**CSCE 5310 - Methods in Empirical Analysis**

**MOST STREAMED SPOTIFY SONG ANALYSIS USING MACHINE LEARNING ALGORITHMS AND JUPYTER NOTEBOOK**

**Team Details:**

1. Nelapati Manideep - 11697590
2. Tarun Preetham Chintada - 11690923
3. Tharun Ramula - 11706360
4. Siva Kishore Reddy Putluru – 11684843

**Goals and Objectives:**

1. **Motivation:**

In this research proposal the problem is related to the project about most streamed songs on Spotify with the usage of machine learning algorithms and jupyter notebooks. The research problem about this topic is detections of songs capacity, detection of music audio data through Spotify and features of acoustics extractions from song databases. This analysis of the problem is important because it will be helping to determine the research topic deeply about the issues faced on Spotify and songs databases. On the other hand, this problem analysis will be helping to predict audio related sets of characteristics for predictions of low and high sounds through machine learning language.

1. **Significance:**

This research is focused on analyzing the most streamed Spotify song by using a machine learning algorithm and Jupiter Notebook. In this context, this research has included the first step which is the data collection. In this context, the official API of Spotify is used with the help of identifying the data set followed by collecting reliable information about the data that includes title artist release date streaming count, and many others. The interpretation of the results is also done industries followed by involving the feature importance analysis to understand the audio feature which is most significant in determining the song's popularity. The research has also illustrated data visualization which has been helpful for understanding the data set and it can be helpful to identify the pattern and friends in the Spotify data analysis has also been found to be very helpful for the purpose of understanding the data as well as the trend in the data set. It has also included different data distributions and summary statistics. In this research, the structure of data is shown after the dropping of some objects in the data set. Information and used for the purpose of cleaning the data set followed by mapping the major and minor data However, the analysis also includes the checking of null values in the data set with the help of using the formula is null. However, the sum of the null values is also identified so that the overall quality of the result can be done. However, this research also included the use of a module for the purpose of configuring the model of logistic regression followed by splitting the data into train and test.

1. **Objectives:**

* The main aim of the project is to investigate the most streamed songs on the Spotify using proper ML approach.
* To determine the factors that define the characteristics of a most streamed song.
* To implement the audio-based approach for identifying the audio features of the songs on Spotify.
* To identify the acoustic features comparison for determining the hyperparameter optimization as well as song predictions on Spotify

1. **Features:**

* Data Collection
* Data preprocessing
* Data Analysis
* Machine learning models (Logistic Regression and Decision Tree)
* Statistical Tests

# Introduction:

Spotify is one of the best music platforms that is used to provide a list of worldwide songs to all users. A list of the songs is displayed to the user by this platform online. This is an online song-providing platform. There are varieties of songs on Spotify which has multiple categories such as English, Hindi, and so on. A variety of songs with multiple categories are provided to the user. Users can choose any option which they like most and play the song through the internet. The listed song assists in understanding the choice of a user. Various user has various choices which highlights the investigation of the streaming of the song. Online streaming on Spotify assists to understand which is the most streamed song in a year. Multiple secondary resources are used to collect the song data with their streaming value. This assists in investigating the streaming of the song. The data investigation approach is used to investigate the most streamed song from a list of Spotify songs. The dataset contains the details of the song with song name, artist name, streaming value, and many other factors. The main functionality of the project defines the overall investigation of the collected data. This assists in understanding the most popular songs on Spotify. The ML approaches define the use of some model creation process such as “Linear regression” (LR) which is used to investigate the Spotify data.

**Background:**

As of now, we've noticed a problem with some online music streaming services: the suggested song selections frequently deviate from the intended tempo continuity, negatively affecting the user experience. Our approach makes use of statistical analysis and predictive modeling grounded in past observations to tackle this problem. With the use of sophisticated statistical models and techniques, we hope to correctly identify trends in song tempos and user preferences. We can improve and optimize song recommendations because of our data-driven approach, which guarantees the accuracy of our analysis. Our objective is to improve the accuracy of tempo-aligned suggestions and give users a seamless and pleasurable listening experience through ongoing iteration and adaptation to user feedback.

**Dataset:**

A screenshot of a computer

Description automatically generated

The Kaggle dataset will be utilized to develop a machine learning model for the analysis and prediction of the most popular songs on Spotify. More than 130 songs and 17 features for obtaining metadata are included in this dataset.

