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. 1 Iss. 1 February 2024, pp: 47-53
ISSN(P): 3032-7466 | ISSN(E): 3032-7474

Comparation of Decision Tree Algorithm, Naive Bayes, K-Nearest Neighbords on Spotify Music Genre

Desvita Hendri[1*], Diana Nadha[2], Faishal Khairi Basri[3],
Muhammad Farid Wajdi[4], Nurul Nadhirah[5]

[1,2,3]Department of Information System, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[4]Department of MA Social Anthropology, School of Global Studies, University of Sussex, England
[5]Department of Computer Science, Islamic International University, Malaysia

E-Mail: 112150320281@students.uin-suska.ac.id, 212150325125@students.uin-suska.ac.id,
312150311981@students.uin-suska.ac.id, [4]nstnfarid@gmail.com, [5]nrlnadhirah101@gmail.com

Received Dec 23th 2023; Revised Jan 14th 2024; Accepted Mar 4th 2024
Corresponding Author: Desvita Hendri

Abstract

Comparison of Decision Tree, Naive Bayes, K-Nearest Neighbords Algorithm on Spotify Music Genre Decision Tree, Naive Bayes, K-Nearest Neighbords This research aims to compare three algorithms Decision Tree, Naive Bayes and K-Nearest Neighbors (K-NN) in classifying Spotify music genres using dataset from Kaggle. The results show that the Decision Tree algorithm produces an accuracy of 23%, Naive Bayes 17%, and K-Nearest Neighbors 19%. This research provides an overview of Spotify music listeners in choosing music genres. Based on research results, the Decision Tree algorithm has the highest accuracy in classifying Spotify music genres, with the Electric Dance Music (EDM) genre being the most popular among Spotify music fans, followed by rap, pop, r&b, Latin and rock. . Meanwhile, the Naive Bayes and K-Nearest Neighbors algorithms show lower accuracy.

Keyword: Classification, Decisien Tree, K-Nearest Neighbors, Naïve Bayes, Spotify

1. INTRODUCTION

The rapid development of technology and globalization has had both good and bad influences on society, such as the emergence of music streaming service media which makes it easier to listen to music [1]. The shift in the habit of playing songs digitally has resulted in many audio platforms emerging to replace the sale of music albums such as CDs and music discs [2] Since the advent of the internet, people can easily search for all the information they need in cyberspace using an internet connection and even listening to music online through digital internet service providers which can be done anywhere, anytime, and of course using devices that are easy to carry [3]. One provider of music, podcast and video streaming services is Spotify which can be accessed digitally, via iOS, Android and PC platforms. At that time Spotify became the most popular streaming platform in the world with 345 million Spotify users [4].

The Clustering method is a classification method that combines several data into one or smallest part, where this method has its own characteristics and criteria. In this method you can categorize several characteristics and criteria according to the cluster. This method is often used in Data Mining material, which is carried out in classifying certain data [4]. This classification aims to review the most popular songs based on accuracy. Decision Tree, Naive Bayes, and K-Nearest Neighbours algorithms are currently widely used in clustering and classification algorithms. The Decision Tree algorithm is a data mining method used in creating a decision tree from data that has interconnected object characteristics [5]. K-Nearest Neighbords or K-NN is an algorithm that functions to classify data based on learning data (train data sets), which are taken from their K nearest neighbours [6]. Naïve Bayes Classifier is a simple probabilistic classifier that calculates a set of probabilities by summing the frequencies and combinations of values from a given dataset. The algorithm uses Bayes' theorem and assumes all attributes are independent or not interdependent given the value of the class variable [7]. Naïve Bayes is one of the most effective and efficient algorithms for classification in data mining. So this algorithm is very reliable to classify music genres quickly and relevantly[8].

