TEAM RATTATA FINAL TECHNICAL REPORT

Necdet Kaan Özdoğan 191180757

Berat Berkay Erken 191180758

Mehmet Ümit Açık 191180001

Zehra Yıldız Açıkgül 191180002

ABSTRACT

Blackpink is a South Korean girl group formed by YG Entertainment in 2016. The group consists of four members: Jisoo, Jennie, Lisa and Rosé. Blackpink music group has proven itself to the whole world with their songs since 2016 and has become the most popular k-pop group of recent times. For this reason, we wanted to examine Blackpink in our project, what is the difference of Blackpink from other groups, we reviewed the features of their songs such as danceability, energy, key, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration etc. We used python libraries, methods and datasets while doing this project.

Keywords

K-Pop, Blackpink, Regression, Classification, Data Science, Machine Learning

1. INTRODUCTION

Recently, the increasing interest and fanbase for Kpop group Blackpink has been the subject of our project. It has been tried to answer questions such as why Blackpink has become so popular, what is the difference of this group from other groups, and why Blackpink's songs are listened to so much. Python libraries, Spotify API, and Kaggle K-pop datasets were helpful in answering these questions. The latest updated versions of machine learning and artificial intelligence were used and questions were answered in the project.

2. METHODOLOGY

We are going to use classification method, which is a supervised learning to distinguish BlackPink from other kpop groups. We are going to use datasets from the web site which name is Kaggle for data collection tools and procedures. We are going to choose Python, which is an object-oriented, interpretative, modular, and interactive high-level programming language to code and Jupyter Notebook, which is an online compiler, IDE, and collaborative coding environment. One of the most frequently utilized forms of unsupervised learning is Clustering. Clustering is the process of separating different parts of data based on common characteristics. Specifically, the average distance of each observation from the cluster center, called the centroid, is used to measure the compactness of a cluster and this makes sense because a good Python clustering algorithm should generate groups of data that are tightly packed together. We are going to be performing to segmentation analysis on the different type of the music using our datasets, which gets from Kaggle.

3. DATASET

Kaggle dataset of K-pop

Context: K-pop (abbreviation of Korean pop) is a genre of popular music originating in South Korea.

Content:

• K-pop Idols

• K-pop Boy Groups

• K-pop Girl Groups

• K-pop Music Videos

API from Spotify data of K-pop and Blackpink.
Relevant data:

• All artists' information under 'k-pop' genre;

• All albums' information of every k-pop artists

• All tracks' information of all albums of every k-pop artists

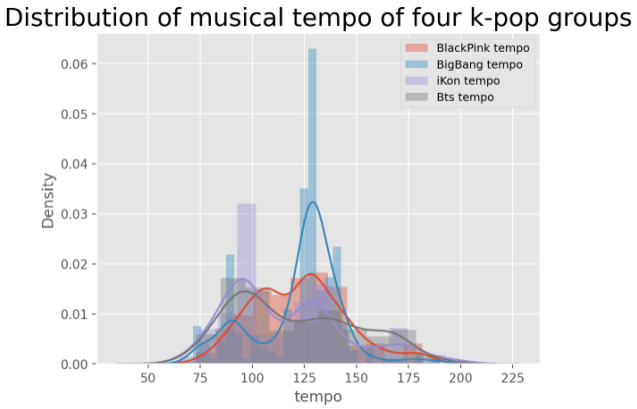
• Audio features provided by Spotify of all tracks

API from Youtube data of K-pop and Blackpink.
Relevant data:

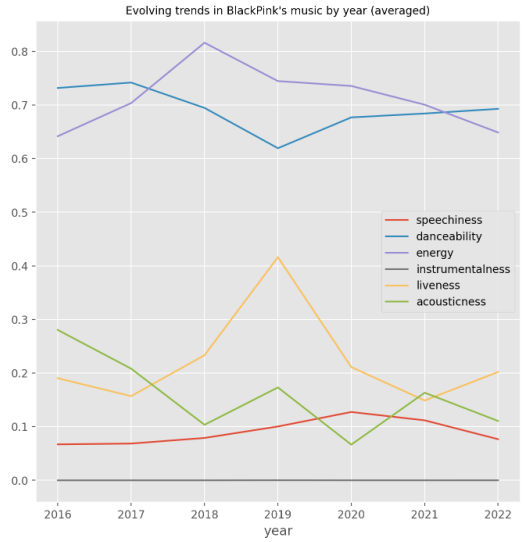
• Details of Youtube channels and videos of k-pop groups and Blackpink

