Regional Music Popularity Prediction

CSE 6242-A

Fall 2023

Team 32

Team Members

Dennis Frank

Chloe Saleh

Snigdha Verma

Ananya Sharma

Sanjana Daté

Team Liaison

Ananya Sharma

Under the Guidance Of

Prof. Duen Horng (Polo) Chau

Introduction

Music has always been a big part of human culture. In an industry where future trends can have massive economic implications, there's an evident gap: the ability to forecast a song's potential popularity before its release and the associated features leading to it.

Problem Definition

By analyzing sound characteristics such as rhythm, melody, and tempo from the Million Songs Dataset, our objective is to identify songs with the potential to become the next big hit for every region. Our project leverages machine learning models, particularly Random Forests to do so. In parallel with this computational approach, the visualization segment aims to provide an interactive user experience, offering a comprehensive view of the predicted most popular song and the most important acoustic features across various regions.

Literature Survey

[1] Using the Spotify API, this paper compares ML algorithms including logistic regression, random forest, and KNN for music popularity feature selection. This will act as a benchmark for relevant music popularity features and algorithms, though we can expand on this with a different data set.

[2] This paper also compares various ML algorithms including random forest, KNN, and linear support vector to predict music popularity and its most relevant features.It uses the Spotify API and provides a good baseline for metrics we can use to measure our popularity algorithms.

[3] This paper utilizes MIDI - an electronic music information format - to predict music popularity. We could extend applications of MIDI in this paper to more low-fidelity whole-song music information of songs available in our database such as segment-related features.

[4] This paper approaches music popularity from a different perspective determining that the early adoption of trends - including an explored music dataset - indicates future popularity. We can see if we can discover trends from our dataset of what might accelerate this initial rapid spread of a trend.

[5] This research focuses on the unique interface between musicology and visualization research. It classifies 129 related works according to the visualized data types and analyzes which visualization techniques were applied for certain research inquiries and to fulfill specific tasks.

[6] This paper evaluates different classification and regression algorithms on their ability to predict popularity and determines the types of features that hold the most predictive power.

[7] Although in the field of health care, this study provides empirical evidence of how domain experts utilize map-based data visualization for generating insights into vitality with respect to health-related concepts. The results also provide guidelines for designing map-based data visualizations that support the decision-making process across various domain experts in the field of vitality.

[8] The study aims to use ML algorithms to determine song popularity based on its audio features. Out of the three models used, Random Forest model achieved highest accuracy.There are five audio features most correlated with song popularity; we can extend the method to other features as the paper suggests that including additional metadata about artists and tracks could enhance prediction accuracy.

[9] The paper explores the features used to predict song popularity on Spotify. It offers valuable information in understanding musical attributes and how they relate to song preferences. However it does not provide an in-depth analysis of the relationships between these features and song popularity, and suggests using machine learning and deep learning techniques for further improvements.

[10] This paper develops a methodology for predicting song’s appearance on Spotify's Top 50 Global ranking two months in advance. The authors employ ML classifiers using historical information from the platform's ranking and acoustic features of the songs. The model using SVM classifier with an RBF kernel achieves good accuracy. Future work could explore using social network data to improve accuracy and further reduce false positives in predictions.

[11] The paper analyzes various metrics for music popularity using acoustic data. It helps in understanding comprehensive music popularity metrics like *debut, max, mean, standard deviation, length, sum, skewness, and kurtosis* derived from chart rankings, essential for predicting music popularity. The

focus on a limited set of metrics and the use of acoustic data alone are limitations. To enhance our model, we will explore additional parameters and features.

[12] This paper explores the possibility to classify songs as hit or non-hit based on audio features and to identify suitable algorithms for this task. It provides specifications of audio features obtained via the Spotify API and visualizes trends, aiding in understanding music popularity dynamics. It establishes the limitation of relying solely on audio features for hit predictions. Our project aims to predict future popularity with better ML models and potentially additional factors beyond audio features.

[13] The paper introduces two new datasets (HSP-S and HSP-L) to overcome limitations in existing datasets for song popularity prediction. These datasets are sourced from AcousticBrainz, Billboard Hot 100, and the Million Song Dataset with rich audio features and play-count data. The datasets provide a broader set of features to evaluate song popularity, crucial for building a robust prediction model.