Muslim Hidayat, et al (2023) [9] compared the Naïve Bayes and K-NN algorithms for classification of scholarship recipients at Mi al - Islamiyah Karangsawah, from 186 student data consisting of 150 training

DOI: https://doi.org/10.57152/IJATIS.v1i1.1219

data and 36 testing data obtained classification results with Naïve Bayes and K-Nearest Neighbor obtained 91.67% and 75.00% respectively. Based on the accuracy values obtained from these two algorithms, the accuracy is considered excellent classification, and the Naïve Bayes algorithm is better in classifying scholarship recipients than the K-Nearest Neighbor algorithm [10]. research conducted by Salma Navisa, et al [3]once carried out a comparison of music genre classification algorithms on Spotify. The classification algorithms used for testing are Naive Bayes, K-NN and Random Forest. From the research conducted, the best accuracy results for the Naive Bayes algorithm were obtained with a value of 58.91% [3]. The best performing algorithms are K-NN and Random Forest with a value of 0.528.

In previous research conducted by Asmaul Husna Nasrullah[11], testing the accuracy of the C4.5 algorithm in classifying best-selling products (private data). The accuracy results of the best-selling product classification model using Decision Tree C4.5 obtained from this research are 90% and the AUC value is 0.709, where this value is included in Good Classification [6]. So it can be concluded that the C4.5 Decision Tree Algorithm data mining classification model is accurate in classifying best-selling products.

In research Yousif et al, 2023[12]also discusses the use of data mining to predict student grades, this study found that the performance of the naive bayes algorithm is better than the k-nearest neighbor algorithm based on the results of the analysis, the accuracy of the naïve bayes algorithm reaches 87% and the accuracy of the k-nearest neighbor algorithm reaches 68%.

There are several studies that have been done for music genre classification. Based on the background and several previous studies, a classification of Spotify music genres will be carried out using data sourced from Kaggle. This research aims to compare the decision tree, naive Bayes, k-nearest neighbors algorithms against the Spotify music genre, then we will get the accuracy value of each algorithm. So it can help music listeners on Spotify in choosing music recommendations. This research can be used to improve the accuracy of music recommendations given to users as well as increase the accuracy of music genre search results.

2. MATERIAL AND METHOD

The research phase carried out goes through several processes, the following research flow will be carried out and can be seen in Figure 1.

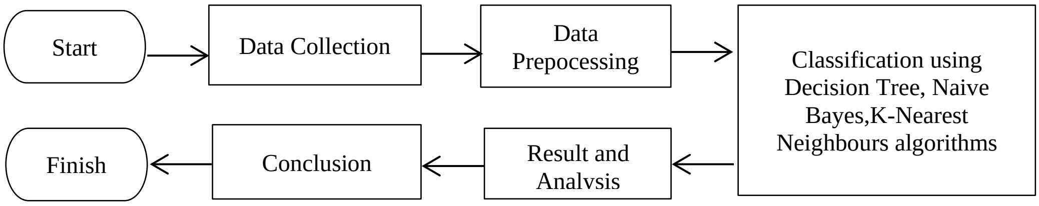
Start
Data Collection
Data Prepocessing
Classification using Decision Tree, Naive Bayes,K-Nearest Neighbours algorithms
Result and Analvsis
Conclusion
Finish

Figure.1 Research Methodology

The research methodology used is shown in Figure 1. The research begins with collecting the necessary data first, then the data preprocessing stage to remove useless attributes and record data that experiences noise, then the classification process is carried out using the Decision Tree algorithm, Naive Bayes, K-Nearest Neighbords. After the classification is done, the classification results are obtained so that conclusions can be drawn from the research carried out.

2.1. Data Mining

Data mining is the process of observing large data and information, which has not previously been detected but can be understood [13]. Data mining methods involve the use of statistical, mathematical, and artificial intelligence techniques to analyze data and identify patterns that can be used to make better decisions, optimize business processes, or reveal new knowledge [14]. Data mining is often used in various fields, including business, health, finance, science and marketing [3].Data mining is a stage of data management that uses statistical techniques, artificial intelligence, machine learning to extract and identify useful information and important indicators from various databases such as Kaggle [15].