The Kaggle dataset can be used to discuss the specifics of the data source, such as the need to first select New Notebook and then the option to add data with input and output options. The file from the Kaggle dataset is in the CSV (Comma Separated Value) file format and is roughly 48 kilobytes in size. Yes, pre-processing steps exist; the Python Anaconda terminal is used to run the codes.

Thus, the first step how raw data looks like and then it must ensure that it is in usable format. In the second step, data duplicity needs to be checked and can be overcome through unique id features. There are five values in total throughout the entire dataset. There are a number of features in those features that are used to construct a predictive model that finds the inaccurate result. In the dataset, we have column variables like the artist\_count, in\_spotify\_playlists, in\_spotify\_charts, in\_apple\_playlists, released\_day, track\_name, artist(s)\_name, released\_year, released\_month, streams and released\_day.

**Analysis and Implementation:**

The machine learning model for prediction and dataset analysis of the most streamed Spotify songs will be used as the metric in this research proposal for evaluating performances with method feature filter selection and logistics regression (Khan, 2023). With the aid of Python and machine learning algorithms, this research proposal's methodology includes filter selection and logistics regression. Thus, Spotify data can be used to achieve an aim based on findings regarding inferring exploratory analysis. We will talk about the selected machine learning metrics as a model to deliver high popularity-based data with Spotify algorithms and popularity genre.

Any quantitative experiments, aside from the evaluation of metrics, rely on music data and its imputed dataset, which includes targeted columns for music popularity. As a result, it can be discovered that 33% of the music data are test data and roughly 66% of the data are for training.

Popular and unpopular songs on the Spotify music dataset exhibit instability, according to qualitative experiments conducted on the dataset. As a result, the experiment's machine learning algorithms are impacted. The make\_calssification function could be used to collect information about coordinated strategies for averting imbalance.

A screen shot of a computer code

Description automatically generated

**Figure 1: Module Import**

The module importation section is used to import necessary Python modules that assist in creating the necessary environment for the investigation. There are multiple modules are used for the investigation such as ‘warnings’ which are used to remove the warning message from Python code execution. On the other hand, ‘numpy’ is used to implement numeric functionality. On the other hand, ‘pandas’ is used to implement the read functionality of the collected data (Beesa *et al.,* 2023). The visualization method modules of the Python coding are also used in this section which is used in the investigation such as ‘matplotlib’, ‘plotly’, and ‘seaborn’. The standard calling functionality can be implemented by using the ‘sklear’ method.

A screenshot of a computer

Description automatically generated

**Figure 2: Read data and create the structure of the data**

(Source: Own Creation)

The reading of the data is the process of reading the collected data with the help of ‘pandas’. The method defines the use of the ‘pd.read\_csv’ functionality. It defines the process of execution of the read data functionality. The read functionality is used for each line of the collected dataset. Here the format of the dataset is ‘csv’. The encoding process is used to encode the dataset to collect the contents of the dataset. Here the encoding process code is ‘cp775’. The ‘head’ functionality is used to define the structure of the dataset. This functionality assists in constructing the details of the collected data.

A screenshot of a computer

Description automatically generated

**Figure 3: Information of the data**

(Source: Own Creation)

The information in the data defines the details of the dataset. This functionality demonstrates the details of the dataset which defines column name, null value count, non-null value count, and the exact type of the data. The type of data defines ‘int64’, and ‘object’. The value defines the count of the integer type data which is 18, and the object type data whose count is 6. The functionality of the ‘info()’ method is demonstrated in this section (Panda *et al.,* 2021). This analysis in the overall investigation and understanding of which types of data can be used for the investigation.

A screenshot of a computer screen

Description automatically generated

**Figure 4: Checking of null**

(Source: Own Creation)

General values are checked by using the formula that has been shown in the above figure followed by providing the information about the sum of null values. It has been found that in\_shazam\_charts and key are found to be having some null values in the data the type of all the variables in the data set are found to be in the form of integers.

A screenshot of a computer screen

Description automatically generated

**Figure 5: Total playlist concerning the key of the sings**

(Source: Own Creation)

The total playlist with respect to the key of the songs is shown in the above figure which represents the Spotify playlist with having different keys. In this context almost 70% of playlists whereas A hash and B hash have a lower number of songs. However, C hash is also near 90% whereas the other keys are also found to be demonstrating the list of songs in Spotify. It has also been identified that the above formula is found to be showing in the above section (Chodos, 2019). The above figure also says the playlist of Spotify and the bar plot related to its data that can be helpful for the purpose of analyzing the total playlist with respect to the key of the songs.

