**International Journal of Science, Technology and Society**
2021; 9(5): 239-244
http://www.sciencepublishinggroup.com/j/ijsts
doi: 10.11648/j.ijsts.20210905.16
ISSN: 2330-7412 (Print); ISSN: 2330-7420 (Online)

# Music Popularity Prediction Through Data Analysis of Music's Characteristics

**Jaehyun Kim**

Cheongshim International Academy, Gyeonggi-do, Korea

**Email address:**
jkhappyvirus@gmail.com

**To cite this article:**
Jaehyun Kim. Music Popularity Prediction Through Data Analysis of Music's Characteristics. *International Journal of Science, Technology and Society*. Vol. 9, No. 5, 2021, pp. 239-244. doi: 10.11648/j.ijsts.20210905.16

**Received**: October 7, 2021; **Accepted**: October 20, 2021; **Published**: October 29, 2021

**Abstract:** Today, the music industry has grown tremendously with the emergence of smartphones and streaming services. In the past, the most of revenue was from the album's sales and concerts. However, these days, streaming services on the web or smartphones have become a huge part of the music industry. Therefore, from an artist's perspective, it is important to rank their music high on streaming services to earn money. While the music industry is growing, the top 1% of artists have gone from earning 26 percent of revenue to between 56% and 77%. This shows the huge income gap among the artists. Large profit on various artist can help to make a better music business. This paper is written in order to analyze the popular music in Spotify, which is one of the most popular music streaming services in the world. To find the factors that popular music has, this paper analyzes data of 2010~2019 top 50 music on Spotify. The paper also presents the table and graph that clearly illustrate the average of many music factors such as beat per minute and duration to investigate how music should be made to rank high on the Spotify. Moreover, the paper utilizes a machine learning model to predict the popularity of music by analyzing the beat per minute, speecheness, loudness, and duration, etc. The prediction model is expected to be used by many artists or music companies before they release their music.

**Keywords:** Data Science, Machine Learning, EDA, Music, Business

## 1. Introduction

Today, the music industry has continued to grow with the development of electronic devices and technology. The emergence of smartphones, smartphone applications, and Bluetooth earphones allow people to listen to music at any time [1]. The global recorded music market grew by 7.4% with total revenue of US$ 21.6 billion. Due to the increase of revenue in the music industry, the number of artists and music companies has been increasing [2, 3].

However, income disparity is a serious problem in the music industry [4, 5]. While famous and popular artists and music companies earn enormous amounts of money, most artists and companies earn less than the minimum cost of living. Since 1982, the top 1% of artists have gone from earning 26 percent of revenue to between 56% and 77% [6]. In the past, music concerts and sale rates of music CDs are the most revenue. However, these days with the emergence of smartphone and music applications, music streaming, which is a way of delivering music without requiring users to download files from the internet. Music streaming revenue has increased dramatically in the last five years growing from $1.9 billion in 2014 to $ 10.1 billion in 2020 [7]. The income of an artist may not be an important issue for some artists, but more income from more artists will help the music industry with the high quality of music.

Spotify is one of the most famous music streaming services which supply numerous kinds of music in the world [8, 9]. Therefore, the artist can earn tons of money if the music ranks high on Spotify. To rank high on Spotify, it is important to understand listeners' tastes and interests [10].

This paper is going to analyze the music of the top 50 in Spotify to investigate which factors of music influence the rank. The paper is going to use not only the genre but the tempo and other variables about the song. Then the paper is going to implement the machine learning model that can predict the popularity or rank of music in Spotify. The model does not consider the name of the artist when it predicts the popularity of an unpopular artist can decide whether their music has the potential to be ranked in Spotify. The research expects that this model can be used by many artists or music companies to predict their rank before they release the music.

The paper contains 5 sections: Introduction, data Exploration, popularity prediction model, and conclusion.

# 2. Data Exploration

Before implementing the model, the paper examine the data that collected from Spotify. This section analyzes the dataset to summarize its characteristics with the visualization.

### *2.1. About Data*

The section analyzes the two different sets of data: Top 10 songs in 2010-2019 and Top 50 songs in 2019. The data set contains 13 columns: name of song, artist's name, genre, BPM (bit per minute), energy, danceability, loudness, liveness, length, acousticness, speechiness, and popularity.