2.2. Decision Trees

Decision Tree is a classification algorithm used in machine learning and data mining. This algorithm describes decisions and consequences based on a series of rules and conditions[11]. The decision tree creation process begins by selecting the most interactive features to divide the data into more homogeneous groups. Feature selection based on metrics such as information gain, gain ratio, or Gini index[11]. Decission Tree algorithm is a classification technique that is widely used by researchers. A decision tree or decission tree is the result of the calculation of the decission tree algorithm[16]. The C4.5 algorithm is a commonly used algorithm for forming decision trees. Decision tree is one of the most popular clarification methods because it is easily interpreted by humans. The advantage of the C4.5 algorithm is that it can produce a decision tree that has an acceptable level of accuracy and is efficient in handling attributes that are discrete or numeric types[15].

$$\text{Entropy }(S) = \sum_{j=1}^{n} \text{pi}.\log 2\text{pi} \tag{1}$$

Information:

S : set of cases
n : number of partitions
pi : proportion of Si to S

The stage of generating a decision tree using the Decision Tree algorithm is as equation 2.

$$\text{Gain }(S, A) = \text{Entropi}(S) \sum_{j=1}^{n} - \frac{|\text{Si}|}{|S|} * \text{Entropi Si} \tag{2}$$

Definition:

S : the set of cases
A : Attribute
n : the number of partitions of attribute A
|Si| : denotes the Number of Cases in the i-th Partition
|S| : denotes the number of cases in S

2.3. Naive Bayes

Naïve Bayes is one of the algorithms that can be used in data classification. Naïve Bayes is a data mining algorithm for classification that does not use rules or decision trees. According to Max Brammer, et al (2007) That's why the Naïve Bayes algorithm is included in the group of non-rule based classification algorithms[8].

Naive Bayes is a classification algorithm used in machine learning and data mining [17]. Naïve Bayes uses probability to classify data into appropriate categories [10]. This algorithm learns patterns from the given training data and then uses these patterns to classify new data [18].

$$P(H/X) = \frac{P(X/H)P(H)}{P(X)} \tag{3}$$

Information:

X : Data whose class is not yet known
H : Data hypothesis X is a specific class
P(H/X) : Probability of hypothesis H based on condition x (posteriori prob.) P(H) = Probability of hypothesis H (prior prob.)
P(X/H) : Probability of X based on these conditions P(X) = Probability of X

2.4. K-Nearest Neighbors

K-Nearest Neighborboard is a classification algorithm used in machine learning to classify objects based on a given training data set [19]. This algorithm works by looking for the K nearest neighbors of the object being classified, and then taking majority of categories from these neighbors to classify the object [20].

$$\text{jn} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4}$$

Information:

d : Distance
x : Training Data
y : Test Data
n : Data dimension i = Data variable

2.5. Classification

Classification is the process of grouping or categorizing objects or data into predetermined classes or categories [21]. The process of classification involves grouping objects, data, or phenomena into specific categories or classes based on their properties or characteristics[22]. The purpose of classification is to identify patterns or characteristics that differentiate each class, so that new objects can be classified based on the same patterns [3] According to (Danny Sebastian, 2019) the number of classes in classification, there are two types of classification: binary classification and multi-class classification. Binary classification is when an object is classified into one of two predefined classes. While multi-class classification is when an object is classified to one or more classes.

3. RESULTS AND DISCUSSION

The following are the results and analysis using Decision Tree, Naive Bayes and K-Nearest Neighbors.

3.1. Prepocessing

Before the data is processed using the algorithm, the data obtained from kaggle needs to be processed first. This is done to ensure the data is ready to use and produce optimal analysis results. After preprocessing the data, the data is ready to be used for analysis with the Decision Tree, Naive Bayes, and K-Nearest Neighbours algorithms. Only a few attributes were used in this study to get maximum results.

3.2. Data

This data was taken from the Kaggle.com website, this data totaled 32,833.