4. EXPERİMENTS

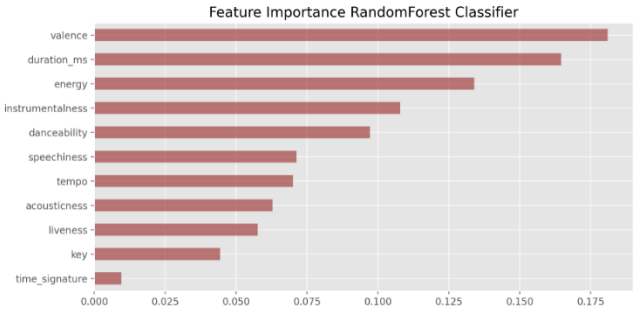
First of all, the data of other groups, including Blackpink, was obtained thanks to the Spotify API and converted to csv file. It was then merged with the existing csv dataset, including BTS. There are 13 groups in total in this merged dataset. To distinguish these groups from each other, we divided them into Blackpink and other.



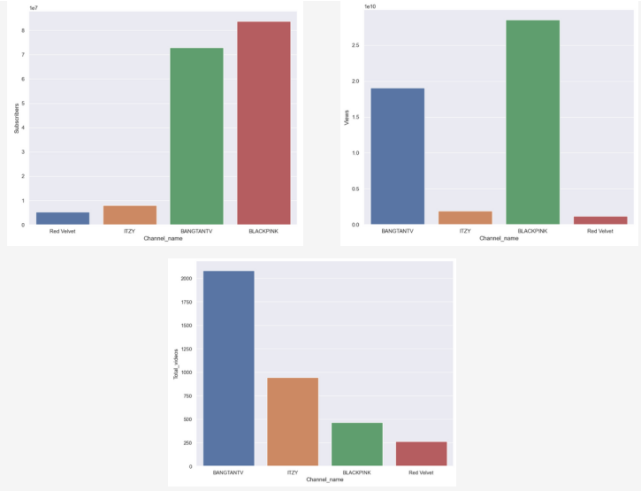
Danceability, energy, key, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_ms, time_signature, year properties of Blackpink group in the combined dataset were examined.



We chose RandomForest Classifier, which is tree-based boosted algorithm. We will also use a oversampling technique (SMOTE) to address the class imbalance issue.



Using the Youtube api, graphs were obtained comparing the number of current views, the number of subscribers and the total number of videos of the blackpink group.



5. RESULTS

We used Youtube Data API to compare Blackpink and the other popular Kpop groups. We compared the subscriber counts of K-pop groups and we see that the group with the highest number of subscribers is Blackpink and also Blackpink is the most viewed group. We used the spotipy library, we got the data of kpop and american pop groups from spotify and we converted it to a csv file. Then we combined the other dataset which is we downloaded from kaggle into a separate python file and we used this combined the data. We got the audio characteristics of the music groups and we see the audio characteristics of the music groups compared to each other for using Spotify. We did a one-v-all approach. As such, our target label will be 1 if BlackPink and 0 if any other group. We used corr() function to find the correlation among the columns in the Dataframe using the 'Pearson' method. We imported machine-learning algorithms libraries and we split 80% of the dataset for training and 20% for testing. We used Smote to oversample the data and e calculated the ROC AUC score with three different machine-learning algorithms. We calculated the mean of the audio features of the groups and showed them as two-dimensional tabular data. Feature Importance refers to techniques that calculate a score for all the input features for a given model — the scores simply represent the "importance" of each feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable. The Feature Impact of LightGBM Classifier on model output is partly on the SHAP values. SHAP value shows us how much each feature individually contributes to the final prediction, negatively or positively.

We compared Korean musical artists with American artists and we did a one-v-all approach. As such, our target label will be 1 if BlackPink and 0 if any other group. We calculated of the mean of the audio features of the groups and we split 80% of the dataset for training and 20% for testing. After that we used Smote to oversample the data we calculated the ROC AUC score with gradient boosting machine-learning algorithm. We added blackpink to american pop dataframe.

As a result of all this data, we understood why Blackpink group is more popular than other popular groups.

6. CONCLUSION

The difference of the recently popular Kpop group Blackpink from other groups has been examined. In the project, which was carried out using machine learning and artificial intelligence, the song features of Blackpink were compared with other Kpop groups, and the results presented us with the features that distinguish Blackpink.

7. REFERENCES

[1] Who is BLACKPINK? https://rachelombok.com/blackpink

[2] Data Analysis of K-pop: Playing with Spotify API https://nancyyanyu.github.io/posts/63adf3bb/

[3] What Is Clustering? https://builtin.com/data-science/data-clustering-python

[4] Extracting Song Data From the Spotify API Using Python https://towardsdatascience.com/extracting-song-data-from-the-spotify-api-using-python-b1e79388d50

[5] How to combine multiple CSV files using Python for your analysis https://medium.com/@stella96joshua/how-to-combine-multiple-csv-files-using-python-for-your-analysis-a88017c6ff9e