### *2.2. Overview*

In this section, the paper first summarizes the top 10 songs between 2010 - 2019. Table 1 shows the average of BPM, evergy, and danceability. Table 2 is average of Loudness (dbm), liveness, and valence. Table 3 shows the average of length, acoustintness, and speechness.

***Table 1.*** *Average BPM, Energy and Danceability in each year.*

| Year | Avg. BPM | Avg. Energy | Avg. Danceability |
|---|---|---|---|
| 2010 | 122 | 78 | 65 |
| 2011 | 119 | 75 | 64 |
| 2012 | 121 | 75 | 66 |
| 2013 | 122 | 74 | 62 |
| 2014 | 123 | 68 | 63 |
| 2015 | 120 | 70 | 64 |
| 2016 | 114 | 67 | 63 |
| 2017 | 117 | 69 | 65 |
| 2018 | 115 | 65 | 67 |
| 2019 | 112 | 65 | 70 |

Beats per minute (MBP) indicate the tempo in the music. In other words, music with higher bpm is faster than the lower one. In table 1, there is a trend of BPM in each year. From 2011 to 2015, the average BPM is higher than 120. However this has decreased to 115 after 2015. There was a similar trend in Energy. Energy of popular songs was high in 2011 to 2015, but it decreased after 2015. Since the BPM is highly related to the energy of songs, they have similar patterns. Unlike BPM and energy there is no huge difference on danceability. However, the popular songs in recent years have more dancability than the past.

***Table 2.*** *Average Loudness, Liveness and Valence in each year.*

| Year | Avg. Loudness | Avg. Liveness | Avg. Valence |
|---|---|---|---|
| 2010 | -5 | 21 | 57 |
| 2011 | -5 | 21 | 54 |
| 2012 | -4.9 | 16 | 64 |
| 2013 | -5.1 | 20 | 53 |
| 2014 | -5.8 | 17 | 52 |
| 2015 | -5.6 | 18 | 53 |
| 2016 | -6.7 | 18 | 45 |
| 2017 | -5.6 | 15 | 52 |
| 2018 | -5.6 | 15 | 49 |
| 2019 | -5.8 | 15 | 51 |

Loudness also has a similar pattern with the BPM and energy because people usually feel energy with the higher bpm and louder beat and sound. Valence describes the musical positiveness in the music. Music with high valence sounds more positive (e.g. happy and cheerful). In contrast, music with low valence sounds more negative (e.g sad and depressed). Usually a positive song has higher BPM with energy while a negative song has lower BPM with lower energy. Therefore, the average of valence is higher in 2010-2015 than the 2016-2019.

***Table 3.*** *Average Length, Acoustitiness and Speecheness in each year.*

| Year | Avg. Length | Avg. Acoustitiness | Avg. Speecheness |
|---|---|---|---|
| 2010 | 230 | 11.6 | 8.9 |
| 2011 | 243 | 13.3 | 9.7 |
| 2012 | 224 | 4.9 | 5.8 |
| 2013 | 234 | 10.3 | 8.3 |
| 2014 | 224 | 17.6 | 8.7 |
| 2015 | 223 | 16.6 | 7.1 |
| 2016 | 220 | 15.9 | 8.4 |
| 2017 | 222 | 16.6 | 9.8 |
| 2018 | 217 | 12.8 | 8.6 |
| 2019 | 200 | 21.7 | 8.1 |

As shown in Table 3, the length of music is getting shorter. The reason that the song is getting shorter is streaming. As explained in the previous section, the most of revenue is from the streaming. In most streaming services, artists can get paid if someone listens to at least 30 seconds of a song. Therefore, a song doesn't have to be long. Also, many artists tend to make 12 short songs in one album than 10 long songs because there is more chance to get streamed if there are more songs in the streaming service. There were no special patterns found in aoucstitiness and speechenss.

### *2.3. Genre*

This section investigates which genres are popular on Spotify. Since there were 50 different genres, they are divided the genre vs. popularity into two different graphs: Figure 1 and Figure 2. The higher popularity means the higher rank in this graph. One of the most popular genres were R&B and escape room. Escape room is the genre that is related to R&B. So from this section can conclude that R&B is the most popular genre in Spotify. Other than R&B, electronic music such as electronic pop are also popular.

### *2.4. Music Trend in 2021*

As explained in the previous section, there is a trend and popular style of music in each year. From the EDA process with the Top 10 data in 2010~2019, the R&B style with low BPM is the most popular style. This section also analyzes the music trend in 2019 by analyzing the Top 50 music on Spotify.