A screenshot of a computer screen

Description automatically generated

**Figure 6: Total playlist with respect to mode**

(Source: Own Creation)

The total playlist with respect to mode is also demonstrated in the figure where the major more consists of 61.4% of the playlist and the minor more consists of 38.6% of the total data set.

A screen shot of a graph

Description automatically generated

**Figure 7: Total sum of the danceability of the song with respect to the year**

(Source: Own Creation)

The total sum of the danceability of the song with respect to the ear is also demonstrated in the above figure where the danceability percentage and released year are shown in the above figure according to which the danceability percentage is 27.623 cases and in the release year 2022.

A graph with a line

Description automatically generated

**Figure 8: Total sum of the speechiness percentage song with respect to key**

(Source: Own Creation)

The total sum of speechless percentage songs with respect to the key is shown in the above figure pair the speechless percentage is 1448 for key C#. The other percentage is also shown in the above figure as per the graph illustration.

A pie chart with numbers and a few words

Description automatically generated with medium confidence

**Figure 9: Most Stream Song Analysis**

(Source: Own Creation)

The most streamed song analysis is shown in the above figure pair the most streamlined song is C which is equivalent to 14% and the G# is equivalent to 10.6%. F# is 8.51% and E is equal to 7.23.

A blue and red pie chart

Description automatically generated

**Figure 10: Most Stream Song Analysis with respect to song modes**

(Source: Own Creation)

The most streamed song analysis with respect to the song is also shown in the above figure according to which the major song is equal to 57.7% whereas minor songs are equal to 42.3%.



**Figure 11: Drop object columns**

(Source: Own Creation)

The drop object columns are shown in the above figure where the different objects in the column are dropped such as the artist’s name track name and others. However, the columns are used for dropping the objects for the purpose of enhancing the quality of the overall data after the analysis (Werner, 2020).

A screenshot of a screen

Description automatically generated

**Figure 12: Structure of the data after drop column**

(Source: Own Creation)

The structure of the data after the drop column is shown in a figure that has been found to demonstrate the artist count released year released month and many others. As per the above figure, the structure of data can be visualized after the drop of the column.

A screenshot of a computer

Description automatically generated

**Figure 13: Information of cleaned data**

(Source: Own Creation)

The information of clean the data found is demonstrated in the above figure in which the memory usage is shown along with providing the non-null count value for the data type are also shown of all data followed by demonstrating the information of the attributes that are used in the data set (Kalustian, and Ruth, 2021).



**Figure 14: Conversion of an object column**

(Source: Own Creation)

The conversion of the object column is also illustrated in the figure which demonstrates the major and minor objects of the data set.

A screenshot of a computer

Description automatically generated

**Figure 15: Details of the data**

(Source: Own Creation)

The details of the data are shown in the above figure which shows the normal count values along with demonstrating the data type and the memory usage. The attributes and the total entries are also found to be shown in the above figure with the help of that that can be understood in an appropriate manner.

A screenshot of a computer screen

Description automatically generated

**Figure 16: Checking of null**

(Source: Own Creation)

The null values are checked by using the above figure in which the identification of null values is done followed by providing the data type (Al-Beitawi *et al.,* 2020).

A white rectangular object with black text

Description automatically generated

**Figure 17: Use modules for model configuration**

(Source: Own Creation)

The modules are used for the models along with demonstrating the importing of logistic regression and train-test split. The classification report and the accuracy score are also the modules that have been imported for the purpose of getting the appropriate data related to it.

A white rectangular object with text

Description automatically generated

**Figure 18: Setting of X and y**

(Source: Own Creation)

The setting of X and Y data is also in the above figure followed by providing the drop of different values. However, the Spotify data are also used for the purpose of dropping the mode from the axis.



**Figure 19: Split of data**

(Source: Own Creation)

The speed of data is done in the above figure according to which the data are split into train and test taking the test size to be 0.2 and the random state be 42. According to this information, data has been found to be demonstrating accurate value and analysis for the data set.

A screenshot of a computer code

Description automatically generated

**Figure 20: Logistic Regression (LR)**

(Source: Own Creation)

The logistic regression is implemented in the data set for getting the train and test value in which the logistic regression is used for the purpose of splitting the data set as well as creating the model.