Table 1. Spotify Music Data

No.	track_name	track_popularity	track_album_release_date	playlist_name	playlist_genre
1	I Don't Care (with Justin Bieber) - Loud Luxury Remix	66	14/06/2019	Pop Remix	pop
2	Memories - Dillon Francis Remix	67	13/12/2019	Pop Remix	pop
3	All the Time - Don Diablo Remix	70	05/07/2019	Pop Remix	pop
4	Call You Mine - Keanu Silva Remix	60	19/07/2019	Pop Remix	pop
5	Someone You Loved - Future Humans Remix	69	05/03/2019	Pop Remix	pop
6	Beautiful People (feat. Khalid) - Jack Wins Remix	67	11/07/2019	Pop Remix	pop
7	Never Really Over - R3HAB Remix	62	26/07/2019	Pop Remix	pop
…	...	…	…	…	…
32833	Typhoon - Original Mix	27	03/03/2014	ƒ? EDM LOVE 2020	edm

3.3. Result on Decision Tree

The following are the results and discussion of the data processing process using Decision Tree using Google Colab tools.

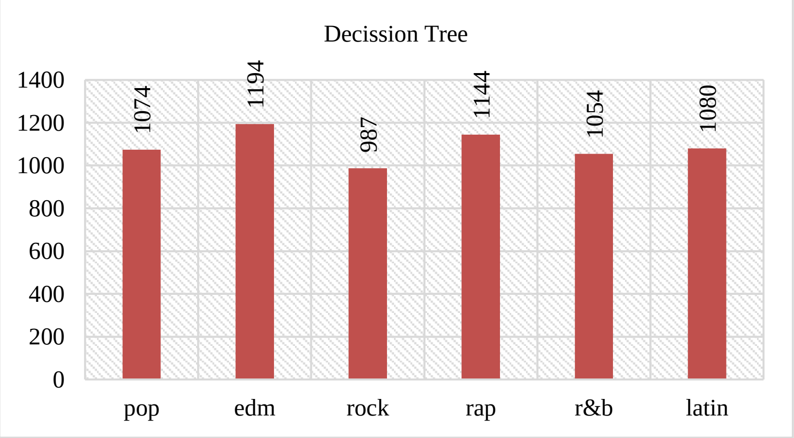

Figure 2. Result on Decision Tree

The accuracy results above show the results of Spotify music data analysis using the decision tree algorithm. From the data processing results, 23% accuracy is obtained. That is, 23% of the music tracks on Spotify can be classified correctly using the decision tree algorithm. Based on the analysis results, the EDM music genre has the highest level of interest among Spotify music fans with a value of 1194. This music genre is followed by rap with 1144, pop with 1074, R&B with 1054, Latin with 1080, and rock with 987.

3.4. Naive Bayes

The following are the results and discussion of the data processing process using Naive Bayes using Google Colab tools.

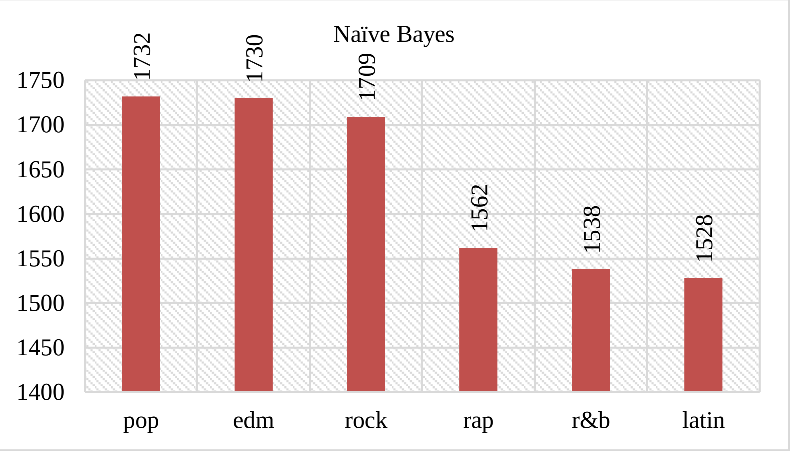

Figure 3. Result on Naïve Bayes

The accuracy result image above shows the results of Spotify music data analysis using the naïve bayes algorithm. From the data processing results, 17.18% accuracy is obtained. This means that 17.18% of the music tracks on Spotify can be classified correctly using the naïve bayes algorithm. Based on the analysis results, the pop music genre has the highest level of interest among Spotify music fans with a value of 1732. This music genre is followed by edm with 1730, rock with 1709, rap with 1562, r&b with 1538, and Latin with 1528.

