International Journal of Data Science and Analysis
2022; 8(5): 149-156
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20220805.15
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Predicting Music Popularity with the Hybrid Approach: K-Means + LGBM

Hyeonsoo Oh

Elective Home Education (EHE/Home-schooling), Seoul, Korea

Email address:
hyeonsoooh6@gmail.com

To cite this article:
Hyeonsoo Oh. Predicting Music Popularity with the Hybrid Approach: K-Means + LGBM. *International Journal of Data Science and Analysis*. Vol. 8, No. 5, 2022, pp. 149-156. doi: 10.11648/j.ijdsa.20220805.15

Received: September 26, 2022; **Accepted**: October 10, 2022; **Published**: October 17, 2022

Abstract: The global revenue from streaming, CD, and digital music sales have exceeded pre-COVID-19 levels since the COVID-19 outbreak. Although other stocks have fallen, stocks relating to the music industry have risen. HYBE entertainment even yielded integrated platform services. Furthermore, there are many people who make music without an agency and post it on platforms such as Soundcloud. Whether the popular music last week can be predicted to be popular this week using the methods we outlined in this paper. We obtained the dataset from Spotify, the main subscription service. The paper has two objectives: predicting popularity and revealing the relationship between K-means and LGBM since there is a paper claiming that the K-means algorithm is efficient in the Spotify dataset. The experiment yielded that the K-means algorithm is not efficient in our dataset by showing less Silhouette score. However, when combining K-means with LGBM, this approach achieved higher performance compared to using LGBM solely. Even if the experiment's result is positive, which could assist in determining whether a composer's songs will be lucrative, we do acknowledge some drawbacks in our methods. For instance, we did not account for the numerous variables introduced by utilizing phony streams to enhance their placement inside the real-time chart. Additionally, we did not include any of the time's top tunes. Christmas theme music, for instance. Throughout the future, we will conduct additional research into this topic to overcome those drawbacks.

Keywords: Music, Machine Learning, K-means, LGBM

1. Introduction

Since the COVID-19 outbreak, other stocks have fallen. However, stocks related to the music industry have risen, such as global revenue from streaming, CD, and digital music sales have surpassed pre-COVID-19 levels. Big Hit Entertainment renamed its company HYBE. For instance, HYBE Entertainment not only increased its sales after Corona but also created a platform by collaborating with IT practitioners to provide convenience to fans [1]. The company grew even more by taking over other agencies during the economic downturn. From 587.22 billion won in 2019 to 796.28 billion won in 2020, sales increased by 35.6 percent [2]. Profits rose by 47%, from 98.74 billion won to 145.51 billion won. Mainstreaming revenue will account for 83% of the total sales in 2021 [3]. The revenue continues to grow from 2010 to 2021. It was $500 million in 2010 and steadily rose to $12.4 billion in 2021 [4].

The most popular streaming app is Spotify. The service is the world's favorite, with 350 million users and 150 million subscribers [3]. This popularity is mainly due to UI and UX, which are superior to other music services. Also, even though music recommendations are now common in other apps, Spotify's system has superior qualities [5]. Therefore, in this work, we focus on Spotify's data set to make forecasts about why the album is popular through machine learning algorithms. As the music industry's influence grows, predicting popularity will be utilized more widely than before. Furthermore, due to the development of music platforms such as Soundcloud, which makes it easier to approach composing, many people compose music these days, and we think this paper can help whether their songs can be successful.