[14] This paper talks about data visualization between various features of songs using available libraries of python. This will inspire our statistical visualization from this paper. The paper mentions using data from Spotify and Billboard charts from the 2010s. The data may not be representative of all music genres, time periods, or regions, which could limit the generalizability of the findings.

[15] This paper incorporates D3 with Ajax to create an interactive map of air quality data. We can do the same as we are doing a location-based prediction of music popularity. The paper describes the visualization platform, but it doesn't delve into user interface design considerations or accessibility features. These aspects are crucial for ensuring that users can effectively interact with the platform.

[16] This paper focuses on predicting the mood of a song using its audio features. We learn how to evaluate relevant features using the Sklearn module of python, and rank them according to their importance, as well as their implementation of logistic regression. The paper does not delve into the interpretability of the specific features that contribute to mood prediction. Understanding which features are most influential could provide valuable insights.

[17] This paper focuses on finding the time varying hit songs based on acoustic features. While the accuracy in this paper is low based on solely acoustic features, it can help us determine how long our data can remain relevant to predicting popularity. We can limit our data input for the previous decade as an improvement on top of this paper. The dataset used in the research consists only of hit songs which limits the generalizability of the findings.

Proposed Method

I. Intuition

The current state-of-the-art models in song popularity prediction involve fitting basic regression techniques like Random Forest Regression to predict a song's popularity. As far as we know, there are no widely recognized models specifically designed to predict song popularity on a per-region basis. This area, however, presents a promising avenue, as regional preferences and cultural influences can significantly impact song popularity and listener trends. Additionally, there has been no attempt to combine the predicted model of song popularity per region and integrate it into a visualization component, such as an interactive map.

II. Approach: Innovation

A. Predicting popularity on a regional basis instead of general popularity, to account for regional nuances in acoustic features and their impact.
B. Evaluation of XGBoost Model for this problem, which has not been done before.
C. Correlation of acoustic features and their impact on popularity.
D. Visualization of acoustic feature correlation using heat map.
E. Use of the Cleveland dot plot to show how similar the acoustic features are between the actual most popular song and predicted most popular song.

III. Computation

We firstly gathered our dataset containing different audio features, available marketplaces, and popularity for songs. After cleaning the data, we selected the most relevant features identified based on our literature survey. These were acousticness, danceability, duration, energy, instrumentalness, liveness, loudness, speechiness, tempo and valence.

We then ran and evaluated 6 models on the dataset, namely: K Neighbors, Decision Trees, Logistic Regression, Random Forests, Linear SVC and XGBoost (Extreme Gradient Boosting). Based on the evaluation methods outlined in the experiments section, we have selected Random Forests as our final model. Our final step was to use the Random Forest Regressor (since our target data is quantitative) to predict the most popular song for each available market. To do so, we fitted a RFR for each available market, and measured their performance individually using the Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2). A lower MSE and MAE and an R^2 closer to 1 indicate a better fit of the model.