As shown in Figure 1, the most popular music was R&B music in 2010-2019. However, as shown in Figure 3, pop music was the most popular genre in 2019. This indicates that the trend of music in recent years is pop music.

The graph also indicates that genre can affect the popularity of music. So, in the later section, the prediction model also uses genre with the other 9 vectors to predict the popularity of music.

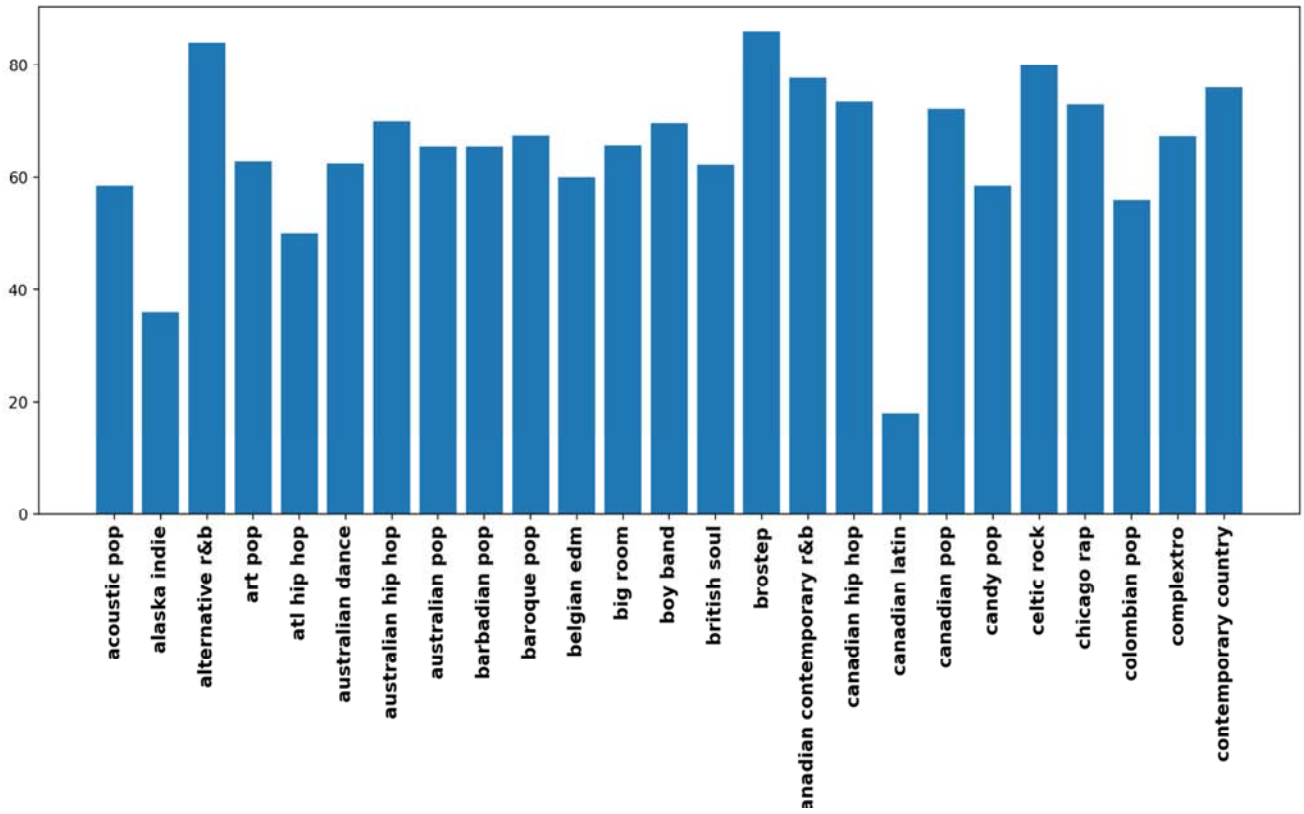


***Figure 1.*** *The genre vs. popularity (1).*

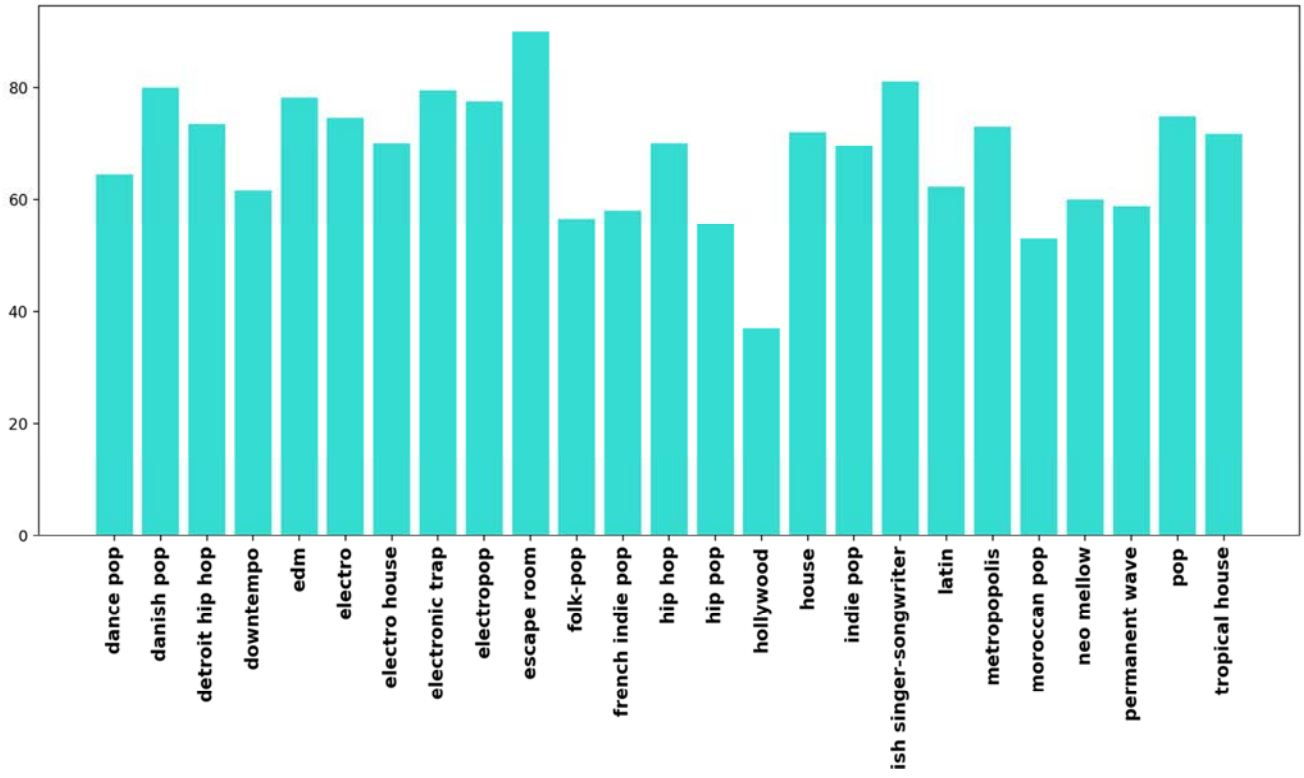


***Figure 2.*** *The genre vs. popularity (2).*

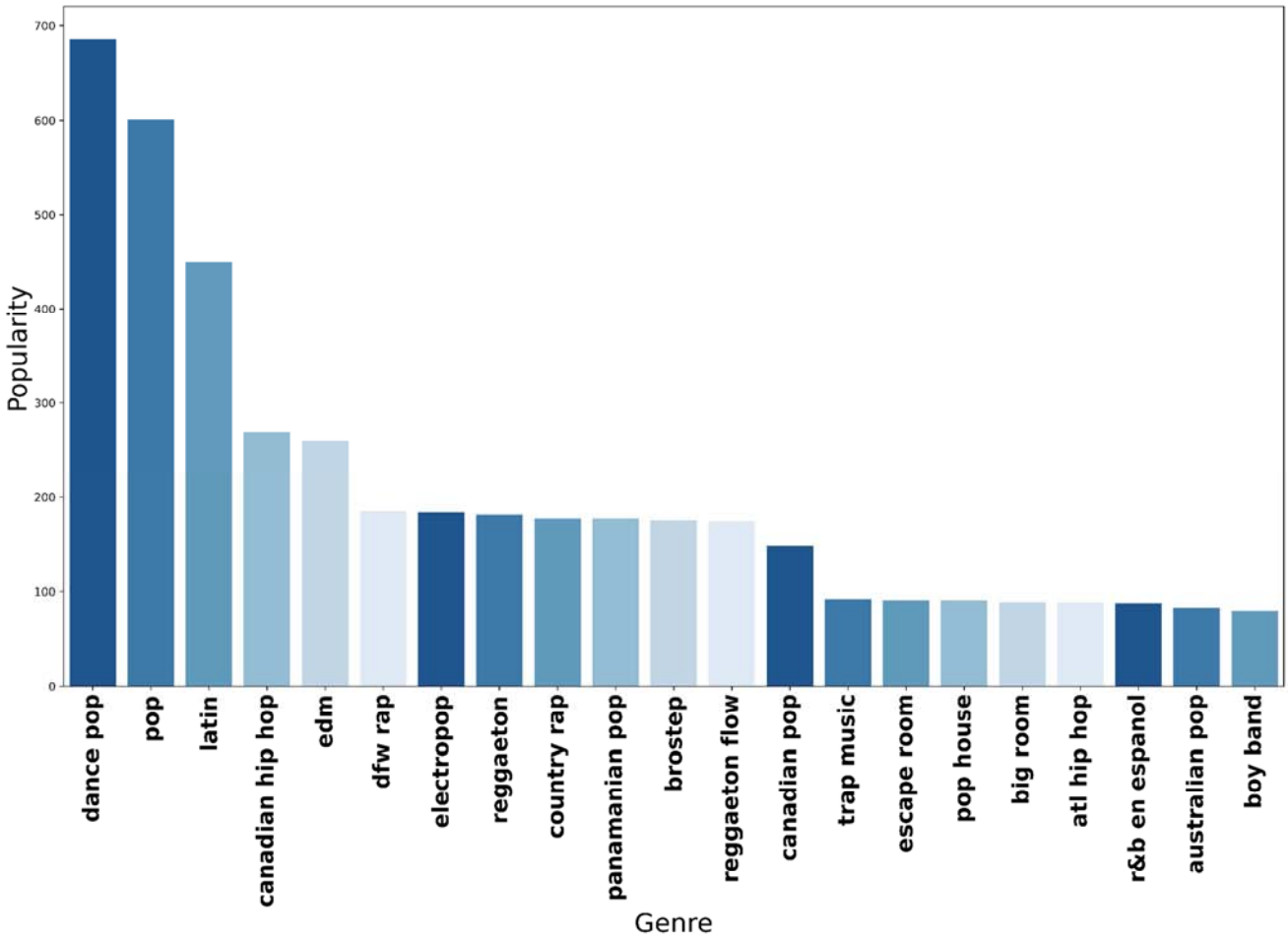


***Figure 3.*** *The genre vs. popularity (2) in 2019.*

## 3. Popularity Prediction Model

This sections shows the implementation the prediction model using machine learning technologies to predict the popularity of music by analyzing the 9 factors of music. The 9 factors of music are already defined in the previous section, and will be explained again in later.