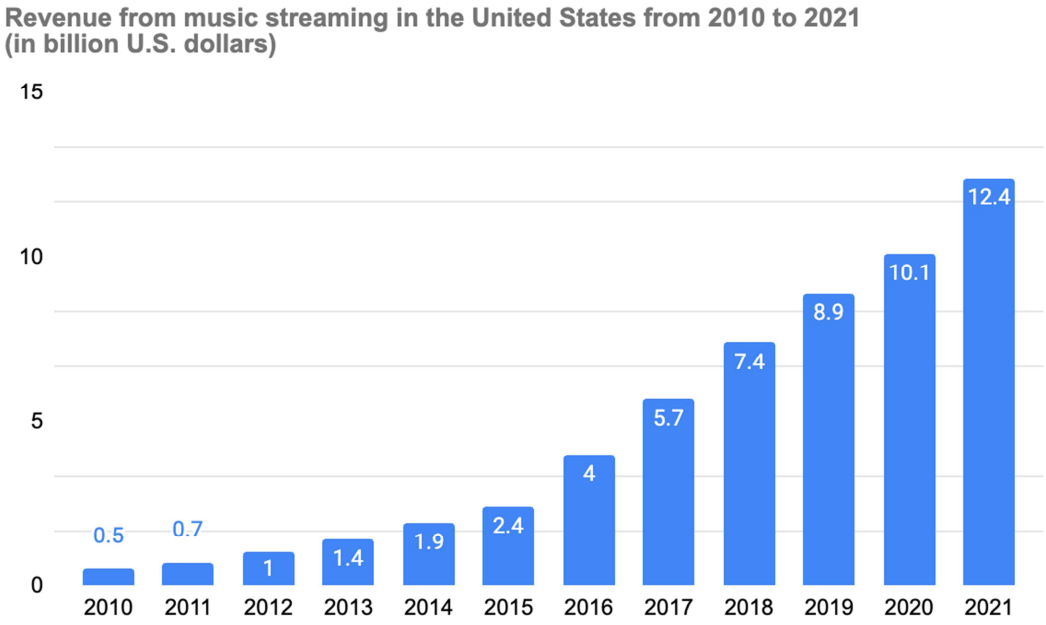

Figure 1. *Revenue from music streaming in the US from 2010 to 2021.*

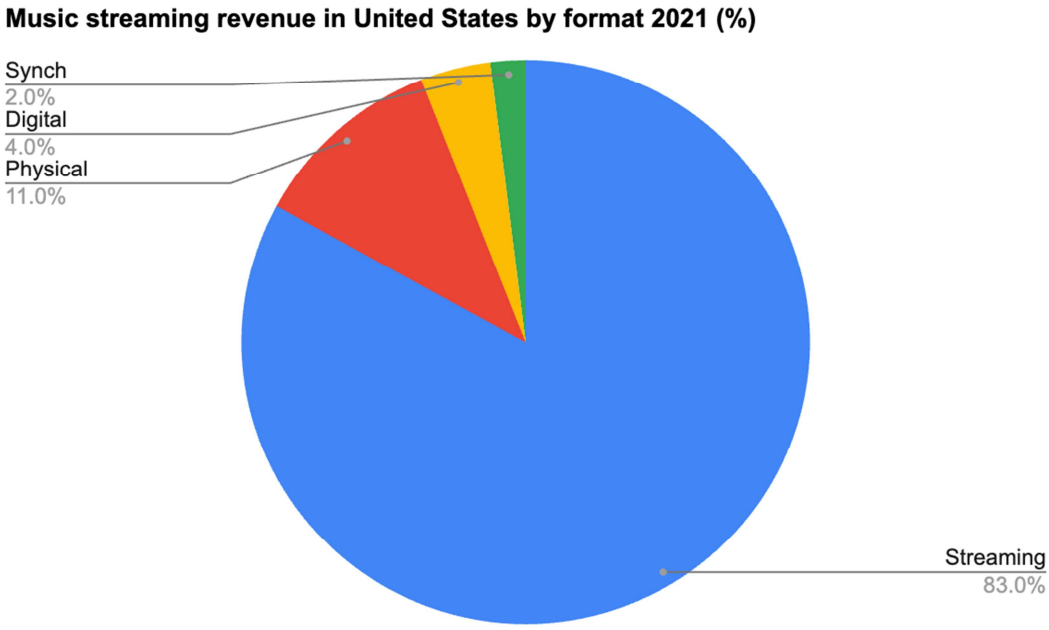

Figure 2. *Music streaming revenue in the US by format 2021.*

2. Related Works

Araujo et al. researched predicting music popularity through machine learning classifiers. They predicted whether the song would be included in the Top 50 charts from Spotify. For the experiment, acoustic features were extracted from the previous songs, and AdaBoost-SAMME, Naive Bayes, random forest, and SVM were utilized as the classifiers. The evaluation was conducted using accuracy, specificity, NPV, and AUC score, and SVM with RBF kernel achieved the highest score, which was higher than 80%. However, they claimed that insufficient datasets were a limitation of this paper, and therefore, they would gain more datasets, including those which require SNS information [6].

Lee et al. proposed that audio signals can predict music popularity patterns. Under this assumption, they collected the datasets from the Billboard Rock Song chart from 2009 to 2014, and 70% of them were used as training datasets, and the rest were for the test sets. These are the target features: Sum, Length, Max, Debut, Kurtosis, and High. Multi-layer perceptrons (MLPs) were used for the classification, and the result from the experiment revealed that the complexity features from the datasets help predict those music popularity patterns [7].

Lee et al. also suggested that music popularity is

significant for both artists and music industries, and therefore, they constructed the experiment to predict music popularity efficiently. The datasets were collected from the Billboard Hot 100 chart and Nielsen Music, and finally, a total of 18,604 songs were collected. Complexity features, including Chroma, Rhythm, Arousal, and MPEG features, were extracted from the datasets. For the experiments, from the extracted features, 70% of them were used as the training datasets, while 30% were used as the test datasets. The SVM was applied to the datasets for binary classification. The result yielded that combining both features (Complexity and MPEG) was more effective than utilizing only certain features. The debut feature was the most important feature for predicting music popularity [8].