IV. Visualization

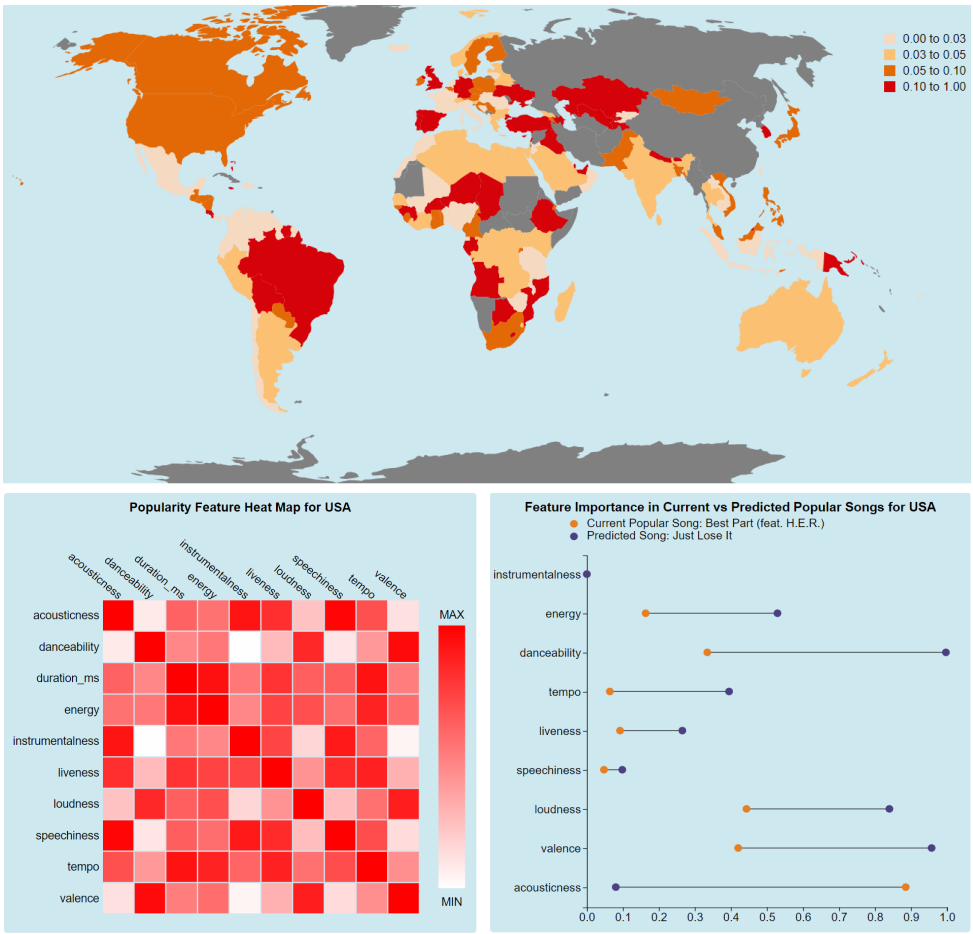


Figure 1. Choropleth, Heat Map, and Cleveland Dot Plot for Regional Music Popularity Prediction

The visualization for our project consists of a choropleth, heat map, and Cleveland dot plot of the most relevant features that influence music popularity as shown above. The user is able to select a feature from a selection tool and the choropleth updates based on the quartile of the normalized importance of that acoustic feature. When a user hovers over a country, a tooltip appears displaying the respective value for that country; if no data is available for that country, the country will be gray and the tooltip will indicate as such. When a user clicks a country with available data, the heat map and dot plot will update beneath

the choropleth to reflect values for that country. The heat map will show the correlation between normalized importance of acoustic features in song popularity. The dot plot shows an orange dot for the value of the normalized importance of each acoustic feature of the actual most popular song and a blue dot for the corresponding value of the predicted most popular song. The y-axis contains features sorted in decreasing order of normalized importance and varies with each region. The heat map allows the user to see which acoustic features correlate together to predict popular songs in a given country to provide guidance in maximizing song success/profit. The dot plot offers a visualization of the difference between predicted and actual acoustic feature values for popular songs in a country to provide the user more insight into which features are most accurate in predicting song popularity for the same purpose as before.

Experiments/Evaluation

I. Computation

List of Computation Questions Our Experiments Are Designed to Answer:

1. How well can machine learning models predict song popularity per region using acoustic features?
2. Which machine learning model performs best in terms of song popularity prediction?
3. Can the selected model perform a prediction for every different region?
4. Which acoustic features impact song popularity the most for different regions?

Detailed Description of Experiments and Observations:

Data Preprocessing: Our first step was to extract data directly from the Spotify API, getting information of around 220,000 songs and their associated features including track details, audio features, and popularity metrics, which are integral to our analysis. We then checked for null values and noticed that there were 58,004 tracks with no assigned markets and 184 with no names. Since we have over 25,161,429 rows and have no way of determining to which market a song belongs to or its name, we decided to drop all the NaN values.

Feature Engineering: After cleaning the data, we selected the most relevant features identified based on our literature survey. These were acousticness, danceability, duration, energy, instrumentalness, liveness, loudness, speechiness, tempo and valence. With this, we ensured avoiding any overfitting. There were 4 categorical variables: available_markets, genre, key, mode, and time signature. We encoded them into numerical format using one-hot encoding and created a binary column for each category.

