Project Report

Genre Classification and Song Data Analysis using Machine Learning

Submitted in partial fulfillment for the award of

Post Graduate Diploma in Big Data Analytics (PG-DBDA),

Know-IT(Pune)

Guided by:

Mrs. Trupti Joshi

Submitted By:

Omkar Pisal (220943025022)
Akshay Patil (220943025024)
Ajinkya Pawar (220943025026)
Tejas Warbhe (220943025050)

CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

This is to certify that

Omkar Pisal (220943025022)
Akshay Patil (220943025024)
Ajinkya Pawar (220943025026)
Tejas Warbhe (220943025050)

Have successfully completed their project on

Genre Classification and Song Data Analysis using Machine Learning

Under the guidance of Mrs. Trupti Joshi

ACKNOWLEDGEMENT

This project **"Genre Classification and Song Data Analysis using Machine Learning"** was a great learning experience for us and we are submitting this work to CDAC ATC Know-IT (Pune).

We all are very glad to mention the name of **Mrs. Trupti Joshi, Mr. Milind Kapse and Mr. Tushar Kute** for their valuable guidance to work on this project. Their guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to **Mr. Shrinivas Jadhav,** Vice-President (Know-IT), **Mrs. Bakul Joshi** (Training Head, PG-DBDA) C-DAC, **Mr. Vaibhav Inamdar**, Placement Head (Know-IT) for their guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS, Pune.

Our most heartfelt thanks go to all the **teaching and non-teaching staff** who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility and extra lab hours to complete the project and throughout the course up to the last day here in C-DAC Know-IT, Pune.

From:

Omkar Pisal (220943025022)
Akshay Patil (220943025024)
Ajinkya Pawar (220943025026)
Tejas Warbhe (220943025050)

TABLE OF CONTENTS

1. **ABSTRACT**
2. **INTRODUCTION**
3. **SYSTEM REQUIREMENTS**
    a. Software Requirements
    b. Hardware Requirements
4. **FUNCTIONAL REQUIREMENTS**
5. **SYSTEM ARCHITECTURE**
6. **METHODOLOGY**
7. **MACHINE LEARNING ALGORITHMS**
8. **DATA VISUALIZATION AND REPRESENTATION**
9. **CONCLUSION AND FUTURE SCOPE**

References

Abstract

Creative art like song has variety of facets. Songs help to express the feelings in words, say the things which can't be expressed in formal words. Song's data contain multiple attributes such as loudness, acoustics, tempo etc. This data can be put to proper use to classify the music. Same data can be used to analyze the songs. Annually there are an estimated 22 million new songs are released which in turn realize an annual revenue of the $25.9 billion worldwide. Considering these things in mind, we thought it would be more exciting if we use this data to classify the songs based on genre and analyze the data to enrich human experience.

INTRODUCTION

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.

- Annually there are an estimated 22 million new songs are released.
- The annual revenue of the global recorded music industry is $25.9 billion as of 2021
- The dataset used in this project is obtained from www.kaggle.com. Data used in this project is unstructured and spans over a period of 14 years from 2008 to 2022. The main goal of the analysis is to be build accurate and robust classification models to classify the Genre of a Song .

Datasets and features:

- This is a dataset of Spotify tracks dataset over a range of 114 different genres
- Total number of rows in this data set are 1,14,000 and these rows are spread over 22 attributes which contains categorical as well as numerical data.
- The data is in CSV format which is tabular and can be loaded quickly. However, the null values, structural errors, no uniformity in data types and outliers are the usual challenges associated with any other dataset.
- Each track has some audio features associated with it. We specifically used 10 to 12 audio features based on the requirement of feature extraction, scaling, analysis and classification part.

SYSTEM REQUIREMENTS

Hardware Requirements:

- RAM – Min 8 GB of RAM
- ROM – At least 120 GB
- Peripheral Devices – Mouse, Keyboard, Monitor
- Internet

Software Requirements:

- Jupyter Notebook(Anaconda Navigator)
- Python 3
- Apache Spark
- Tableau
- **OS – Windows**

FUNCTIONAL REQUIREMENTS

(1) Jupyter Notebook:

- Jupyter Notebook is an open-source, web-based interactive environment.
- It allows you to create and share documents that contain live code, mathematical equations, graphics, maps, plots, visualizations, and narrative text.
- It integrates with many programming languages like Python, PHP, R, C#, etc.