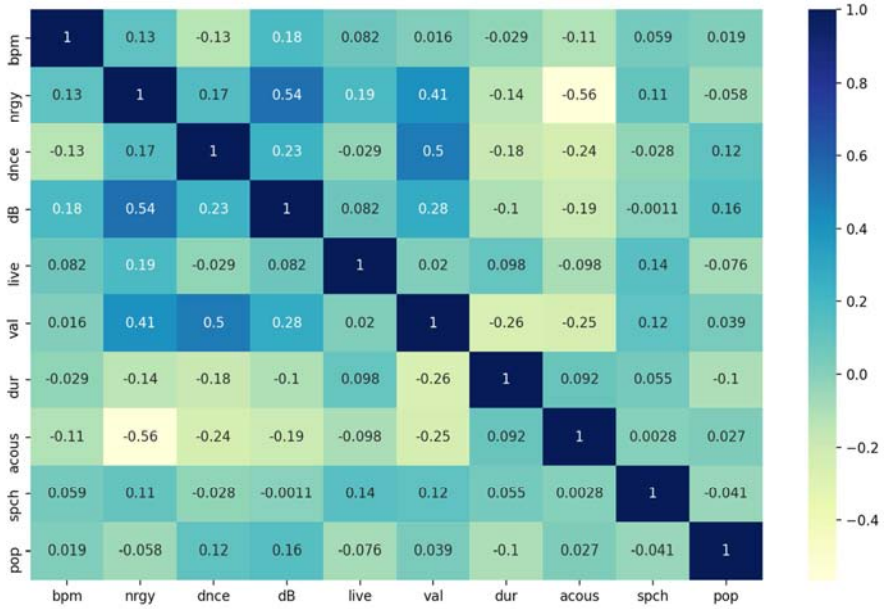


***Figure 4.*** *The heatmap of the 9 factors and popularity.*

### *3.1. Correlation of Factors*

Figure 4 depicts the correlation and relationship between the various factors. As we described in the previous section, music contains 9 factors: bpm (BPM), nrgy (energy), dnce (danceability), loudness (dB), live (liveness), val (valence), dur (length), acoustics (acous), and speech (speechness). 'Pop' in the heatmap indicates the popularity (rank). In the heatmap, if the number inside of each box is closer to 1, the more correlated the two factors are. For example, since the heatmap value for the region that represents the relationship of the loudness (db) and the danceability is 0.16 and is very close to 1, it is considered as very correlated. The factor that has the highest correlation with the popularity is loudness (dB). Danceability, valence and BPM are also important factors for the high rank at Spotify. However, as shown in Figure 4, length (duration), speechness, and liveness are not correlated with popularity (rank).

### *3.2. Model Implementation*

This section shows the implementation a prediction model that can predict the popularity of music by analyzing the 10 factors. The one additional factor is genre. We transformed the genre into integer type (e.g. pop to 1), so genre could be used as a factor to predict the popularity. We used the following algorithms and compare the performances (accuracy) of the prediction:

Linear Regression: Linear regression is the model that predicts the popularity by fitting linear equations in variables. The 9 factors will be transformed into one vector variable, and this will be used as explanatory variables, and the popularity is considered to be the dependent variable [11, 12].

The k-nearest Neighbor regression: K-neighbor regression is a non-parametric method performed using Python in order to analyze the relationship between two variables. As similar to Linear regression, 9 factors of music were transformed into one variable [13].

Random Forest: In random forest, multiple decision trees are created. Each tree predicts the value (popularity) by learning the simple decision rules. Random forest combines the results of those trees to get more accurate and stable predictions of popularity [14, 15].