Model Selection: We tested and evaluated six different machine learning models to determine their predictive performance on the whole cleaned dataset. These models included K-Neighbors, Decision Trees, Logistic Regression, Random Forests, Linear SVC, and XGBoost (Extreme Gradient Boosting). In terms of overall model performance, the Random Forest Classifier emerged as the top performer, achieving the highest accuracy with a score of 0.947904. This was closely followed by the Decision Tree Classifier, which registered an accuracy of 0.888340. When considering the Area Under the Curve (AUC) as a performance metric, the Random Forest Classifier again led the pack with an AUC of 0.811065, closely tailed by the Decision Tree Classifier with an AUC of 0.804608. This highlights that the Random Forest Classifier not only surpassed other models in terms of accuracy but also in AUC, making it the most suitable model for our task of analyzing the Spotify dataset.

	Model	Accuracy
1	RandomForestClassifier	0.947904
3	DecisionTreeClassifier	0.888340
5	XGBClassifier	0.864707
0	LogisticRegression	0.862961
2	KNeighborsClassifier	0.841966
4	LinearSVC	0.781000

	Model	AUC
1	RandomForestClassifier	0.811065
3	DecisionTreeClassifier	0.804608
2	KNeighborsClassifier	0.577932
4	LinearSVC	0.576775
5	XGBClassifier	0.506714
0	LogisticRegression	0.500000

Figure 2: Accuracy and AUC score for all six ML models

Our objective is to now use the selected best model to predict the most popular song per region and the associated feature importance. We will be using a Random Forest Regressor since our target data is quantitative. Every available marketplace has over 130,000 data points associated with it which allows us to split the dataset and fit a random forest model for each one without worrying about insufficient data. For each marketplace we have done the following:

- **Splitting the Dataset:** We split the data into features (X) and target (y) using an 80/20 standard split. Here, our target variable is the song's popularity.
- **Model Training/prediction:** We trained our random forest on the training dataset then evaluated the model's performance on the testing dataset.
- **Feature Importance:** After training, we checked which features were most important in predicting the song's popularity to provide the information in the visualization map. The importance is computed as the (normalized) total reduction of the criterion brought by that feature.

Some statistics for our Random Forest Regressor (these were computed by averaging the values over all different regions):

Evaluation Statistics	**Achieved Value**
Mean Squared Error	146.1895112309048
Mean Absolute Error	9.399691239661642
R-squared	0.6347769383527415

II. Visualization

List of Visualization Questions Our Experiments Are Designed to Answer:

1. Which UI elements are most effective to convey song popularity and potential song profit to a user?
2. Does a heat map and/or dot plot allow a user to easily detect trends in popularity features for a country?
3. How do users interact with the visualization? Which aspects are most engaging?
4. Does the choropleth allow a user to visualize trends in song popularity in different parts of the world?
5. Is the performance of the visualization optimal (i.e. small loading time) even with a large dataset?

A/B Testing: We used user testing to choose a Cleveland dot plot over a grouped bar chart, to display the differences in importance of features in actual vs. predicted popular songs per region. We implemented a Cleveland dot plot since it was less cluttered due to the limited number of data points. It was more readable and simple, thus minimizing chartjunk. We also performed tests to finalize the color palette.

Country-Level Analysis: When a user hovers over a country, the cursor changes to a pointer only if the country has available data for that feature. When the user clicks on the country, we drill down in the dataset and show visualizations of correlation heatmap and dot plot that visualize the relationship between different audio features within that specific country. Thus we show patterns in features and their impact on the actual popular song and the predicted popular song in the country.

User Interaction Analysis: We tracked user interactions with the dashboard. The users explored the heatmaps and the dot plot for multiple countries. The combination of the choropleth along with the heat map and dot plot proved to be engaging for the users.

Geospatial Analysis: We explored the choropleth map to identify countries where specific audio features are strongly correlated with song popularity. This can help discover interesting patterns and preferences in different parts of the world such as the USA and Mexico, and India and Australia, sharing preferences of features.

Performance Optimization: The input data to the visualization contains the final calculated values of normalized importance of acoustic features for predicted and actual popular songs by country. Since these are the final "static" calculated values, this is a relatively small dataset, so the visualization is able to update upon user interaction virtually instantaneously to result in a smooth user experience.

Conclusion and Discussion

Our project, centered on predicting regional music popularity, highlighted the efficacy of the Random Forest Regressor among various models, achieving an average R^2 value of 0.63478. Using data from Spotify's API, our model demonstrated its capability to predict music preferences across 180+ countries. We were able to effectively visualize the results using a choropleth map, heatmap and Cleveland dot plot. Since we were able to achieve satisfactory results, it indicates that acoustic features do genuinely have a correlation with popularity of a song, and that it can vary by region based on cultural and social perceptions of the area.

