**PG-DBDA: September 2022 batch**

**Group no: 11**

**Know-It, CDAC-Pune**

**Genre Classification and Song Data Analysis using Machine Learning**

**INDEX**

|  |  |  |
| --- | --- | --- |
| Sr. No. | Contents | Page No |
| 1 | Project Title and Objective | 3 |
| 2 | About the Dataset | 4 |
| 3 | Content, Usage | 4 |
| 4 | Column Description | 5,6 |
| 5 | Data Collection | 7 |
| 6 | Data Importing and Cleaning steps | 7 |
| 7 | Exploratory Data Analysis: |  |
| 8 | Feature Extraction, Selection, Reduction, Data Transformation |  |
| 9 | Model Selection and building, Comparing different algorithms, Hyperparameter tuning, Cross validation. |  |
| 10 | Comparison of different types of models. Identify the reasons behind the selection of the algorithm, Explore the results in graphical pattern. |  |
| 11 | Model Deployment |  |
|  |  |  |

**Project Title:** Genre Classification and Song Data Analysis using Machine Learning

**Description:** Over the years song has become an integral part of human life. Presenting the music in a way more interactive way enriches the taste of music.

* The objective of this project is to automatically classify songs based on genre and analyzing patterns in the song dataset(Spotify Tracks Dataset).
* The project implements machine learning algorithms to classify the data based on requirement and utilizes Apache Spark to analyze the dataset.
* The project will include tableau to present the analyzed results.

Through this project, we will also measure the songs popularity, top 5 artists, top 5 popular songs, top 5 genres etc.

DATASET: **Spotify Tracks Dataset (20.12 MB)**

**About Dataset:**

**Content:**

This is a dataset of Spotify tracks over a range of 125 different genres. Each track has some audio features associated with it. Although the data is in CSV format which is tabular*, it contained null values, missing values, outliers, some structural errors.* Data engineer or analyst needs to work on it before using this dataset for further use. The collected data ranges over the period form 07/10/2008 to 22/10/2022.

**Usage:**

The dataset can be used for:

* Classification purposes based on audio features and available genres.
* Finding out top performing artists, genres, songs, etc.
* Building a Recommendation System based on some user input or preference(We have not included this in our project)

**Column Description: This dataset contains 22/21 columns out of which 14 are used for classification and analysis in ‘which’ algorithm.**

1. **Index\_id:** It is the serial number column*. ???*
2. **track\_id:** It is the Spotify id for track.
3. **artists:** This column contain names of artists who performed the song.
4. **album\_name:** This column contains name of the album in which the track appears.
5. **track\_name:** This column contains name of the track.
6. **popularity:** This column contains popularity of a track which is a value between 0 and 100, with **100 being the most popular**. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past*. Duplicate tracks (e.g. the same track from a single and an album) are rated independently.* Artist and album popularity is derived mathematically from track popularity.
7. **duration\_ms:** This column shows track length in milliseconds.
8. **explicit:** Whether or not the track has explicit lyrics (true = yes it does; false = no it does not OR unknown)
9. **danceability:** Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable
10. **energy:** Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale.
11. **key:** The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
12. **loudness:** The overall loudness of a track in decibels (dB)
13. **mode:** 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
14. **speechiness:** Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
15. **acousticness:** A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
16. **instrumentalness:** Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.
17. **liveness:** Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
18. **valence:** A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
19. **tempo:** The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
20. **time\_signature:** An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of 3/4, to 7/4.
21. **track\_genre:** The genre in which the track belongs.
22. **spotify\_release\_date:** This column contains the date on which the song was released on the spotify.

**Step 1: Data Collection:**

The data was collected from spotify’s android app. It was downloaded through kaggle.com.

Link: <https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset>

**Step 2: Data importing and cleaning steps:**

* **Null Values Removal:**
* **Missing values:**
* **Outlier Removal:**
* **Structural Error removal:**

**Step 3: Exploratory Data Analysis:**

* Use of tableau to crate 10 graphs
* Listing out names of those 10 graphs and looking for the outcomes through those grpahs.

**Step 4: Feature Extraction, Selection, Reduction, Data Transformation**

* **Feature Extraction**
* **Selection**
* **Reduction**
* **Data Transformation**

**Step 5: Model Selection and building, Comparing different algorithms, Hyperparameter tuning, Cross validation.**

* **Model Selection and building**
* **Comparing different algorithms**
* **Hyperparameter tuning**
* **Cross validation**

**Step 6: Comparison of different types of models. Identify the reasons behind the selection of the algorithm, Explore the results in graphical pattern.**

* **Comparison of different types of models**
* **Reasons behind the selection of the algorithm**
* **Exploring result in the graphical pattern**

**Step 7: Model Deployment:**