Institut Riset dan Publikasi Indonesia (IRPI)
IJATIS: Indonesian Journal of Applied Technology and Innovation Science
Journal Homepage: https://journal.irpi.or.id/index.php/ijatis
Vol. 2 Iss. 1 February 2025, pp: 7-12
ISSN(P): 3032-7466 | ISSN(E): 3032-7474

Implementation of Supervised Learning Algorithm on Spotify Music Genre Classification

Muhammad Fiqri[1*], Farhan Bin Siddik Lizen[2], Muhammad Ikrom[3]

[1]Faculty of Education and Literature, Universitas of Tabuk, Saudi Arabia
[5]Faculty of Scince, Nowab Foyzunnesa Govt Collage, Bangladesh
[123]Dapartment of Information System, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia

Email: [1]Muhammadfiqri587@gmail.com,
[2]Farhanlizen@gmail.com, [3]12050316301@students.uin-suska.ac.id,

Received Dec 27th 2023; Revised Nov 16th 2024; Accepted Nov 28th 2024; Available Online Feb 28th 2025, Published Feb 28th 2025
Corresponding Author: Muhammad Fiqri

Abstract

Spotify is a music streaming application that has been around since 2008. In the application, users can compile a playlist of songs they want to listen to. Users can determine the name of the singer, type of music, music genre and tempo of the music they want to listen to play as needed. The genre received by each user from his device will produce different recommendations, this is due to the classification process based on music listening behavior, such as songs that are often, rarely, or even never listened to or played at all by users. Therefore, the process of classifying music genres on spotify with the help of machine learning using supervised learning algorithms with algorithms namely Naïve Bayes, K-Nearest Neighbors (K-NN), Random Forest and Decision Tree with the aim of comparing the accuracy of each algorithm so as to get the best model for calcification. The results of this study obtained Random Forest has the highest accuracy value of 79.40%, followed by Decision Tree at 79.30%. In the next position Naïve Bayes with an accuracy value of 77.28%, the algorithm with the lowest accuracy is K-NN with an accuracy value of 60.74%. Meanwhile, evaluation with the t-test algorithm with the best performance is obtained from the Random Forest algorithm with a value of 0.794. It can be concluded that the best algorithm in music genre classification on Spotify is using Random Forest.

Keyword: Decision Tree, K-NN, Music Genre, Naïve Bayes, Random Forest, Spotify

1. INTRODUCTION

Spotify is a business that provides music streaming services located in Sweden, founded in 2008. Many customers are satisfied with the services provided by Spotify [1]. In addition, Spotify is also a highly innovative application that successfully utilizes modern technological advances to continuously improve user satisfaction. This is achieved through intelligent data collection, such as the automatic generation of personalized playlists every week [2]. These playlists are composed with different songs that are tailored to the mood of each playlist and also take into account the history of songs that have been heard by customers [3].

Music consumption today is highly customizable and accessible through various devices like smartphones, laptops, and portable tablets. Users' music collections are aggregated into a centralized database managed by service providers. Sorting and recommending music involve criteria such as artist/singer name, album/single type, release year, music genre label, and curated playlists. However, current methods may fail to provide accurate recommendations due to variations in user behavior, where some songs are frequently listened to while others are rarely or never played. This discrepancy necessitates a more precise classification system to tailor recommendations effectively.

Previous research has explored music genre classification using machine learning algorithms with varying degrees of success. For instance, Naïve Bayes achieved 58.91% accuracy for Spotify-based classifications, followed by Random Forest at 52.81%, and K-Nearest Neighbors (KNN) at just 33.30%. To address these limitations, researchers employ machine learning techniques to analyze complex patterns in user behavior and musical attributes. Studies on spatial modeling and zoning of landslide vulnerabilities have demonstrated high accuracies ranging from 86% to 89% using similar algorithms. Additionally, chronic kidney disease predictions showed superior performance with Random Forest compared to Naïve Bayes and KNN, indicating the robustness of certain machine learning approaches.

The primary motivation behind conducting this research lies in developing a sophisticated system capable of offering personalized and reliable music recommendations. By leveraging advanced machine learning algorithms, we can ensure that users receive tailored suggestions that align closely with their preferences, thereby enhancing overall musical experiences through improved customization and relevance. This study aims to address the challenges inherent in modern music consumption by harnessing the capabilities of machine learning techniques such as Naïve Bayes, Random Forest, and KNN. By optimizing these methods based on previous successes across diverse applications, we aim to create a robust framework capable of delivering highly accurate and personalized music recommendations tailored specifically to individual tastes.