(2) Python 3:

- Python is a general purpose and high level programming language.
- It is used for developing desktop GUI applications, websites and web applications.
- Python allows to focus on core functionality of the application by taking care of commonprogramming tasks.
- Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68,Small Talk, and Unix shell and other scripting languages.

(3) Apache Spark:

Apache Spark is a lightning-fast unified analytics engine for big data and machine learning. It was originally developed at UC Berkeley in 2009.

- Apache spark can use to perform batch processing.
- Apache spark can also use to perform stream processing. For stream processing, we were using Apache Storm / S4.
- Spark is also useful to perform graph processing. Neo4j / Apache Graph was using for graph processing.
- Spark can process the data in real-time and batch mode.

(4) Tableau:

- Data visualization is the graphical representation of information and data.
- It helps create interactive elements like charts, graphs, and maps, data visualization toolsprovide an accessible way to see and understand trends, outliers, and patterns in data.
- Tableau is widely used for Business Intelligence but is not limited to it.
- It helps create interactive graphs and charts in the form of dashboards and worksheets togain business insights.
- All of this is made possible with gestures as simple as drag and drop.

SYSTEM ARCHITECTURE

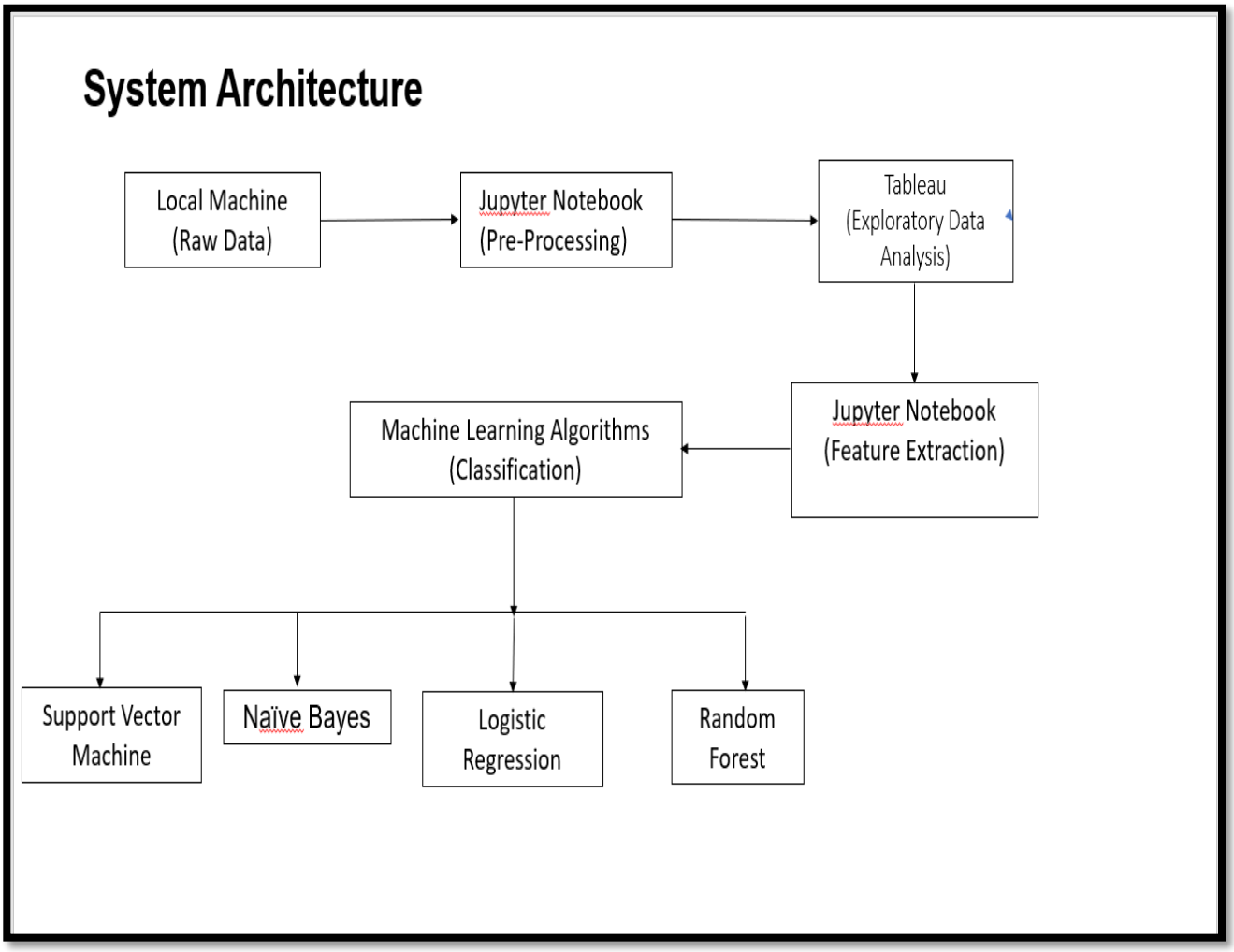

Fig: System Architecture and Work Flow

METHODOLOGY

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:

- **Data importing:** To import the data, following code is used: music = spark.read.csv(r'path\filename.csv')
- **Null Values Removal:**
  - Our dataset contained total 135 null values.
  - We deleted those values by using following command, i.e. music.na.drop(how='any').count()
  - Total rows were, 1,14,000, after deleting the null value rows, 1,13,865 rows remained in our dataset.
- **Outlier Removal:** We used anomaly detection to find out outliers, however, our dataset has 22 attributes. However, features were 114 genre, hence the dataset is bound to have large number of outliers, hence we decided to go ahead with the outliers as it will not hamper the classification of songs based on genre.
- **Structural Error removal:**