Martín-Gutiérrez et al. also predicted music popularity with the AI algorithms. Even though the importance of music information retrieval (MIR) has grown, the existing datasets could not fulfill the demands. Therefore, they proposed SpotGenTrack Popularity Dataset (SPD) as a replacement. Furthermore, they also constructed novel deep learning architecture, HitMusicNet, consisting of two stages: MusicAENet and MusicPopNet. The MusicAENet, based on an autoencoder, compressed the input datasets, and the MusicPopNet, a deep neural network, predicted the music's popularity. Both classification and regression were conducted for the prediction, and the proposed model surpassed the others by showing 0.0118 MSE in regression and 83.02% of the F1 score in classification [9].

Privandhani and Sulastri utilized the K-means and K-medoids algorithm to conduct clustering on the pop songs from the Spotify datasets. For the experiment, features such as danceability, Acousticness, and loudness were extracted from the music. With these features, they utilized both algorithms and then successfully clustered the given datasets. Even though the clustering results are pretty different from the genres of music, the result is still meaningful since the clustering is conducted based on the specific features from the datasets [10].

Those related works proposed various machine learning approaches to music datasets, especially from Spotify. Since the research from Privandhani and Sulastri proved that clustering is meaningful for the streaming datasets, we first conduct clustering on the datasets, extract the features, and merge them into given datasets. Then, machine learning algorithms are applied to the modified datasets to predict the popularity of each music. Combining those two methods, this paper aims to figure out how well clustering works and how clustering affects the performance of the supervised learning algorithms.

3. Materials and Methods

3.1. Dataset Description

The Spotify dataset was obtained from the Kaggle website, consisting of hit songs from 1960 to 2019. The overall dataset can be found in Table 1, and it extracts multiple features from Spotify API, such as 'acousticness,' 'danceability,' 'liveness,' 'energy,' and 'target.' Even though the dataset implies 'track,' 'artist,' and 'url,' those features were not used in the experiment since this paper aims to predict the popularity only with the music-related features. The "target" is a target column in the dataset, and '1' means the song was featured last week, while '0' denotes a flop. Since the target only consists of two values (0 and 1), binary classification was conducted on the dataset [11].

Table 1. *Overview of the dataset used in the experiment..*

acousticness	danceability	liveness	energy	· · ·	target
0.490	0.417	0.0779	0.620	· · ·	1
0.018	0.498	0.1760	0.505	· · ·	0
0.846	0.657	0.1190	0.649	· · ·	0
0.706	0.590	0.061	0.545	· · ·	0

3.2. K-means

The K-means algorithm is one of the machine learning algorithms which belongs to unsupervised learning. Since it is unsupervised learning, it does not require targets for training. The K-means algorithm is conducted as follows: firstly, it sets k number of clusters. Based on the center of k, each data is allocated to those centers based on Euclidean distance. After allocating every data to each cluster, the center k is moved to the average of each cluster. Those steps are repeated since the center k does not move, which means the location of k is equal to the mean value of the cluster. The elbow method can be utilized to find the optimal value k. When the elbow function is utilized, a kink point in the graph indicates the optimal value [12]. These algorithms and functions can be implemented through the Scikit-learn package with Python.

3.3. Experimental Setup

The experiment consists of three different steps. Firstly, the K-means algorithm was applied to the dataset to figure out whether the given dataset could be clustered evenly. Then, multiple machine learning algorithms were utilized with the clustered datasets to evaluate the performance. Lastly, those algorithms were utilized again to the original datasets and compared the results to the second experiment.

4. Result

4.1. Clustering

The figures below are the output from the K-means

algorithms. Figure 3 depicts the counts of four different clusters from the K-means clustering, and Figure 4 exhibits the Sihouette plot. From those figures, it could be concluded that the clustering did not perform well. since there is a large variation between clusters (most of the data belongs to Cluster 3), and the Silhouette score was close to 0. Even though the research from Privandhani and Sulastri showed that the K-means clustering performed well on the Spotify datasets, different results were yielded from this experiment.