3.5. Result on K-Nearest Neighbors

The following are the results and discussion of the data processing process using K-Nearest Neighbors using Google Colab tools.

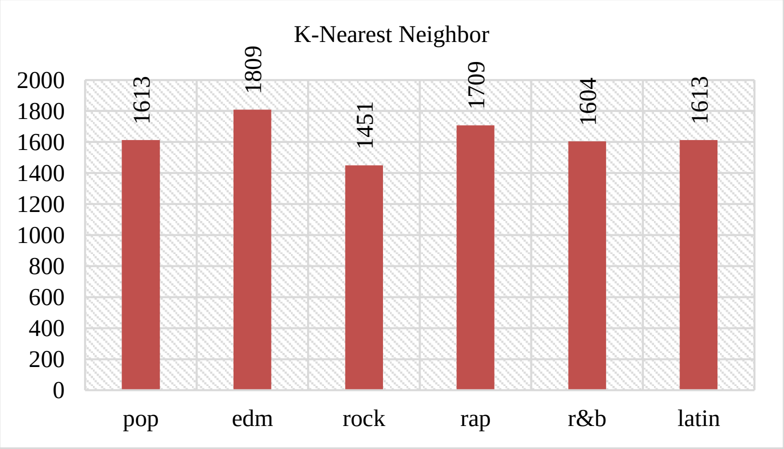

Figure 4. Result on K-Nearest Neighbors

The accuracy result image above shows the results of Spotify music data analysis using the K-NN algorithm. From the data processing results, 19% accuracy is obtained. That is, 19% of the music tracks on Spotify can be classified correctly using the decision tree algorithm. Based on the analysis results, the edm music genre has the highest level of interest among Spotify music fans with a value of 1809. This music genre is followed by rap with 1709, pop with 1613, Latin with 1613, r&b with 1604, and rock with 1451.

3.6. Comparison Result of Decision Tress, Naive Bayes and K-Nearest Neighbors

The classification results of the Decision Tree algorithm using Google Colab tools obtained an accuracy value of 23% with the results of the edm music genre having the highest level of interest as much as 1194. For the Naïve Bayes classification algorithm using the confusion matrix technique to represent prediction class instances with an accuracy value of 17.18% with the results of the pop music genre having the highest level of interest as much as 1732. While the K-Nearest Neighbors classification algorithm uses the confusion matrix technique with an accuracy value of 19% with the result that the edm music genre has the highest level of interest as much as 1809.

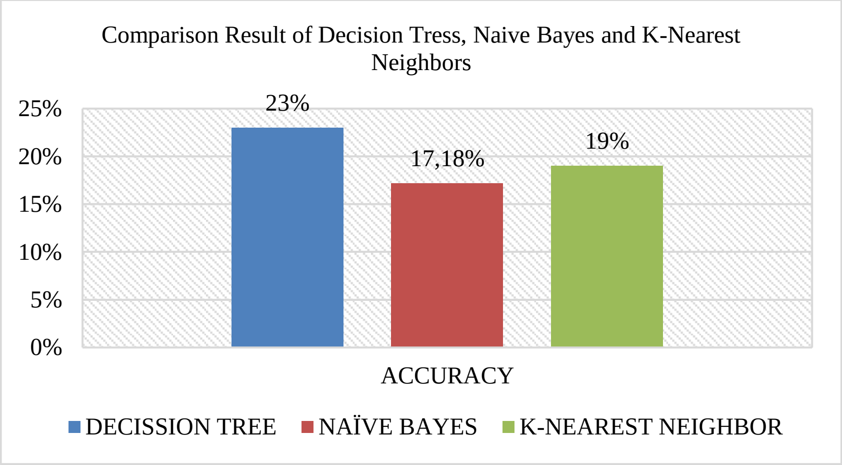

Figure 5. Comparison Result of Decision Tress, Naive Bayes and K-Nearest Neighbors

