recommendation system by segmenting songs into genres based on their auditory properties. By utilizing data analysis techniques and clustering algorithms, the project contributes to a more accurate and personalized music experience for Spotify users, ultimately improving user satisfaction and engagement on the platform.

Problem Statement:

To develop an automated system that can segment songs into different genres based on their audio properties and recommend similar songs to users based on their preferences. The goal is to utilize the available data on Spotify songs to create a recommendation system that can understand the auditory characteristics of different songs and group them into clusters for genre classification.

Aim:

The aim of the Spotify Songs' Genre Segmentation project is to develop an automated system that can accurately segment songs into different genres based on their audio properties and create a recommendation system that suggests similar songs to users based on their preferences.

DATA DESCRIPTION

The Spotify Songs' Genre Segmentation project utilizes a dataset that contains a collection of songs and their associated features. The dataset provides valuable information about each song, allowing for in-depth analysis and genre segmentation.

The dataset consists of the following key features:

- **track_id**: A unique identifier for each song track.
- **track_name**: The name or title of the song.
- **track_artist**: The artist or group associated with the song.
- **track_popularity**: A measure of the song's popularity, typically ranging from 0 to 100.
- **track_album_id**: A unique identifier for the album containing the song.
- **track_album_name**: The name of the album.
- **track_album_release_date**: The release date of the album.
- **playlist_name**: The name of the playlist to which the song belongs.
- **playlist_id**: A unique identifier for the playlist.
- **playlist_genre**: The genre of the playlist.
- **playlist_subgenre**: The subgenre of the playlist.
- **danceability**: A measure of how suitable the song is for dancing, ranging from 0 to 1.

- **energy**: The energy level of the song, representing intensity and activity, ranging from 0 to 1.
- **key**: The key in which the song is composed (e.g., C major, D minor).
- **loudness**: The loudness of the song in decibels (dB).
- **mode**: Indicates whether the song is in a major or minor key (0 for minor, 1 for major).
- **speechiness**: A measure of the presence of spoken words in the song, ranging from 0 to 1.
- **acousticness**: The acoustic quality of the song, ranging from 0 to 1 (0 for highly electronic, 1 for highly acoustic).
- **instrumentalness**: A measure of whether the song is instrumental or contains vocals, ranging from 0 to 1.
- **liveness**: A measure of the presence of a live audience in the recording, ranging from 0 to 1.
- **valence**: The musical positivity of the song, ranging from 0 to 1 (0 for negative, 1 for positive).
- **tempo**: The tempo of the song in beats per minute (BPM).
- **duration_ms**: The duration of the song in milliseconds.

The dataset provides a diverse range of song features, including both audio characteristics and metadata. These features allow for comprehensive analysis, clustering, and genre segmentation.

Data preprocessing steps are performed to handle any missing values and convert categorical variables into numerical representations. This ensures the dataset is ready for further analysis and model development.

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a crucial step in understanding the dataset and extracting meaningful insights. In the Spotify Songs' Genre Segmentation project, EDA techniques are applied to gain a deeper understanding of the songs dataset and uncover patterns, distributions, and relationships among the features. The following EDA steps were performed:

1. **Data Summary:** The dataset is initially examined by displaying the first few rows using the `head()` function. This provides a glimpse of the data structure and helps in understanding the available features.

2. **Data Types:** The data types of each column are examined using the `dtypes` attribute. This helps identify whether any feature needs to be converted to a different data type for analysis.

3. **Missing Values:** The presence of missing values in the dataset is assessed using the `isnull().sum()` function. This step helps determine if any imputation or removal of missing values is required for further analysis.

4. **Histograms:** A histogram is plotted to visualize the distribution of the 'track_popularity' feature. The `hist()` function with specified bins is used for this purpose. Histograms provide insights into the popularity distribution of songs, highlighting any skewedness or patterns.

5. **Scatter Plots:** A scatter plot is created to explore the relationship between 'danceability' and 'energy' features. The scatter plot is colored by 'playlist_genre' to observe any genre-related patterns. This visualization helps identify potential clusters or separations among songs based on these features.

6. **Correlation Matrix:** A correlation matrix is generated using the `corr()` function to understand the relationships between numerical features. The matrix is then visualized using a heatmap, allowing for the identification of strong positive or negative correlations among features. This step helps in feature selection and understanding feature interactions.

The exploratory data analysis provides valuable insights into the dataset, including the distribution of track popularity, relationships between features, and potential clusters based on auditory properties. These insights serve as a foundation for further analysis and model development in the project.

CLUSTERING ANALYSIS

Clustering analysis is a crucial step in the Spotify Songs' Genre Segmentation project to identify distinct groups of songs based on their auditory properties. This analysis is performed using the K-means algorithm, a popular unsupervised machine learning algorithm. The following steps outline the clustering analysis:

1. **Feature Selection:** The features used for clustering are selected based on their relevance to genre segmentation. In this project, the features selected are 'danceability', 'energy', 'loudness', and 'tempo'. These features capture important aspects of a song's auditory characteristics.

2. **Data Scaling:** Before performing clustering, the dataset is preprocessed by scaling the numerical features. This step ensures that all features are on the same scale and have equal importance during clustering. In this project, the MinMaxScaler from scikit-learn is used to scale the features to a range between 0 and 1.

3. **K-means Clustering:** The K-means algorithm is applied to the preprocessed dataset. The number of clusters, k, is predetermined based on prior knowledge or experimentation. In this project, k is set to 3 as an example. The algorithm aims to minimize the within-cluster sum of squared distances by iteratively assigning data points to the nearest centroid and updating the centroids based on the assigned points.

4. **Cluster Label Assignment:** After performing K-means clustering, each song is assigned a cluster label based on its proximity to the cluster centroids. These cluster labels indicate the segment or genre to which each song belongs.

5. **Visualize Clusters:** The clusters are visualized using a scatter plot, where 'danceability' and 'energy' are plotted on the X and Y axes, respectively. Each point in the scatter plot represents a song, and the color represents the assigned cluster label. This visualization helps identify any distinct clusters or separations among songs based on their auditory properties.

6. **Cluster Characteristics**: The mean values of the numerical features within each cluster are calculated to determine the characteristics of each cluster. This provides insights into the average danceability, energy, loudness, tempo, and other features for each segment or genre.

The clustering analysis helps identify and define different segments or genres based on the auditory properties of songs. These segments serve as the basis for building a recommendation system that suggests similar songs to users based on their preferences and the identified clusters.