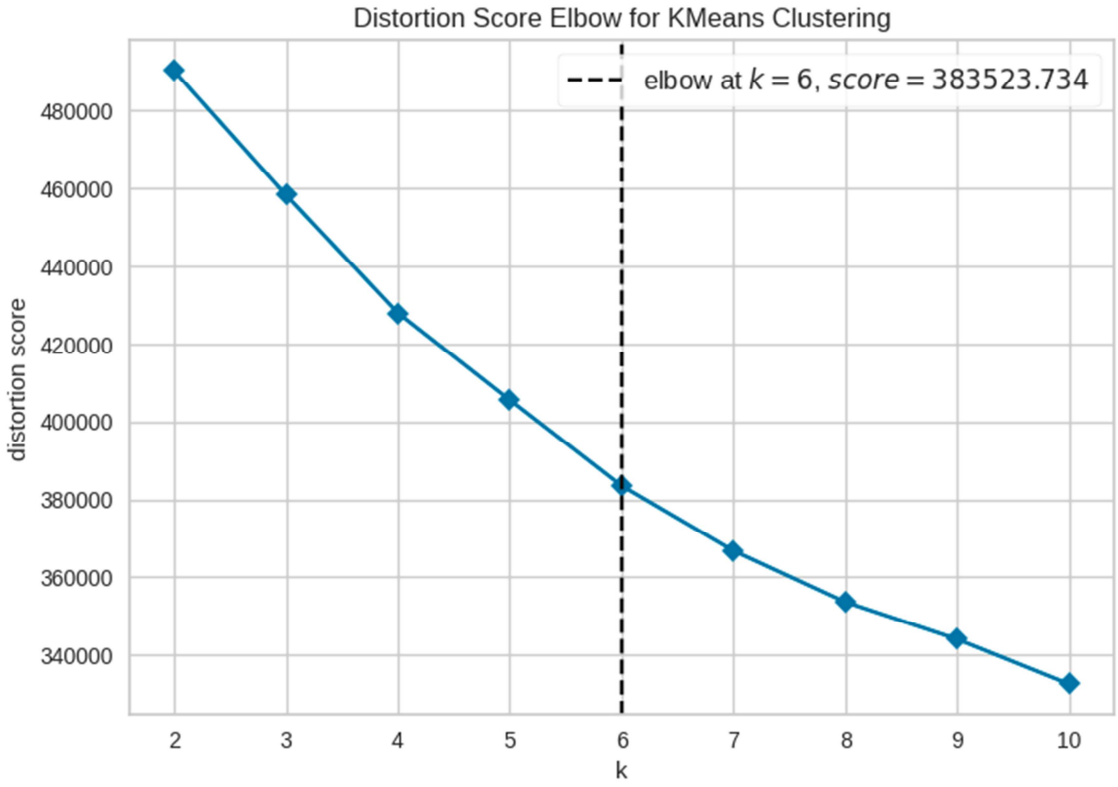

Figure 3. *Example of the result from the elbow method: 6 is an optimal number of k in figure.*

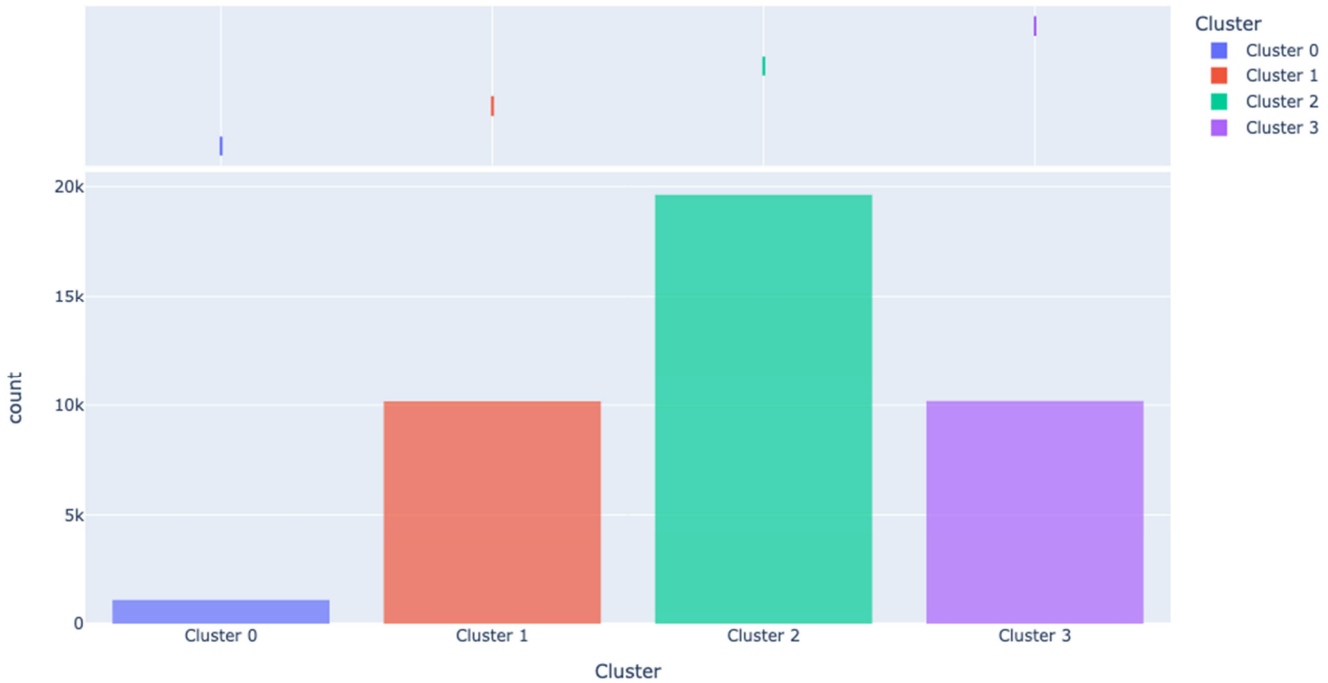

Figure 4. *Visualization of the counts of each cluster obtained by K-means clustering.*