4. CONCLUSION

This research aims to compare three Decision Tree algorithms, Naive Bayes and K-Nearest Neighbors in classifying Spotify music genres using datasets from Kaggle. The results show that the Decision Tree algorithm produces an accuracy of 23%, Naive Bayes 17%, and K-Nearest Neighbors 19%. This research provides an overview of Spotify music listeners in choosing music genres. Based on the research results, the Decision Tree algorithm has the highest accuracy in classifying Spotify music genres, with the electric dance music (EDM) genre being the most popular among Spotify music fans, followed by rap, pop, r&b, Latin, and rock. Meanwhile, the Naive Bayes and K-Nearest Neighbors algorithms show lower accuracy. This research provides important insight into music genre selection for Spotify users, with the Decision Tree algorithm standing out as the most accurate choice. These results can provide benefits for music listeners in choosing music recommendations on the Spotify platform.

REFERENCES

[1] U. L. Musyarofah, S. N. Alima, and D. S. Y. Kartika, "Klasifikasi Top 50 Spotify Tahun 2010-2019 Menggunakan Metode K-Means Clustering," Pros. Semin. Nas. Teknol. dan Sist. Inf., vol. 2, no. 1, pp. 215–220, 2022, doi: 10.33005/sitasi.v2i1.300.

[2] C. W. Harto, V. C. Mawardi, and N. J. Perdana, "Website Rekomendasi Dan Klasifikasi Lagu Menggunakan Metode Weighted K-Nearest Neighbor," J. Ilmu Komput. dan Sist. Inf., vol. 11, no. 1, 2023, doi: 10.24912/jiksi.v11i1.24074.

[3] S. Navisa, Luqman Hakim, and Aulia Nabilah, "Komparasi Algoritma Klasifikasi Genre Musik pada Spotify Menggunakan CRISP-DM," J. Sist. Cerdas, vol. 4, no. 2, pp. 114–125, 2021, doi: 10.37396/jsc.v4i2.162.

[4] B. A. Firdaus, D. E. Ratnawati, and B. T. Hanggara, "Klusterisasi Popularitas Artist pada Playlist Today's Top Hits Menggunakan Metode K-Means dengan Integrasi Spotify Web API dan Teknologi Amazon SageMaker," vol. 5, no. 1, pp. 2548–964, 2021, [Online]. Available: http://j-ptiik.ub.ac.id

[5] L. Nurhalimah, T. I. Hermanto, and I. Kaniawulan, "Analisis Prediksi Mood Genre Musik Pop Menggunakan Algoritma K-Means dan C4.5," JURIKOM (Jurnal Ris. Komputer), vol. 9, no. 4, p. 1006, 2022, doi: 10.30865/jurikom.v9i4.4597.

[6] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," Indones. J. Data Sci., vol. 1, no. 2, pp. 29–33, 2020, doi: 10.33096/ijodas.v1i2.11.

[7] S. Sahar, "Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Clasiffier Pada Dataset Penyakit Jantung," Indones. J. Data Sci., vol. 1, no. 3, pp. 79–86, 2020, doi: 10.33096/ijodas.v1i3.20.

[8] Thoriq Nurchaidir, Widodo, and Bambang Prasetya Adhi, "Klasifikasi Genre Musik Menggunakan Algoritma Naïve Bayes Classifier Untuk Layanan Streaming Youtube," PINTER J. Pendidik. Tek. Inform. dan Komput., vol. 7, no. 1, pp. 1–6, 2023, doi: 10.21009/pinter.7.1.1.