Potential Use Cases:

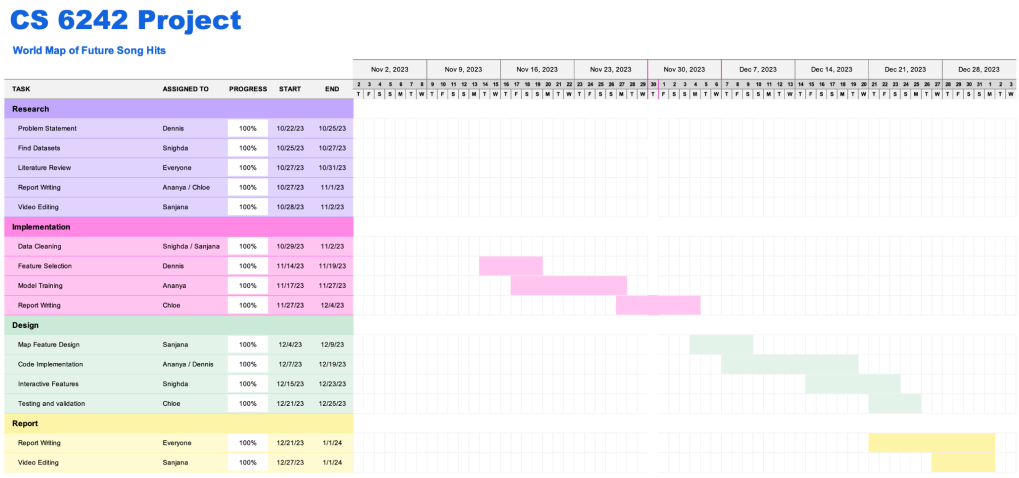
This analysis benefits: Music companies (invest in hit songs), artists (refine songs based on popular metrics), music platforms (curate playlists), and content creators (select popular songs for wider reach). All stakeholders will be able to select better and more profitable songs based on desired product region.

Possible Future Improvements:

1. Leverage Genius Lyrics API for Deeper Insights: Integrate the Genius lyrics API to explore the correlation between song lyrics and popularity. This addition can provide valuable textual context, enriching the predictive power of the model.
2. Incorporate Social Network Data for Enhanced Sentiment Analysis: Explore the integration of social network data to improve popularity prediction. Recognizing the strong connection between social sentiment and song popularity, this approach can offer a more comprehensive understanding of user preferences.
3. Dynamic Playlist Generation from Regional Popularity Metrics: Implement a feature for playlist generation based on current song popularity features by country. By dynamically adapting to musical preferences, the system can provide users with curated playlists that provide insight and enjoyment based on current regional trends.

Contributions

The work split of all team members is displayed in the following Gantt Chart:



TASK	ASSIGNED TO	PROGRESS	START	END
Research				
Problem Statement	Dennis	100%	10/22/23	10/25/23
Find Datasets	Snighda	100%	10/25/23	10/27/23
Literature Review	Everyone	100%	10/27/23	10/31/23
Report Writing	Ananya / Chloe	100%	10/27/23	11/1/23
Video Editing	Sanjana	100%	10/28/23	11/2/23
Implementation				
Data Cleaning	Snighda / Sanjana	100%	10/29/23	11/2/23
Feature Selection	Dennis	100%	11/14/23	11/19/23
Model Training	Ananya	100%	11/17/23	11/27/23
Report Writing	Chloe	100%	11/27/23	12/4/23
Design				
Map Feature Design	Sanjana	100%	12/4/23	12/9/23
Code Implementation	Ananya / Dennis	100%	12/7/23	12/19/23
Interactive Features	Snighda	100%	12/15/23	12/23/23
Testing and validation	Chloe	100%	12/21/23	12/25/23
Report				
Report Writing	Everyone	100%	12/21/23	1/1/24
Video Editing	Sanjana	100%	12/27/23	1/1/24

All team members have contributed a similar amount of effort

References

[1] Khan, F., Tarimer, I., Alwageed, H. S., Karadağ, B. C., Fayaz, M., Abdusalomov, A. B., & Cho, Y.-I. (2022). Effect of Feature Selection on the Accuracy of Music Popularity Classification Using Machine Learning Algorithms. *Electronics*, *11*(21), 3518. http://dx.doi.org/10.3390/electronics11213518