4.2. Classification: LGBM

Multiple machine learning algorithms were utilized for the classification, such as an extra tree (ET), light gradient boosting machine (LGBM), gradient boosting classifier (GBC), ada boost classifier (ADA), and logistic regression (LR). Since the binary classification was conducted, the algorithms were evaluated by accuracy score, precision score, recall score, AUC score, and F1 score. However, our proposed model, LGBM, achieved the highest performance by showing 78.60, 86.49, 84.94, 75.38, and 79.87, respectively. This result, in particular, is 20% higher in most scores than logistic regression.

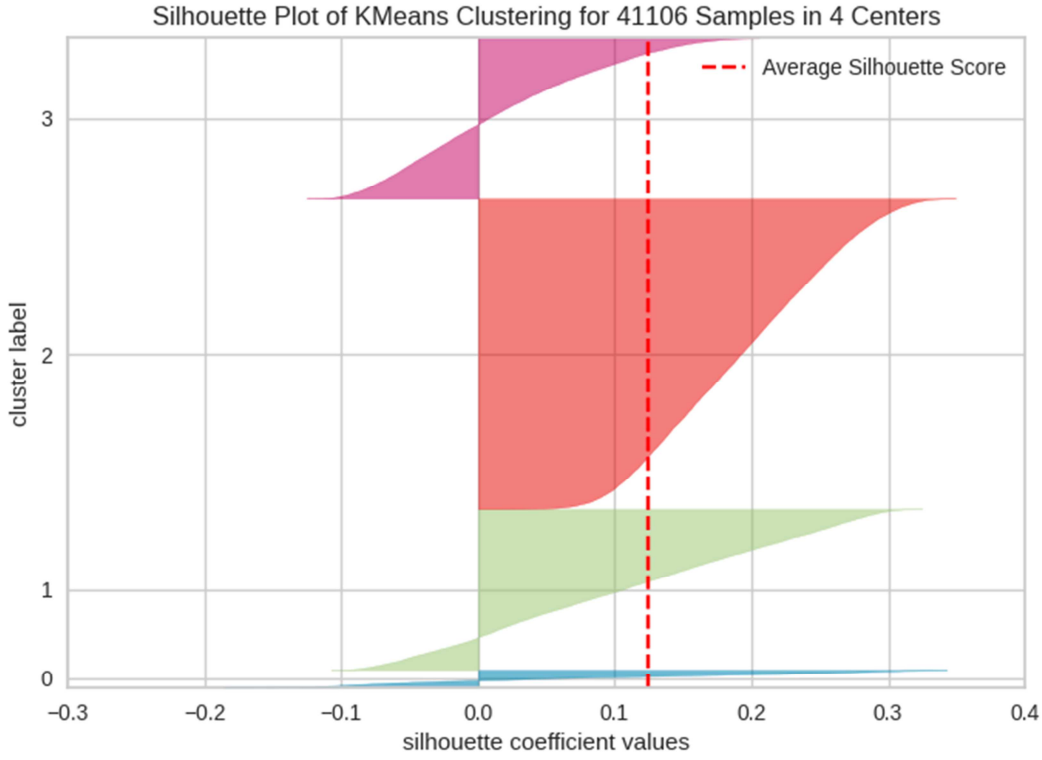

Figure 5. *Silhouette plot of K-means clustering.*

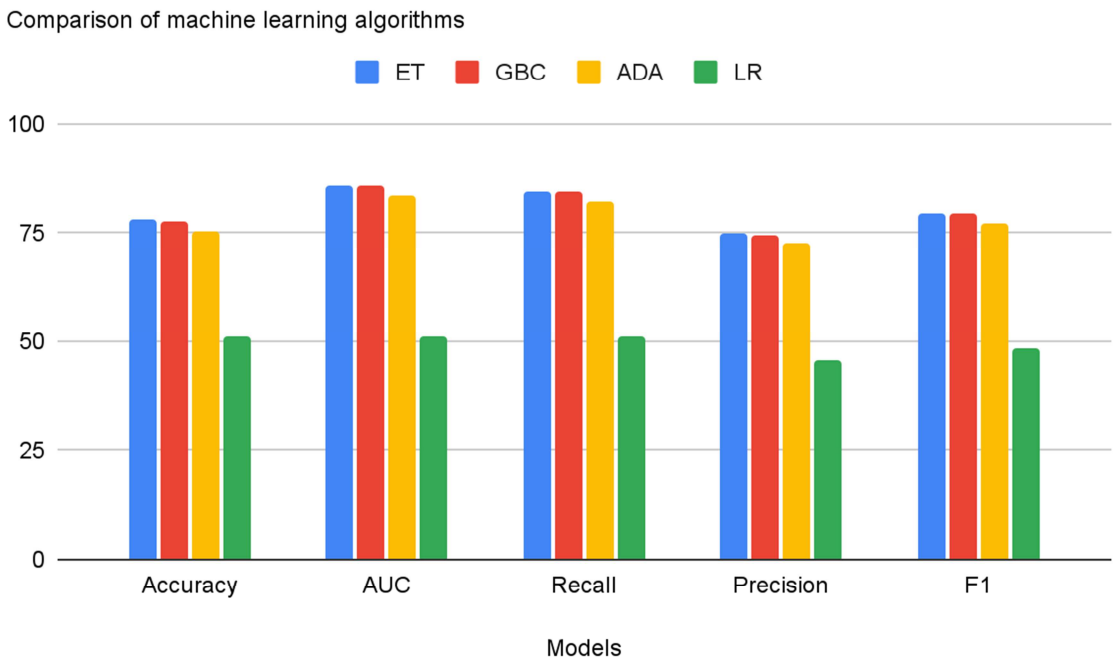

Figure 6. *Comparison of machine learning algorithms: extra tree, gradient boosting classifier, ada boost classifier, and logistic regression.*

4.3. Classification: K-means + LGBM

