GAZI UNIVERSITY ENGINEERING FACULTY
COMPUTER ENGINEERING CENG 313 DATA SCIENCE
FINAL PROJECT REPORT

TEAM NİDOQUEEN

1st Author
Serdem ERGÜN
181180028

2nd Author
Feyyaz KAVUN
171180039

3rd Author
Şeyma SERTER
171180060

ABSTRACT

Here, by using Spotify Data, it is aimed to make detailed analysis of the unique characteristics of the music listened to according to the relationships between the songs. In this data analysis study of Spotify listeners, the data of the music they listen to are recorded instantly. As a team, we evaluated them using some algorithms and compared them among themselves. We also deepened our analysis by using visualization.

1. INTRODUCTION

Services like Spotify haven't just changed the way they consume music. Beyond that, with the help of technology, it also has the power to guide the user's music listening habits and music preferences. Spotify, a music application founded in 2006 and nowadays has 286 million users [1]. Each activity of these users within the application is recorded instantly, this data is analyzed with the help of artificial intelligence, the data is collected and then the user can offer accurate music suggestions, used for.

As a team, what are the relationships between them? What relationships are linked between each other? If the danceability and instrumentality of the music is higher, is it energetic? Or if their duration is longer, is it possible to talk about a high pace? We have answered these questions.

In this research, various analyzes were attempted with the information obtained from the Spotify data set. Questions were asked about the relationship between features in the data set, and these questions were answered using methods such as visualization, classification and regression.

2. THE APPROACH

2.1. Data Visualization

In this part shows us effects of all factors on this data set.

2.1.1. Feature analysis of liked and disliked songs

The "Target" feature is a feature that indicates whether the person preparing the data set likes that song or not. The "target = 1" status means that the song is liked, and the "target = 0" status means that the song is not liked.

Thanks to the given data, information can be obtained about the criteria of the songs in the data set affecting whether they are liked or not, by using histogram.

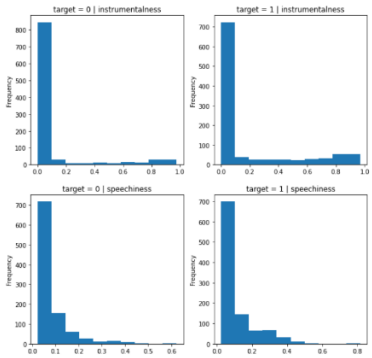

Figure 1: "Speechness" and "instrumentalness" histograms grouped according to the "target" feature

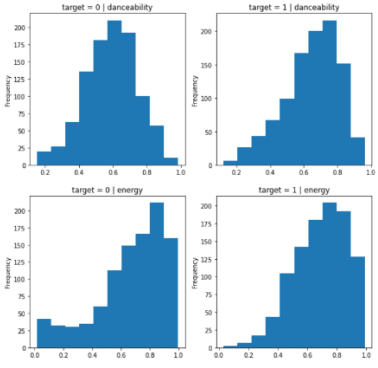

Figure 2: "Danceability" and "energy" histograms grouped according to the "target" feature

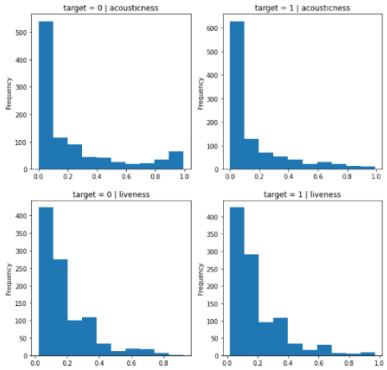

Figure 3: "Acousticness" and "liveness" histograms grouped according to the "target" feature

A lot of information can be obtained from the histograms given above. For example, although there is not much difference between the dislikes and likes of other intervals for the "acousticness" feature, the dislikes rate of those with a value between 0.9 and 1.0 is higher than the rate of like. Similarly, for the "energy" feature, although there is not much difference between the dislikes and likes of other ranges, the dislikes rate of those with a value between 0.0 and 0.2 is higher than the rate of like.