  - **Header Name:**
    - The headers in the file by default contained alphanumeric code as header names, we replaced it with exact column names.
    - music= spark.read.option('header','true').csv(r'path\filename.csv')
  - **Data type:**
    - File contained same data type for every column, i.e. even if a column contained numeric values, it was showing its data type as String.
    - music.printSchema(): this code is used to find out the data types.
    - By using typecasting code, data types are changed.

Step 3: Exploratory Data Analysis:

- Use of tableau to create 10 graphs.
- Listing out names of those 10 graphs and looking for the outcomes through those graphs.

Step 4: Feature Extraction, Selection, Scaling

- **Feature Extraction :** 12 out of 22, attributes are extracted for further process of machine learning. 10 Redundant columns are dropped.
- **Selection:** All the 12 extracted feature are required for the classification.
- **Scaling the data:**
    - In scaling we used the Standard scaling.

Step 5: Model Selection and building, comparing different algorithms, Hyperparameter tuning.

- **Model Selection and building:**
    - **Random Forest:** This algorithm is applied using all the attributes.
    - **Logistics Regression:** This algorithm is applied using all the attributes.
    - **Naïve Bayes** This algorithm is applied using all the attributes.
    - **SVM Classifier:** This algorithm is applied using all the attributes.
- **Comparing different algorithms:**
    - **Random Forest:** This algorithm accrued the 82.94% accuracy.
    - **Logistic Regression:** This algorithm accrued the 75.65% accuracy.
    - **Naïve Bayes:** This algorithm did not accrue any accuracy as Spotify dataset contained negative numeric values besides positive and this algorithm does not work with negative values.
    - **SVM Classifier:** This algorithm did not yield result even after waiting for around 3.5 hours.
- **Hyperparameter tuning:** Test and Train data using different algorithms.

Step 6: Selecting the most suitable algorithm:

- This project classifies the genre of songs based on different attributes with more pace as it uses a random forest algorithm. The accuracy of using a random forest algorithm is 82.94%, which overcomes the drawback of lack of accuracy in other algorithm.