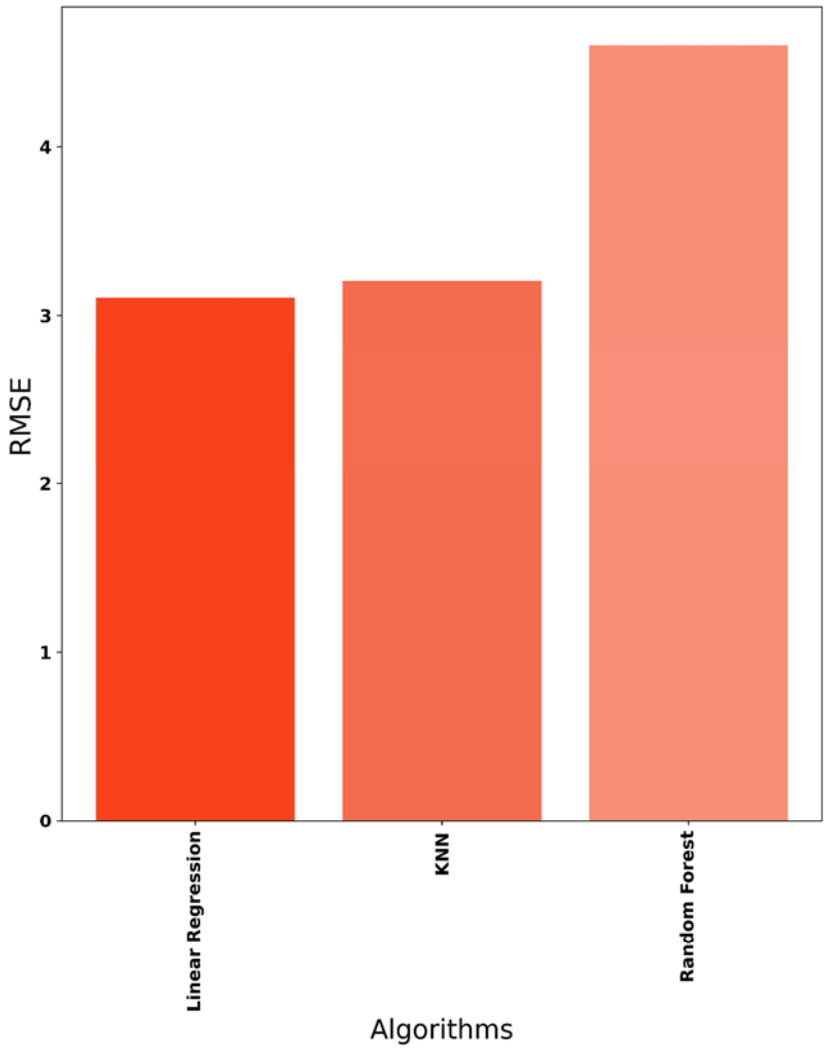


***Figure 4.*** *RSME of three machine learning algorithms in the popularity prediction model.*

Root Mean Square Error is the standard deviation of prediction errors. Lower RMSE means higher accuracy. As shown in Figure 4, Linear regression model has the lowest RSME error which is 3.12 while KNN and Random forest got 3.3 and 4.5. However, the prediction of all three algorithms succeeded in predicting the popularity of music with low RSME. For example, the actual popularity of the song "Beautiful People" is 86. For the prediction popularity, Linear regression model got 88, while KNN and Random Forest model got 88.5 and 81. Therefore, we expect that artists can use this prediction model to predict the popularity in Spotify.

# 4. Conclusion & Future Works

With the emergence of smart phone and streaming services, ranking high in streaming services has become an important factor for better profits and popularity. The trend of music changes every year, so it is crucial to understand the trend to rank high in the streaming services. To solve this issue, we analyze the top 50 musics in 2010-2019 to see the trend of popular music. We also implemented the machine learning model to predict the popularity by analyzing 9 factors of the music. We made the model with the three different machine learning algorithms, and we got the model with 90% of accuracy. We expect that this model can be used by many artists or companies for predicting their music before release. For future works, we are going to expand this project that can predict the popularity of music by analyzing the audio of music instead of text data.

# References

[1] M. Barata, P. Coelho "Music streaming services: understanding the drivers of customer purchase and intention to recommend", Heliyon, August 2021.

[2] Álvarez, Ricardo. "The music industry in the dawn of the 21st century". 2017. 10.13140/RG.2.2.32360.67847.

[3] M. Inter, K. Ka, L. Wang, J. Yang, Z. Yu and N. Ko "Musical trends and predictability of success in contemporary songs in and out of the top charts", May 16 2018.

[4] A. Varshavsky, "Analysis of income inequality impact on the musical art", Journal of the New Economic Association, New Economic Association, 2020.

[5] P. Dicola, "Money from Music: Survey Evidence on Musicians' Revenue and Lessons About Copyright Incentives, Northwestern university School of Law, 2019.

[6] M. Mai, "Death of the Music Long Tail", silpayamanant.2014, https://silpayamanant.wordpress.com/2014/03/07/death-of-the-musical-long-tail/.

[7] "U.S. music streaming revenue 2020", Statista, 2021. https://www.statista.com/statistics/437717/music-streaming-revenue-usa/

[8] Seth A. Carver "Changing the Industry, Spotify", University of Tennessee, 2016.

[9] Fleicher, Rasmus & Snickars, Pelle. "Discovering Spotify - A Thematic Introduction. Culture Unbound": Journal of Current Cultural Research. 9. 2017. 130-145. 10.3384/cu.2000.1525.1792130.

[10] Araujo, Carlos & Cristo, Marco & Giusti, Rafael. "Predicting Music Popularity on Streaming Platforms". 2019. 141-148. 10.5753/sbcm.2019.10436.

[11] scikit-learn developers.. *sklearn. neighbors. kneighborsregressor*. scikit learn. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html.

[12] Kumari, Khushbu & Yadav, Suniti. "Linear regression analysis study". Journal of the Practice of Cardiovascular Sciences. 2018 4. 33. 10.4103/jpcs.jpcs_8_18.

[13] Teixeira-Pinto, A. *2 "*K-nearest Neighbours Regression | Machine Learning for Biostatistic*"*. Biostatistics Statistics Collaboration of Australia. 2021. https://bookdown.org/tpinto_home/Regression-and-Classification/k-nearest-neighbours-regression.html

[14] Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. "Random Forests and Decision Trees". 2012. International Journal of Computer Science Issues (IJCSI). 9.

[15] Gesrad. B "Analysis of a Random Forest Model, Journal of Machine Learning Research, 2012.