In this study, researchers will classify music genres on spotify using the Naïve Bayes, K-NN, Random Forest and Decision Tree algorithms which in addition to the results of the classification of music genre types will compare the accuracy of each algorithm.

2. MATERIAL AND METHOD

In this research, the data used is Kaggle data in the form of Song of Spotify datasets. The attributes used in this study are 12 criteria which include danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, livenees, valence, tempo, and duration_ms. After the initial dataset is collected, the preprocessing stage is carried out, before the classification stage is carried out. Next, the data will be divided using Cross Validation, then enter the classification process using Naïve Bayes, K-NN, Random Forest, and Decision Tree. Then the results of each experiment are compared so that the best results are obtained based on accuracy and Kappa. The stages of the research methodology can be seen in Figure 1.

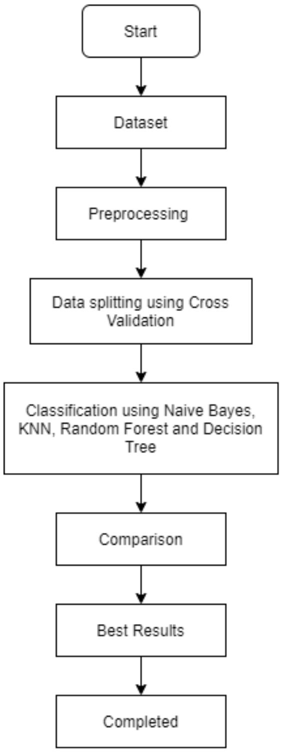

Figure 1. Research Methodology

2.1. Naïve Bayes

Naïve Bayes is a data mining algorithm that applies Bayes theory for classification. Naïve Bayes has accuracy when applied to data owners with large enough data [7]. Naïve Bayes shows fast performance and high accuracy when applied to large datasets. When performing classification, this algorithm looks for the highest probability of all document categories being tested. [8]. The initial stage in the classification process is done by calculating the mean value and standard deviation of the features in the training dataset for each class using equation 1 [9].

$$P(Ci|X) = \frac{P(X|Ci).P(Ci)}{P(X)} \quad (1)$$

2.2. K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) is a standard instance-based machine learning method that contains large-scale data sets [10]. K-NN uses a clustering approach to identify categories that are used to classify new objects based on training data that is closest to them [11]. Simply put, K-NN classifies the test object by considering the predicted value as the majority result of the categories of training objects that are in the K closest group [12]. The algorithm operates by finding the closest proximity, or Euclidean distance of the data being processed, which can be calculated using equation 2.

$$d_{euclidean}(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \quad (2)$$

2.3. Random Forest

Random Forest is a technique often used in classification and clustering based on the merging of multiple decision trees [13]. Classification decisions are made based on the majority vote of all trees in the ensemble. In Random Forest, decision trees are built in a randomized manner, allowing for variation in the formation of each tree [14]. The advantages of Random Forest include its ability to overcome overfitting, good performance on large datasets, and the ability to handle both categorical and numerical data [15].

2.4. Decision Tree

Decision Tree is an algorithm in computer science and statistics that is used to make decisions based on given conditions [16]. This algorithm works by dividing the dataset into smaller subsets based on existing features [17]. This method attempts to find a function that has a discrete value and is resistant to data errors (data interference) and can learn disjunctive expressions [18]. The equation for finding the value of the decision tree is using the gain and entropy values. The following equation 3 and equation 4 calculates the entropy and gain values.

$$\text{Entropy} = \sum_{i=1}^{n} pi * \log_2 pi \quad (3)$$

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * \text{Entropy}(S_i) \quad (4)$$

2.5. Cross Validation

Cross-validation is a statistical algorithm technique that involves dividing the dataset into two main parts, namely training data and testing data. The main objective of this method is to validate the proposed model using standardized testing to assess the error rate [19]. Cross Validation is used to evaluate model performance and serves as a determinant of the division ratio between training and testing data. The use of the Cross Validation Method can provide reliable accuracy estimates for a data set. In this research, 10-Fold Cross Validation is used to produce reliable accuracy estimates and reduce potential bias. [20]. 10-Fold Cross Validation works by training the model on a subset of training data and testing the model on a subset of testing data. By dividing the data into 10-folds of equal size, 10-Fold Cross Validation creates 10 subsets of data to evaluate the performance of the model. This process is repeated 10 times for comprehensive testing [21].

3. RESULTS AND DISCUSSION

3.1. Data Processing

In this research, the dataset used comes from Kaggle and is a dataset of songs from Spotify, with a total of 32,834 data. This dataset includes 7 types of genres, including pop, rock, r&b, rap, Latin, and edm. Next, data analysis was conducted to understand the variables before conducting thorough data processing. After collecting the dataset, the next step is to perform data preprocessing, which includes data cleaning and data normalization. The datasets used in this study can be seen in Table 1.