MACHINE LEARNING ALGORITHMS

In this project we applied various different types of Classification and Regression Algorithms such as Logistics Regression, Random Forest etc. During the implementation we analyzed the accuracy of all the algorithms.

Machine learning is the research that explores the development of algorithms that can learn from data and provide classifications based on it. The methods commonly used include SVM, Naïve Bayes, Random forests, Logistics Regression. They were mainly used for classification and prediction. In this project we use various machine learning algorithms which are as follows:

1. Support Vector Machine (SVM)

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

Pros:

- SVM works relatively well when there is a clear margin of separation between classes.
- SVM is more effective in high dimensional spaces.

Cons:

- SVM algorithm is not suitable for large data sets. It works too slow on large datasets. For instance, wrt to our dataset it more than 3 hours but did not yield any result.
- SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.

2. Naïve Bayes:

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the “naive” assumption of conditional independence between every pair of features given the value of the class variable.

Pros:

- When assumption of independent predictors holds true, a Naive Bayes classifier performs better as compared to other models.
- Naive Bayes requires a small amount of training data to estimate the test data. So, the training period is less.

Cons:

- If categorical variable has a category in test data set, which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as Zero Frequency.
- Main imitation of Naive Bayes is the assumption of independent predictors. Naive Bayes implicitly assumes that all the attributes are mutually independent. In real life, it is almost impossible that we get a set of predictors which are completely independent

3. Random Forest:

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

Pros:

- The predictive performance can compete with the best supervised learningalgorithms.
- They provide a reliable feature importance estimate
- They offer efficient estimates of the test error without incurring the cost of repeatedmodel training associated with cross-validation

Cons:

- An ensemble model is inherently less interpretable than an individual decision tree
- Training a large number of deep trees can have high computational costs (but can beparallelized) and use a lot of memory.

4. Logistic Regression:

Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of certain classes based on some dependent variables. In short, the logistic regression model computes a sum of the input features (in most cases, there is a bias term), and calculates the logistic of the result.

Pros:

- Logistic regression is easier to implement, interpret, and very efficient to train.
- It makes no assumptions about distributions of classes in feature space.
- It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.
- It not only provides a measure of how appropriate a predictor (coefficient size) is, but also its direction of association (positive or negative).

Cons:

- The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variable.
- Logistic Regression requires average or no multicollinearity between independent variables.

DATA VISUALIZATION AND REPRESENTATION

Figure 1: Total Number of Genre in Word Cloud (2008-2022)