2.1.2 Like/Dislike analysis of songs according to their "tempo" feature

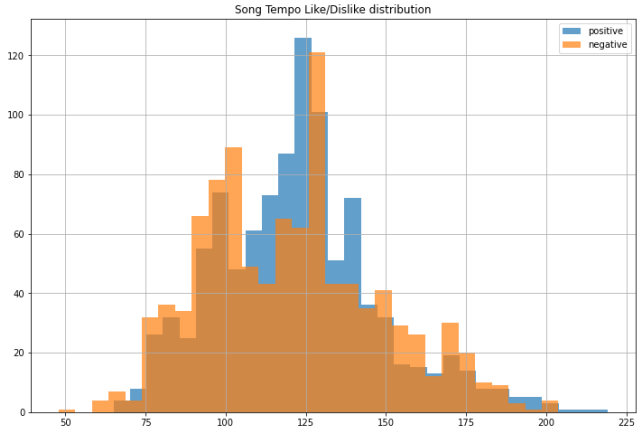

Figure 4: Histogram of liked or disliked songs according to the "tempo" feature

As it is clearly seen in the figure, the rate of liking for the songs with "tempo" between 100-150 is higher than the rate of disliking. On the contrary, for other ranges, dislikes are more than likes.

2.1.3. Heat map that includes every feature

Heat maps are used to show whether any variables are similar to each other and to determine whether there is any correlation between them.

The use of heat maps makes perfect sense for this data set. Perhaps the most basic question to be asked is, does the change of one feature affect the other feature and how much does it affect it.

By looking at the heat map, the interactions of the features with each other can be seen very clearly.

As the linear relationship between the features in the heat map, shown in the next figure, decreases; the color changes from green to red. Many results can be drawn from the heat map prepared with all the data of our data set. For example, there is an opposite relationship between "acousticness" feature and "loudness" and "energy" features, while "energy" feature and "loudness" feature are in a linear connection with each other.

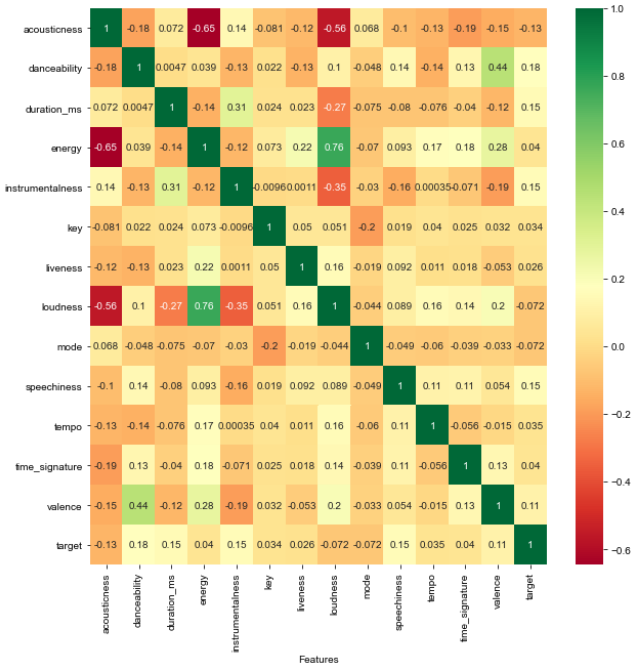

Figure 5: Heat map for all the features

2.1.4. Analysis for "tempo" and "mode" features

The "mode" property represents modality. If "mode = 1" the track is major, if "mode = 0" the track is minor.