**Figure 21: Prediction implementation**

(Source: Own Creation)

The prediction of implementation is done in the above figure that provides the prediction figure X test data.

A white box with black text

Description automatically generated

**Figure 22: LR accuracy**

(Source: Own Creation)

The accuracy of the logistic regression is shown in the above figure which provides the value to be 57%.

A screenshot of a computer

Description automatically generated

**Figure 23: Classification report**

(Source: Own Creation)

The classification report is shown in the above figure in which the precision value is seen to be zero for zero and .57 for 1. Macro average value and weighted average value are also shown above. It has also been found that the recall value is also shown. The F1 score values of the accuracy, macro average, and weighted average are also shown in the above figure.

**Preliminary Results:**

A screenshot of a computer code

Description automatically generated



A white box with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

The logistic regression is implemented in the data set for getting the train and test value in which the logistic regression is used for the purpose of splitting the data set as well as creating the model.

The prediction of implementation is done in the above figure that provides the prediction figure X test data. The accuracy of the logistic regression is shown in the above figure which provides the value to be 57%. The classification report is shown in the above figure in which the precision value is seen to be zero for zero and .57 for 1. Macro average value and weighted average value are also shown above. It has also been found that the recall value is also shown. The F1 score values of the accuracy, macro average, and weighted average are also shown in the above figure.

**Project Management:**

1. **Work Completed**

* **Description**

We preprocessed and cleaned the data, presented the results using data visualization techniques, and used a variety of models like Logistic Regression to estimate how accurate the outcomes would be. We also performed several statistical tests to look at sample means for the whole population.

* **Responsibility**

Data preprocessing – Manideep Nelapati

Data cleaning & visualization – Tarun Preetham Chintada

Data modeling – Tharun Ramula

Statistical approach – Siva Kishore Reddy

* **Contribution**

Nelapati Manideep – 25%

Tarun Preetham Chintada – 25%

Tharun Ramula – 25%

Siva Kishore Reddy Putluru – 25%

1. **Work To Be Completed**

* **Description**

We will use the Decision Tree algorithm to conduct a few more statistical analyses in the final section. It is a supervised learning algorithm that can handle regression and classification tasks. Additionally, it is a tree structure with leaf, internal, and root nodes. Additionally, we'll use a variety of visualization techniques for data visualization.

* **Responsibility**

Data preprocessing – Manideep Nelapati

Data cleaning & visualization – Tarun Preetham Chintada

Data modeling – Tharun Ramula

Statistical approach – Siva Kishore Reddy

* **Issues**

The technical challenges associated with the research topic of Spotify music streaming and its analysis using machine learning algorithms via Jupyter notebooks. Acoustics features comparison, engineering, and hyperparameter optimization, as well as attribute-based Spotify audio features and song prediction from the Spotify music database are examples of technical issues that can be discussed.

This research project can employ correlation values and machine learning models to evaluate efficient music search methods. The most searched songs are found using logistic regression and feature filter selection in the context of the most streamed songs on Spotify. The method for improving music evaluation through machine learning and correlation coefficients.

**References:**

Al-Beitawi, Z., Salehan, M. and Zhang, S., 2020. What makes a song trend? Cluster analysis of musical attributes for Spotify top trending songs. Journal of Marketing Development and Competitiveness, 14(3), pp.79-91.

Álvarez, P., García de Quirós, J. and Baldassarri, S., 2023. RIADA: A Machine-Learning Based Infrastructure for Recognising the Emotions of Spotify Songs.

Araujo, C.V.S., De Cristo, M.A.P. and Giusti, R., 2019, December. Predicting music popularity using music charts. In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) (pp. 859-864). IEEE.

Beesa, P., Naregavi, V., Imandar, J. and Thatte, S., 2023. Songs Popularity Analysis Using Spotify Data: An exploratory study. Vidhyayana-An International Multidisciplinary Peer-Reviewed E-Journal-ISSN 2454-8596, 8(si7), pp.211-223.

Chodos, A.T., 2019. What does music mean to Spotify? An essay on musical significance in the era of digital curation. INSAM Journal of Contemporary Music, Art and Technology, 1(2), pp.36-64.

Dawson Jr, C.E., Mann, S., Roske, E. and Vasseur, G., 2021. Spotify: You have a Hit!. SMU Data Science Review, 5(3), p.9.