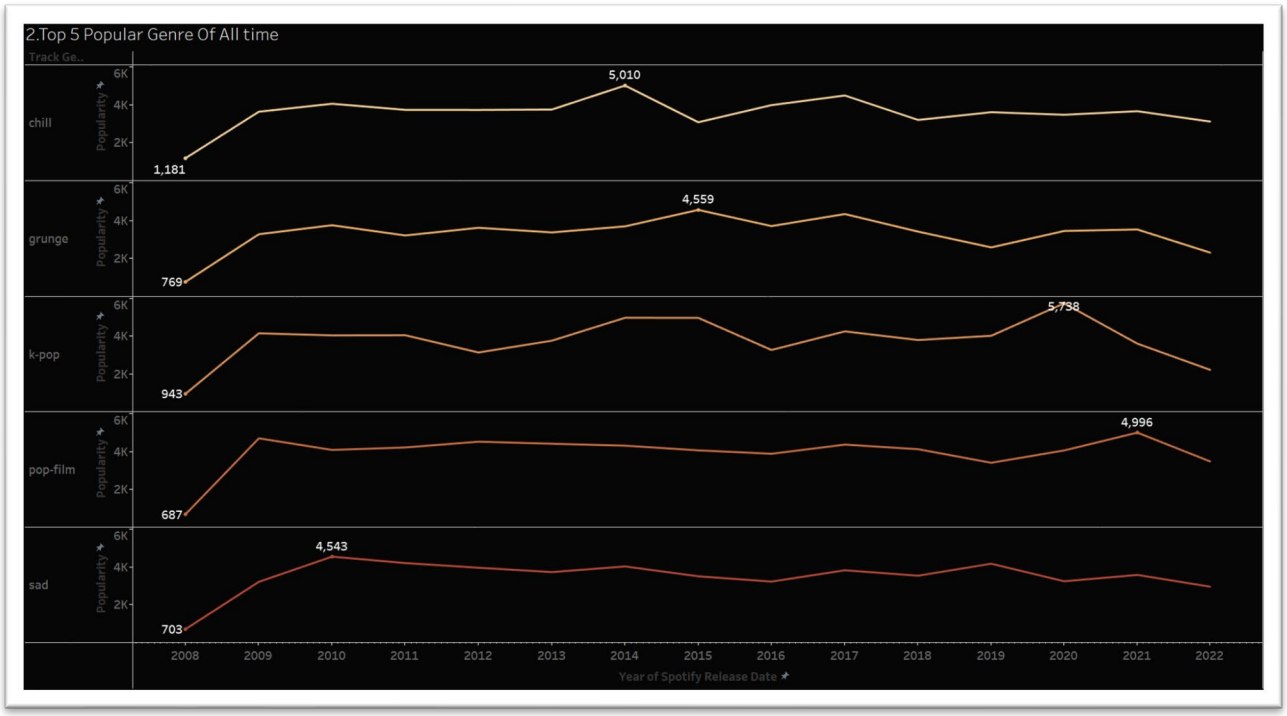

Figure 2: Top 5 Genre (2008-2022).