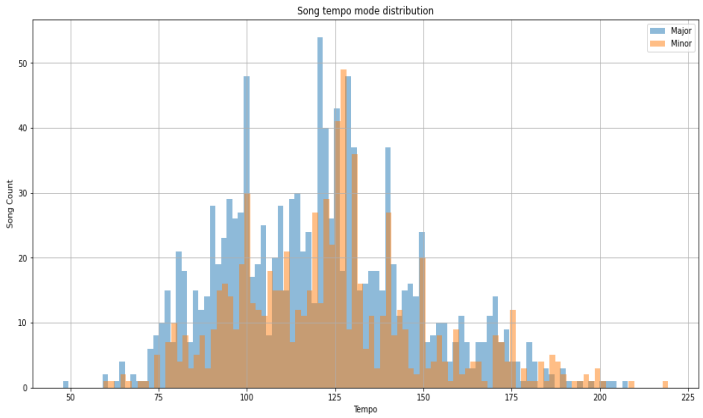

Figure 6: Histogram prepared according to "mode" and "tempo" features

This histogram shows how many songs are in certain "tempo" intervals, whether the number of minor tracks or major tracks is dominant. However, the point that should be considered while looking at this figure is that it should not be decided for "mode" by looking at "tempo" intervals. Although there is a dominant "mode" value for each range, there are 782 minor tracks and 1235 major tracks in the data set. In other words, if the research is about the effect of the "tempo" feature on the "mode" feature, the number of minor and major songs should be kept the same so that a healthy result is obtained. Therefore, when looking at this figure, such a conclusion cannot be drawn, only the number of pieces distributed according to the "tempo" and "mode" features can be seen.

2.1.5 3D Modelling

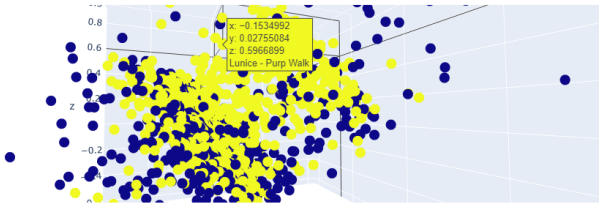

Figure 7: 3D model by “target”

In this model, prepared with "energy", "liveness", "tempo", "valence", "loudness", "speechiness", "acousticness", "danceability", "instrumentalness"columns with “target” column. We can tell which is liked song or not easily. Yellows are liked songs blues are not liked. We can understand that user likes musics that generally has higher x value. Also he listens every music. Because marks are interbedded.

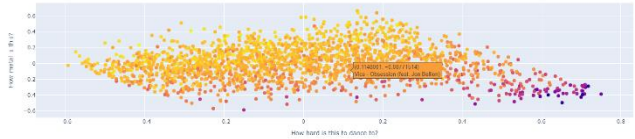

Figure 8: 2D model “how the metal is or how hard is this to dance to”

This model shows us which musics are in the metal class and how hard is this to dance. This model prepared with "energy", "liveness", "tempo", "valence" columns with “loudness”. This columns selected cause of share commons features of metal music and music that will make you dance.

2.1.6 Classification by Target Column

We told before “target” column is liked or not liked songs by user. We classified songs by “target” column with 2 algorithm.

Firstly we start with Decision Tree algorithm. Minimum samples of leaf is 50 with 10 random state. We got %69.8 accuracy for test data and we got %76.1 accuracy for train data.

Secondly we continued with Random Forest algorithm. After several try we decided number estimator is 120 ideal. With this model we got %77.6 accuracy.

2.1.7 Classification by Artist

In this section we classified songs features by artist. Firstly we found 10 most frequent artist in dataset. Then we selected 4 of them. These are 'Kanye West','Drake','Rick Ross','Disclosure' . We have 49 data about them. Then we deleted some columns that make trouble us These are “target”, “song_title”, “key”, “mode”, “time_signature”. “target is not necessary, “song_title” is specific and “key”, “mode”, “time_signature” columns do not have distinctive features.

Secondly we classified them firstly with Naïve-Bayes. We used Gaussian Naïve Bayes model. After classification we put results in confusion matrix.

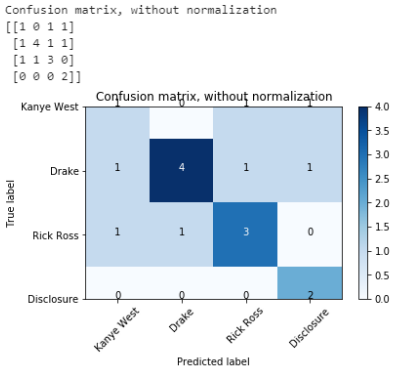

Figure 9: Gaussian Naïve Bayes Confusion Matrix

With this model our learning rate is 10/17 of train data.

Our second model was Random Forest Classifier. We used 19 estimator with random state = 0.

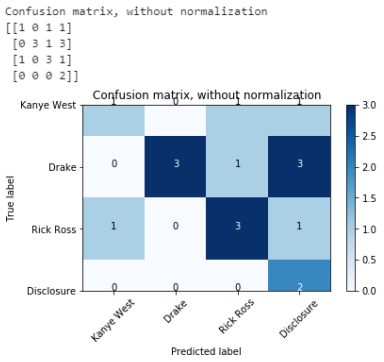

Figure 10: Random Forest Confusion Matrix

With this model our learning rate is 9/17 of train data

Our third model was k-nn model. We used k as 5 because of our train datas’ number is 32.