Table 1. Dataset Song of Spotify

playlist_genre	danceability	Energy	key	loudness	mode	speechiness	acousticness	…	duration_ms
pop	0.748	0.916	6	-2.634	1	0.0583	0.102	…	194754
pop	0.726	0.815	11	-4.969	1	0.0373	0.0724	…	162600
pop	0.675	0.931	1	-3.432	0	0.0742	0.0794	…	176616
pop	0.718	0.93	7	-3.778	1	0.102	0.0287	…	169093
pop	0.65	0.833	1	-4.672	1	0.0359	0.0803	…	189052
pop	0.675	0.919	8	-5.385	1	0.127	0.0799	…	163049
pop	0.449	0.856	5	-4.788	0	0.0623	0.187	…	187675
pop	0.542	0.903	4	-2.419	0	0.0434	0.0335	…	207619
pop	0.594	0.935	8	-3.562	1	0.0565	0.0249	…	193187
…	…	…	…	…	…	…	…	…	…
edm	0.609	0.989	2	-3.51	1	0.0867	0.4678	…	171697
edm	0.545	0.78	7	-4.867	0	0.0436	0.0309	…	255093
edm	0.559	0.916	11	-3.05	1	0.0626	0.0453	…	215295
edm	0.645	0.832	2	-5.595	1	0.0294	0.00106	…	188371
edm	0.581	0.64	5	-8.367	1	0.0365	0.0266	…	196993
edm	0.428	0.922	2	-1.814	1	0.0936	0.0766	…	204375

playlist_genre	danceability	Energy	key	loudness	mode	speechiness	acousticness	…	duration_ms
edm	0.522	0.786	0	-4.462	1	0.042	0.00171	…	353120
edm	0.529	0.821	6	-4.899	0	0.0481	0.108	…	210112
edm	0.626	0.888	2	-3.361	1	0.109	0.00792	…	367432
edm	0.603	0.884	5	-4.571	0	0.0385	0.3567	…	337500

3.2. Modeling

In the modeling stage, the data processing process uses Rapidminer tools, in this study using four classification algorithms to compare the reliability in processing the dataset. The initial step involves preparing the data that will enter the modeling stage. In this research, the dataset used is the Spotify song dataset. Incomplete or imperfect data is then corrected by adding operators to declare missing values and replace empty values. In the modeling stage, researchers used four classification algorithms namely Naive Bayes, K-NN, Random Forest, and Decision Tree, with the aim of generating decisions from the classification of music genres on Spotify. Data processing can be described by the design in Figure 2.

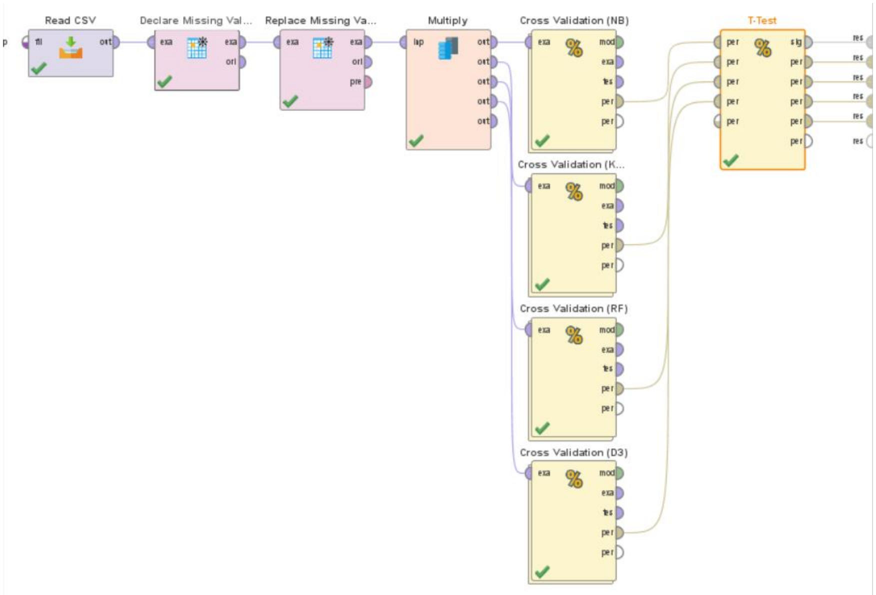

Figure 2. Data Mining Modeling for RapidMiner

3.3. Evaluation

The assessment of the classification model relies on trials to estimate the correct and incorrect objects. The output of the cross-validation testing step is the accuracy value obtained from the confusion matrix. The results of processing and testing using the Cross Validation technique produce the following evaluation can show table 2.