[9] M. Hidayat, A. N. Fuadi, D. P. Utomo, and ..., "Studi Komparasi Algoritma Naïve Bayes Dan K-Nn Untuk Klasifikasi Penerimaan Beasiswa Di Mi Al–Islamiyah Karangsawah," … J. Ilm. Tek. …, vol. 2, no. 4, pp. 172–180, 2023, [Online]. Available:

https://journal.literasisains.id/index.php/storage/article/view/2865%0Ahttps://journal.literasisains.id/index.php/storage/article/download/2865/1339

[10] P. D. Rinanda, B. Delvika, S. Nurhidayarnis, N. Abror, and A. Hidayat, “Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes pada Ibu Hamil,” MALCOM Indones. J. Mach. Learn. Comput. Sci., vol. 2, no. 2, pp. 68–75, 2022, doi: 10.57152/malcom.v2i2.432.

[11] A. H. Nasrullah, “Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik,” J. Pilar Nusa Mandiri, vol. 7, no. 2, p. 217, 2021.

[12] P. Journal, “Performance comparison between Naïve Bayes and k-Nearest Neighbor in predicting student grades,” Humanit. Nat. Sci. J., vol. 4, no. 7, 2023, doi: 10.53796/hnsj476.

[13] S. K. P. Loka and A. Marsal, “Perbandingan Algoritma K-Nearest Neighbor dan Naïve Bayes Classifier untuk Klasifikasi Status Gizi Pada Balita,” MALCOM Indones. J. Mach. Learn. Comput. Sci., vol. 3, no. 1, pp. 8–14, 2023, doi: 10.57152/malcom.v3i1.474.

[14] A. C. Sitepu, W. Wanayumini, and Z. Situmorang, “Analisis Kinerja Support Vector Machine dalam Mengidentifikasi Komentar Perundungan pada Jejaring Sosial,” J. Media Inform. Budidarma, vol. 5, no. 2, p. 475, 2021, doi: 10.30865/mib.v5i2.2923.

[15] J. J. Pangaribuan, C. Tedja, and S. Wibowo, “Perbandingan Metode Algoritma C4.5 Dan Extreme Learning Machine Untuk Mendiagnosis Penyakit Jantung Koroner,” J. Informatics Eng. Res. Technol., vol. 1, no. 1, pp. 9–15, 2019.

[16] F. M. Hana, “Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5,” J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan), vol. 4, no. 1, pp. 32–39, 2020, doi: 10.47970/siskom-kb.v4i1.173.

[17] A. D. Cahyo, “Metode Naive Bayes Untuk Klasifikasi Masa Studi Sarjana,” J. Teknol. Pint., vol. 3, no. 4, 2023, [Online]. Available: http://teknologipintar.org/index.php/teknologipintar/article/view/385%0Ahttp://teknologipintar.org/index.php/teknologipintar/article/download/385/370

[18] N. Nurhachita and E. S. Negara, “A Comparison Between Naïve Bayes and The K-Means Clustering Algorithm for The Application of Data Mining on The Admission of New Students,” J. Intelekt. Keislaman, Sos. dan Sains, vol. 9, no. 1, pp. 51–62, 2020, doi: 10.19109/intelektualita.v9i1.5574.

[19] D. Sebastian, “Implementasi Algoritma K-Nearest Neighbor untuk Melakukan Klasifikasi Produk dari beberapa E-marketplace,” J. Tek. Inform. dan Sist. Inf., vol. 5, no. 1, pp. 51–61, 2019, doi: 10.28932/jutisi.v5i1.1581.

[20] D. Cahyanti, A. Rahmayani, and S. A. Husniar, “Analisis performa metode K-NN pada Dataset pasien pengidap Kanker Payudara,” Indones. J. Data Sci., vol. 1, no. 2, pp. 39–43, 2020, doi: 10.33096/ijodas.v1i2.13.

[21] M. Al Khadafi, Kurnia Paranitha Kartika, and Filda Febrinita, “Penerapan Metode Naïve Bayes Classifier Dan Lexicon Based Untuk Analisis Sentimen Cyberbullying Pada Bpjs,” JATI (Jurnal Mhs. Tek. Inform., vol. 6, no. 2, pp. 725–733, 2022, doi: 10.36040/jati.v6i2.5633.

[22] K. N. N. Dan and A. Genetika, “Sistem rekomendasi musik spotify menggunakan K-NN dan algoritma genetika,” vol. 7, no. 4, pp. 2585–2591, 2023.