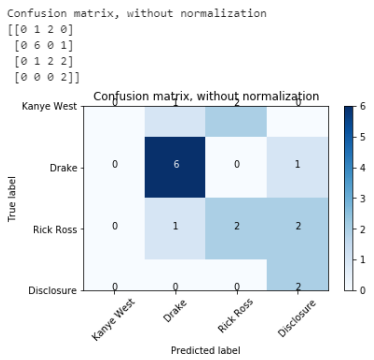

Figure 11: K-NN Confusion Matrix

With this model our learning rate is 10/17 of train data

As a result our models are successful for classify Drake's songs. Otherwise it not more successful.

2.1.8 Regressions

We tried regression between two columns by correalation maps. We was searching for if correalation rate is high between two features then linear regression is good for them.

First we selected loudness and energy. Their rate was high. Then we got results shown by figure 12.

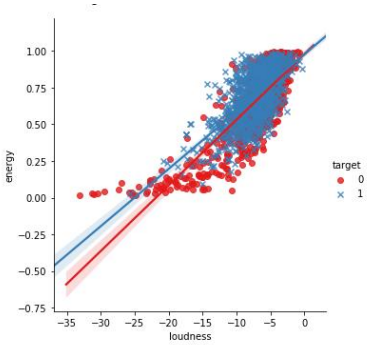

Figure 12: Linear Regression

As seen, it is successful. Then we tried lower correalation rate. These was loudness and acousticness. Then we got result shown by figure 13.

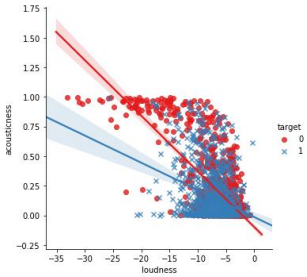

Figure 13: Linear Regression-2

As seen we got worse result then other. So our hypothesis can be true.

After linear regression we tried apply Multiple Linear Regression on this dataset. In first try we got very bad R-Squared result. Then we apply backward elimination. We dropped 3 columns that has high P>|t| value. We calculated again R-Squared. Then we got worse value than first. So We decided Multiple Linear Regression is not successful for this dataset.

3. RESULTS

Some steps were followed within the scope of the project. It is possible to list these stages as follows. First the problem and purpose are determined. The aim of our project was Detailed analysis of Spotify Song Attributes. Then relevant data werecollected .Later, some work has been done for data analysis. In this step, seaborn, matplotlib libraries are used for data visualization. Graphs such as histograms and bars were preferred to see the relationships between the data of: acousticness, danceability, duration, energy, instrumentalness, liveness, loudness, speechiness, tempo and valance, which are in our dataset. During this preference, graphics where the effects of the factors can be seen in the best way were preferred. After we realized the recognition of our data, we went to the last step. In this step, we wrote down the algorithms that model our data. We examined the algorithms that we think can be successful and suitable for our data by regression and classification. We have applied algorithms such as Random Forest, Decision Tree, Gaussian Naive Bayes, K-NN, Linear Regression, Multiple Linear Regression for our data. We defended hypothesis of if there is a positive correalation between two features, its linear regression's accuracy will be high.

Finally, we saw in histograms by target which features is effective on users music choice, how is it easy to dance, how metal is it, which relationships have they have. Then we decided for classification by target best is Random Forest. Also we decided for classification by artist bests are K-NN and Random Forest. We saw that our hypothesis about linear regression is can be true. Then we noticed about multiple linear regression after 3 backward elimination is not successful for this dataset.

4. CONCLUSIONS

In this study, we followed steps such as data analysis, data visualization and success test of algorithms until reaching a conclusion. The steps we followed at the end of the whole study showed how we can make a detailed analysis of the data we have. Throughout the study, we experienced how we should approach and interpret data.

5. REFERENCES

[1] https://www.businessofapps.com/data/spotify-statistics/#:~:text=during%20Q3%202018-,Spotify%20User%20Statistics,Premium%20subscribers%20in%20Q4%202019..