Table 2. Evaluation algorithm

Performance	Naïve Bayes	K-NN	Random Forest	Decision Tree
Accuracy	77.28%	60.74%	79.40%	79.30%
Kappa	0.534	0.195	0.576	0.575

Based on the data above, it can be seen that Random Forest has the highest accuracy value of 79.40%, followed by Decision Tree at 79.30%. The next position is Naïve Bayes with an accuracy value of 77.28%, the algorithm with the lowest accuracy is K-NN with an accuracy value of 60.74%.

Based on the data presented, we can observe that Random Forest achieves the highest accuracy value of 79.40%, making it the most effective algorithm among those tested. Random Forest, an ensemble learning method, typically performs well because it combines the predictions of multiple decision trees, helping to mitigate overfitting and improve the generalization of the model. This characteristic likely contributes to its superior accuracy in this particular classification task.

The second-best algorithm is Decision Tree, with an accuracy of 79.30%. While slightly lower than Random Forest, Decision Tree remains a strong performer, as it is a simple yet effective model for classification tasks. However, Decision Trees are prone to overfitting, especially with complex datasets, which may explain why its accuracy is just marginally lower than that of Random Forest. Decision Trees are also sensitive to the choice of splitting criteria and depth, which can further impact performance.

The third position is held by Naïve Bayes, which achieved an accuracy of 77.28%. Naïve Bayes performs well for problems with strong conditional independence between features. However, in real-world data, features are often correlated, which may limit the performance of Naïve Bayes compared to more complex algorithms like Random Forest and Decision Tree. Despite its simplicity and speed, the Naïve Bayes model struggles when the independence assumption is violated, which might explain the relatively lower accuracy compared to the top two models.

The algorithm with the lowest accuracy is K-NN, with an accuracy of 60.74%. K-NN, a non-parametric method, can be quite sensitive to the choice of the number of neighbors (k) and the distance metric used. Additionally, K-NN is computationally expensive and may not perform well when the dataset is large or contains noisy data. The lower accuracy of K-NN could be attributed to these factors, as well as the potential difficulty in distinguishing between classes when the feature space is not well-separated.

In conclusion, Random Forest emerges as the most reliable algorithm in terms of accuracy, likely due to its ensemble approach and ability to handle complex data. Decision Tree is also a strong performer but is more susceptible to overfitting. Naïve Bayes is efficient but less effective when feature dependencies exist, and K-NN, while simple, is the least accurate, likely due to its sensitivity to hyperparameters and computational challenges. The choice of the most suitable algorithm should therefore depend on the characteristics of the dataset and the computational constraints.

4. CONCLUSION

Based on the results of a series of music genre classification experiments on Spotify using Naive Bayes, K-NN, Random Forest, and Decision Tree algorithms, the best performance was obtained using the Random Forest algorithm with an accuracy of 79.40%. It can be concluded that the best algorithm for music genre classification on Spotify is Random Forest. The strength of this research is the ability to compare different algorithms in the context of music classification, providing insight into the performance of each algorithm and helping to select the optimal algorithm for music genre classification applications. In addition, this research offers a better understanding of the use of statistical evaluation techniques such as t-tests to assess the significance of performance differences between algorithms. Suggestions for the future include conducting further experiments with additional algorithms, including deep learning techniques such as neural networks, which may provide better results for more complex music datasets. In addition, adjustments to the parameters of each algorithm can be made to optimise its performance. Future research could also focus on enhancing the features used in the model to improve the accuracy of music genre classification.

REFERENCES

[1] A. S. Rahayu, A. Fauzi, and Rahmat, “Komparasi Algoritma Naïve Bayes Dan Support Vector Machine (SVM) Pada Analisis Sentimen Spotify,” J. Sist. Komput. dan Inform., vol. 4, no. 2, pp. 349–354, 2022, doi: 10.30865/json.v4i2.5398.

[2] A. R. Zaidah, C. I. Septiarani, M. S. Nisa, A. Yusuf, and N. Wahyudi, “KOMPARASI ALGORITMA K-MEANS , K-MEDOID , AGGLOMEARTIVE CLUSTERING TERHADAP GENRE SPOTIFY,” J. Ilm. Ilmu Komput., vol. 7, no. 1, pp. 49–54, 2021.

[3] S. Navisa, L. Hakim, and A. Nabilah, “Komparasi Algoritma Klasifikasi Genre Musik pada Spotify Menggunakan CRISP-DM,” J. Sist. Cerdas, vol. 04, no. 02, pp. 114–125, 2021.