The third experiment combines K-means and LGBM algorithms to compare the results to the second experiment. Figure 7 shows that the proposed model (K-means + LGBM) achieved slightly higher scores in all evaluation indices. For the F1 score, the proposed approach increased by 3% and 2% for the rest. From this result, it could be concluded that combining the K-means algorithm could enhance the performance of the LGBM. Even if the K-means algorithm did not perform as expected, this slight improvement could show a possibility if clustering performance was better, combining it with other supervised learning algorithms would increase the accuracy. Figure 8 depicts the ROC curve, and Figure 9 shows the confusion matrix of the result, which are both related to those evaluation indices.

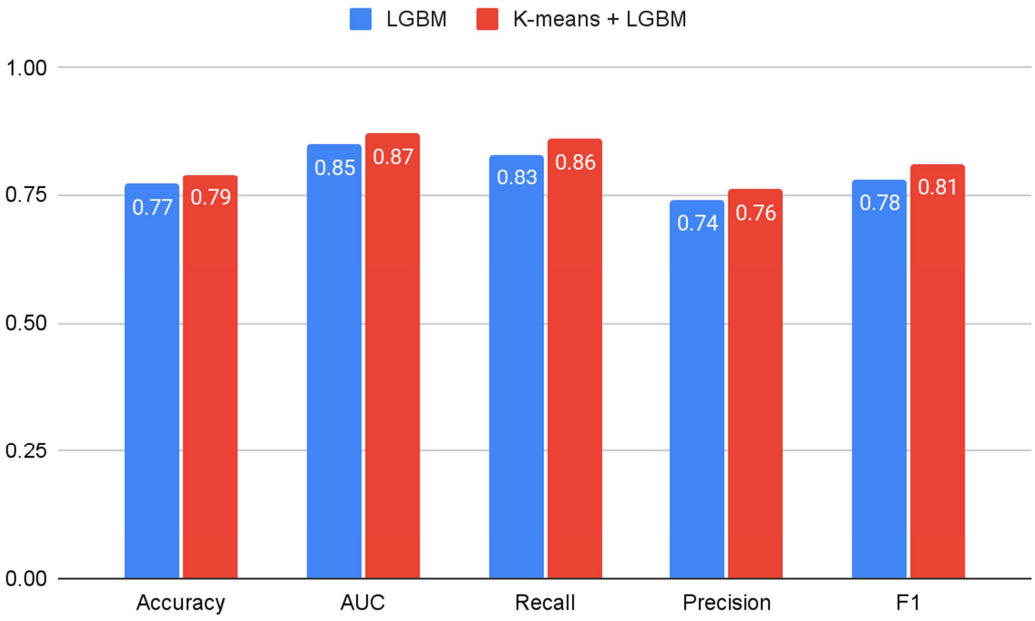

Figure 7. *Comparison of the results between the second and third experiments: blue indicates the second one and red for the third one.*