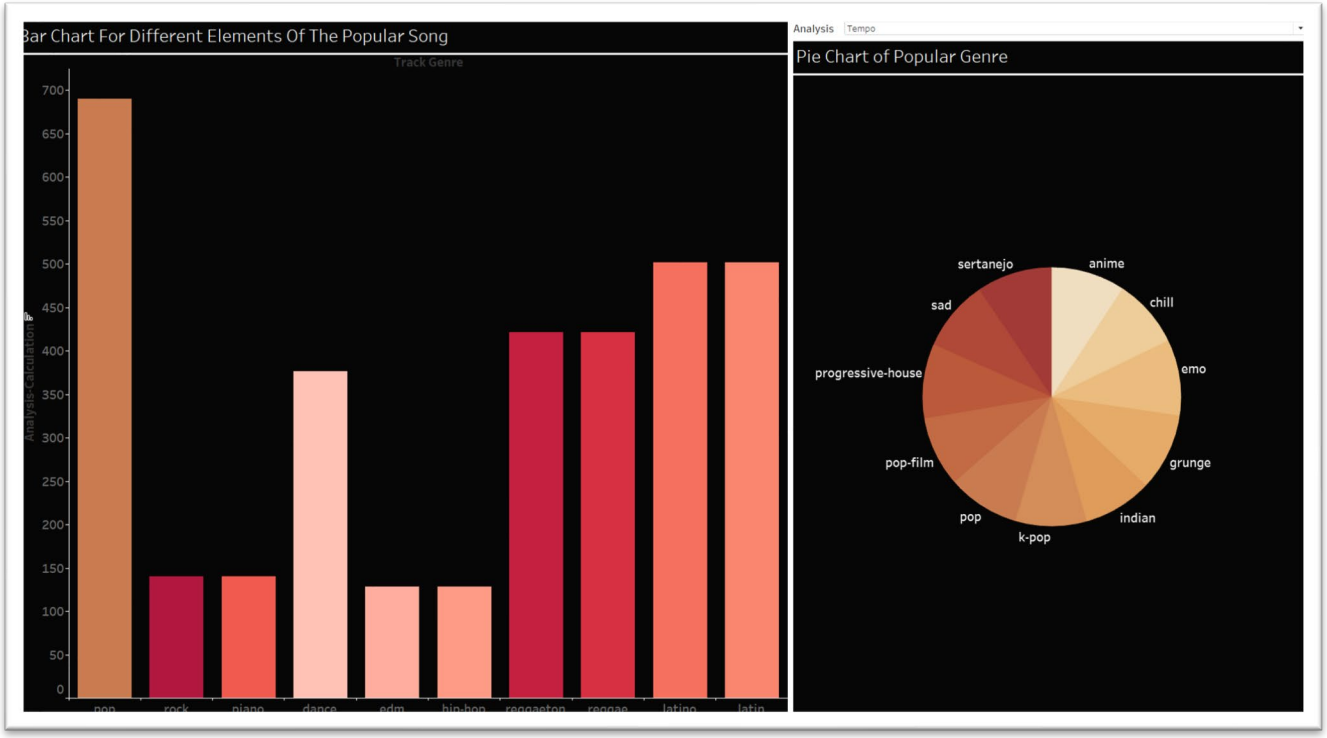

Figure 3: Genre-wise Elemental Distribution

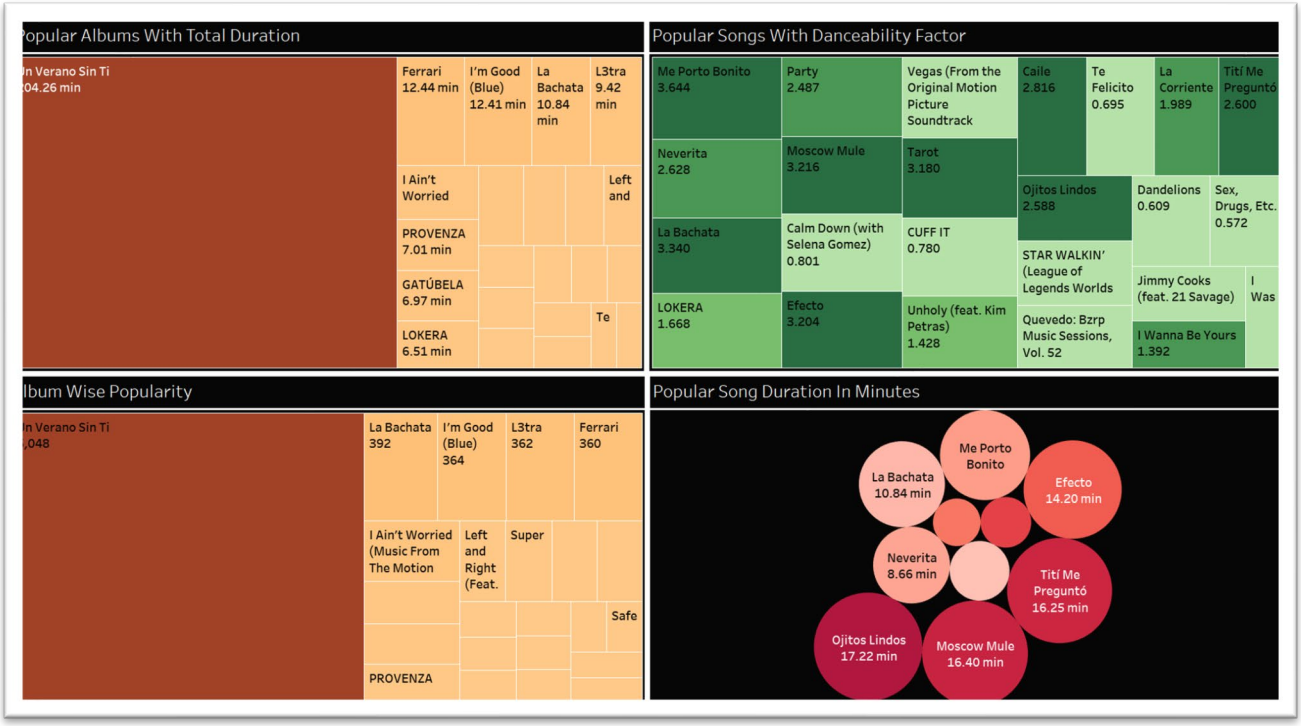

Figure 4: Element wise Popularity

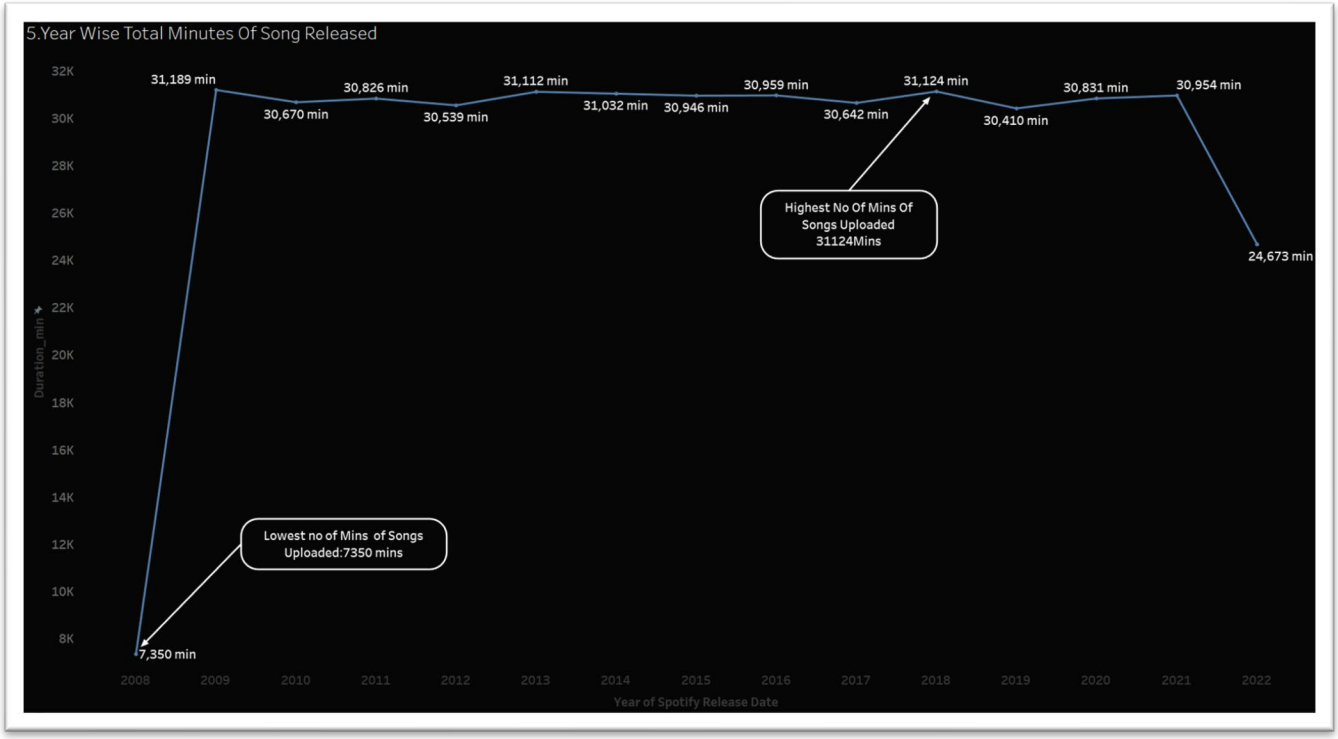

Figure 5: Year wise total number of minutes of songs uploaded (2008-2022).

CONCLUSION AND FUTURE SCOPE

To improve accuracy, optimal attribute selection is used, which in turn helps to speed up the process. Song data analysis helps to understand top performing genres based on different attributes, which in turn will help to showcase the top genre-based song according to a variety of features.

This project classifies the genre of songs based on different attributes with more pace as it uses a random forest algorithm. The accuracy of using a random forest algorithm is 82.94%, which overcomes the drawback of lack of accuracy in other algorithm.

References

1. https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset.
2. https://ieeexplore.ieee.org/document/9320778.
3. (114) Tutorial 1-Pyspark With Python-Pyspark Introduction and Installation – YouTube.
4. Breiman, L., “Random forests,” Machine learning, Vol. 45, No. 1, 2001, pp. 5–32.
5. Machine Learning Using Python By Manaranjan Pradhan And U Dinesh Kumar (www.wiley.com).