[4] S. Ahmed, A. El, M. Sk, A. Ali, and Q. Bao, “Spatial modeling and susceptibility zonation of landslides using random forest , naïve bayes and K - nearest neighbor in a complicated terrain,” Earth Sci. Informatics, pp. 1227–1243, 2021, doi: 10.1007/s12145-021-00653-y.

[5] R. Devika, S. V. Avilala, and V. Subramaniyaswamy, “Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes , KNN and Random Forest,” 2019 3rd Int. Conf. Comput. Methodol. Commun., no. Iccmc, pp. 679–684, 2019.

[6] A. Purnamawati, W. Nugroho, D. Putri, and W. F. Hidayat, “Penyakit Daun pada Tanaman Padi Menggunakan Algoritma Decision Tree , Random Forest , Naïve Bayes , SVM dan KNN,” InfoTekJar J. Nas. Inform. dan Deteksi, vol. 1, no. 1, pp. 212–215, 2020.

[7] A. M. Rahat, A. Kahir, A. Kaisar, and M. Masum, “Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset,” 8th Int. Conf. Syst. Model. Adv. Res. Trends, 2019.

[8] S. Chen, G. I. Webb, L. Liu, and X. Ma, “Knowledge-Based Systems A novel selective naïve Bayes algorithm ☆,” Knowledge-Based Syst., no. xxxx, p. 105361, 2019, doi: 10.1016/j.knosys.2019.105361.

[9] I. Wickramasinghe and H. Kalutarage, “Naive Bayes : applications , variations and vulnerabilities : a review of literature with code snippets for implementation,” Soft Comput., no. 1989, 2020, doi: 10.1007/s00500-020-05297-6.

[10] N. Hidayati and A. Hermawan, “K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation,” J. Eng. Appl. Technol., vol. 2, no. 2, pp. 86–91, 2021.

[11] R. Andrian, M. A. Naufal, B. Hermanto, A. Junaidi, and F. R. Lumbanraja, “k-Nearest Neighbor (k-

NN) Classification for Recognition of the Batik Lampung Motifs," J. Phys. Conf. Ser. Pap., pp. 1–6, 2019, doi: 10.1088/1742-6596/1338/1/012061.
[12] D. M. P. Murti, A. P. Wibawa, and M. I. Akbar, "K-Nearest Neighbor (K-NN) based Missing Data Imputation," Int. Conf. Sci. Inf. Technol. K-Nearest, pp. 1–6, 2019.
[13] C. M. Yesilkanat, "Chaos , Solitons and Fractals Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm," Chaos, Solitons and Fractals, vol. 140, 2020, doi: 10.1016/j.chaos.2020.110210.
[14] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," Expert Syst. Appl., vol. 134, pp. 93–101, 2019, doi: 10.1016/j.eswa.2019.05.028.
[15] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression , Random Forest and KNN Models for the Text Classification," Augment. Hum. Res., vol. 5, no. 12, 2020, doi: 10.1007/s41133-020-00032-0.
[16] H. Lu and X. Ma, "Chemosphere Hybrid decision tree-based machine learning models for short-term water quality prediction," Chemosphere, vol. 249, p. 126169, 2020, doi: 10.1016/j.chemosphere.2020.126169.
[17] A. D. Purwanto, K. Wikantika, A. Deliar, and S. Darmawan, "Decision Tree and Random Forest Classification Algorithms for Mangrove Forest Mapping in Sembilang National Park , Indonesia," Remote Sens., vol. 16, no. 151–31, 2023.
[18] W. Baswardono, D. Kurniadi, A. Mulyani, and D. M. Arifin, "Comparative analysis of decision tree algorithms : Random forest and C4 . 5 for airlines customer satisfaction classification," J. Phys. Conf. Ser., 2019, doi: 10.1088/1742-6596/1402/6/066055.
[19] C. A. Ramezan, T. A. Warner, and A. E. Maxwell, "Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification," Remote Sens., vol. 11, no. 185, pp. 1–21, 2019, doi: 10.3390/rs11020185.
[20] J. Sakhnini, H. Karimipour, and A. Dehghantanha, "Smart Grid Cyber Attacks Detection using Supervised Learning and Heuristic Feature Selection," proceeding IEEE SEGE 2019, 2019.
[21] M. Koklu and I. A. Ozkan, "Multiclass classi fi cation of dry beans using computer vision and machine learning techniques," Comput. Electron. Agric., vol. 174, no. May, p. 105507, 2020, doi: 10.1016/j.compag.2020.105507.