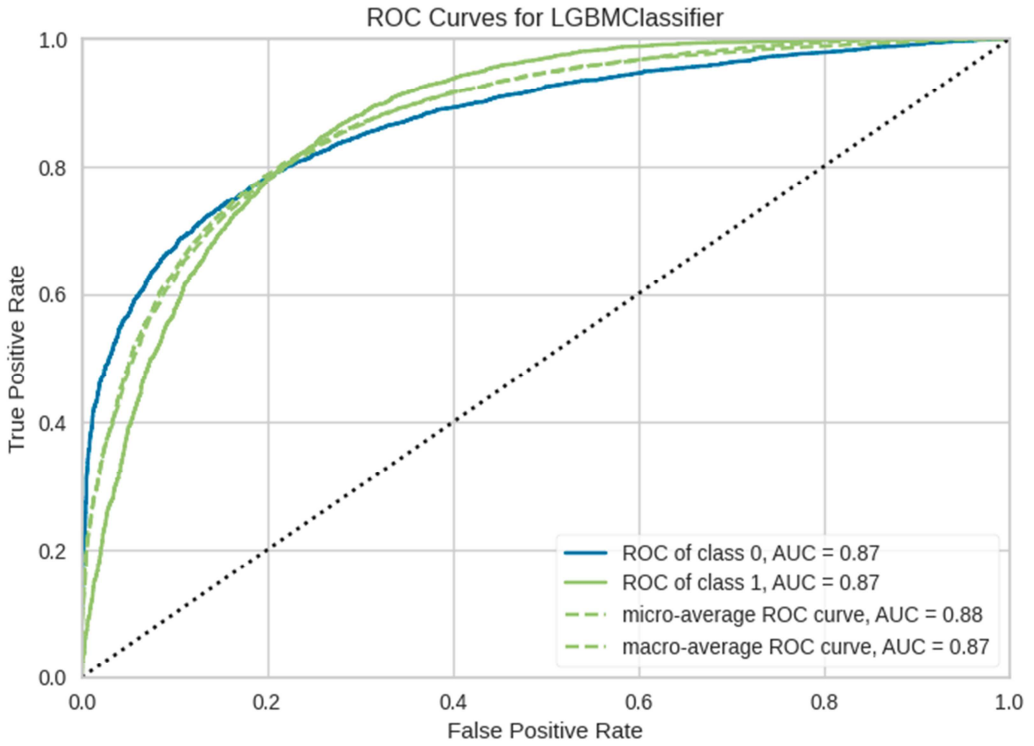

Figure 8. *ROC curve for the proposed approach.*

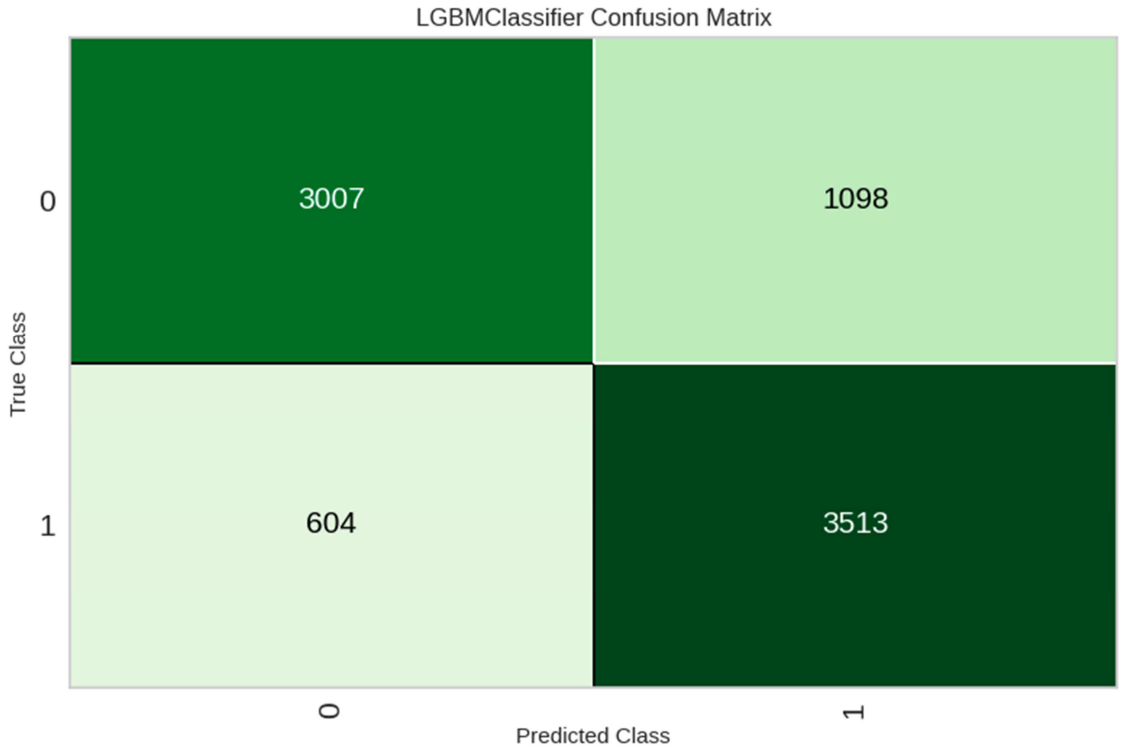

Figure 9. *Concussion matrix of the results from the proposed model.*

5. Discussion

Through the proposed algorithm, this paper successfully predicts the popularity of the songs. Furthermore, this paper is meaningful since we combine two machine learning approaches: unsupervised and supervised learning. Since the proposed methods (K-means + LGBM) achieved a high score, and the important features included "Cluster_Cluster 1" and "Cluster_Cluster 3", which were extracted from the K-means, it could be concluded that K-means enhance the performance of the prediction. Furthermore, this could prove the statement from Privandhani and Sulastri's research that the K-means are effective in the music datasets. However, there are still limitations in our research. Firstly, there can be many variables inside the music charts, people only check out the top charts, and the Artists and musicians are tempted to stream themselves to rank up their rankings on streaming sites. As a result, real-time charts and rankings have a high potential for steam manipulation. Secondly, we did not reflect on the favored songs of that time. For example, at Christmas, music related to Christmas that fits the mood is popular, and there are times when the music becomes famous and favored by many people after the program is over.

6. Conclusion

In this paper, we presented a methodology to predict whether the music popular last week will be popular this week. We collected the data from the streaming platform Spotify because it is the most popular music platform these days. The experiment consists of three steps: clustering through K-means, applying LGBM, and combining K-means and LGBM. The proposed algorithm (K-means + LGBM) yielded 79%, 87%, 86%, 76%, and 81% for accuracy, AUC, recall, precision, and F1-score, respectively. This result revealed that combining those two methods could enhance the performance and be applied more widely to different datasets. Even if the result of the experiment is successful, which could be helpful for whether a composer's songs can be successful, we do identify some limitations in our methodology. For instance, we did not reflect that using fake streams to boost their ranking inside the real-time chart introduces many variables. Additionally, we did not include popular songs from that period. Music with a holiday theme, for example. We will research this topic in further research.

References