[2] Pareek, P., Shankar, P., Pathak, P., & Sakariya, N. (2022). Predicting Music Popularity Using Machine Learning Algorithm and Music Metrics Available in Spotify. *Journal of Development Economics and Management Research Studies, 9*(11), 10-19. https://www.cdes.org.in/wp-content/uploads/2022/01/Predicting-Music-Popularity.pdf

[3] Rajyashree, R., Anand, A.., Soni, Y., & Mahajan, H. (2018). Predicting Hit Music using MIDI features and Machine Learning. *2018 3rd International Conference on Communication and Electronics Systems*. 94-98. https://doi.org/10.1109/CESYS.2018.8724001

[4] Shulman, B., Sharma, A., & Cosley, D. (2021). Predictability of Popularity: Gaps between Prediction and Understanding. *Proceedings of the International AAAI Conference on Web and Social Media*, *10*(1), 348-357. https://doi.org/10.1609/icwsm.v10i1.14748

[5] Khulusi, R., Kusnick, J., Meinecke, C., Gillmann, C., Focht, J., & Jänicke, S. (2020). A survey on visualizations for musical data. *Computer Graphics Forum, 39*(6), 82-110 https://doi.org/10.1111/cgf.13905

[6] Pham, J., Kyauk, E., & Park, E. (2016). Predicting song popularity. *Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep*, *26*. https://cs229.stanford.edu/proj2015/140_report.pdf

[7] Wada, K., Wallner, G., & Vos, S. (2022). Studying the Utilization of a Map-Based Visualization with Vitality Datasets by Domain Experts. *Geographies*, *2*(3), 379-396. https://doi.org/10.3390/geographies2030024

[8] Gulmatico, J. S., Susa, J. A. B., Malbog, M. A. F., Acoba, A., Nipas, M. D., & Mindoro, J. N. (2022). SpotiPred: A machine learning approach prediction of Spotify music popularity by audio features. *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)*, 1-5. https://doi.org/10.1109/ICPC2T53885.2022.9776765

[9] Grace, A. S. (2022). Song and Artist Attributes Analysis For Spotify. *2022 International Conference on Engineering and Emerging Technologies (ICEET)*, 1-6. https://doi.org/10.1109/ICEET56468.2022.10007360

[10] Araujo, C. V. S., De Cristo, M. A. P., & Giusti, R. (2019). Predicting music popularity using music charts. *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 859-864. https://doi.org/10.1109/ICMLA.2019.00149

[11] Lee, J., & Lee, J. S. (2018). Music popularity: Metrics, characteristics, and audio-based prediction. *IEEE Transactions on Multimedia*, *20*(11), 3173-3182. https://doi.org/10.1109/TMM.2018.2820903

[12] Raza, A. H., & Nanath, K. (2020). Predicting a Hit Song with Machine Learning: Is there an apriori secret formula? *2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*, 111-116. https://doi.org/10.1109/DATABIA50434.2020.9190613

[13] Vötter, M., Mayerl, M., Specht, G., & Zangerle, E. (2021). Novel datasets for evaluating song popularity prediction tasks. *2021 IEEE International Symposium on Multimedia (ISM)*, 166-173. https://doi.org/10.1109/ISM52913.2021.00034

[14] Lončar, P. (2022). Internet of Musical Things and Music Data Visualization. *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, 1501-1506. https://doi.org/10.23919/mipro55190.2022.9803404

[15] Zeng, Y. R., Chang, Y. S., & Fang, Y. H. (2019). Data visualization for air quality analysis on bigdata platform. *2019 International Conference on System Science and Engineering (ICSSE)*, 313-317. https://doi.org/10.1109/ICSSE.2019.8823437

[16] Dalida, M. R., Aquino, L. B., Hod, W. C., Agapor, R. A., Huyo-a, S. L., & Sampedro, G. A. (2022). Music Mood Prediction Based on Spotify's Audio Features Using Logistic Regression. *2022 IEEE 14th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, 1-5. https://doi.org/10.1109/HNICEM57413.2022.10109396

[17] Nikas, D., & Sotiropoulos, D. N. (2022). A Machine Learning Approach for Modeling Time-Varying Hit Song Preferences. *2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 1-8. https://doi.org/10.1109/iisa56318.2022.9904376