[1] BBC News. (2022, February 23). K-pop: BTS agency Hybe grows profits by 31%. Retrieved September 26, 2022, from https://www.bbc.com/news/business-60488623

[2] K-pop's biggest agencies see rise in profits despite pandemic. (2021, March 30). Retrieved September 26, 2022, from https://koreajoongangdaily.joins.com/2021/03/30/entertainment/kpop/SM-entertainment-JYP-entertainment-YG-entertainment/20210330160916443.html

[3] Music Streaming App Revenue and Usage Statistics (2022). Business of Apps. (2022, September 12). Retrieved September 26, 2022, from https://www.businessofapps.com/data/music-streaming-market/

[4] K-pop's biggest agencies see rise in profits despite pandemic. Korea joongAng Daily. (2021, March 30). Retrieved September 21, 2022, from https://koreajoongangdaily.joins.com/2021/03/30/entertainment/kpop/SM-entertainment-JYP-entertainment-YG-entertainment/20210330160916443.html

[5] Watkins, C. (2022, January 19). How Spotify's user experience is helping them win the streaming wars. Medium. Retrieved September 26, 2022, from https://uxdesign.cc/ux-ui-analysis-spotify-31f3855a1740

[6] Soares Araujo, C. V., Pinheiro de Cristo, M. A., & Giusti, R. (2019, December). Predicting Music Popularity Using Music Charts. 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA). https://doi.org/10.1109/icmla.2019.00149

[7] Lee, J., & Lee, J. S. (2015). Predicting Music Popularity Patterns based on Musical Complexity and Early Stage Popularity. Proceedings of the Third Edition Workshop on Speech, Language &Amp; Audio in Multimedia - SLAM '15. https://doi.org/10.1145/2802558.2814645

[8] Lee, J., & Lee, J. S. (2018, November). Music Popularity: Metrics, Characteristics, and Audio-Based Prediction. IEEE Transactions on Multimedia, 20 (11), 3173–3182. https://doi.org/10.1109/tmm.2018.2820903

[9] Martin-Gutierrez, D., Hernandez Penaloza, G., Belmonte-Hernandez, A., & Alvarez Garcia, F. (2020). A Multimodal End-to-End Deep Learning Architecture for Music Popularity Prediction. IEEE Access, 8, 39361–39374. https://doi.org/10.1109/access.2020.2976033

[10] Privandhani, N. A. (2022). Clustering Pop Songs Based On Spotify Data Using K-Means And K-Medoids Algorithm. Jurnal Mantik, 6 (2), 1542-1550.

[11] The Spotify Hit Predictor Dataset (1960-2019). (2020, April 26). Kaggle. Retrieved September 26, 2022, from https://www.kaggle.com/datasets/theoverman/the-spotify-hit-predictor-dataset

[12] Yuan, C., & Yang, H. (2019, June 18). Research on K-Value Selection Method of K-Means Clustering Algorithm. J, 2 (2), 226–235. https://doi.org/10.3390/j2020016