Journal of Student Research Exploration

https://shmpublisher.com/index.php/josre
p-ISSN 2964-1691 | e-ISSN 2964-8246

Application of k-nearest neighbor algorithm in classification of engine performance in car companies using Rapidminer

Irendra Lintang Keksi[1], Apri, Dwi Lestari[2], Budi Prasetiyo[3]
[1,2,3]Department of Computer Science, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received April 18, 2024
Revised July 4, 2024
Accepted July 5, 2024

Keywords:

K-nearest neighbor
Classification
Machine performance
Rapidminer
Automotive industry

ABSTRACT

Implementation of the k-Nearest Neighbor (k-NN) algorithm in the classification of CAR Car company engine performance using RapidMiner software. The company's engine performance is a very important aspect in the automotive industry that greatly affects operational efficiency and customer satisfaction. As an effort to monitor and improve engine performance, classification is an important key to identify machines that are feasible and require repair. The dataset used is a generated dataset from the AI Chat GPT bot whose prompts have been adapted to the research needs. The k-NN algorithm was chosen due to its ability to produce accurate predictions. The k-NN classification method utilizes training and testing data and calculates the distance between the data to determine the appropriate class. The results of this study show excellent performance in terms of accuracy, precision, and recall. The highest accuracy is 90.62% at the value of k = 2. The highest precision and recall are 100% and 93.75% at the values of k = 2, k = 4, and k = 7.

1. Introduction

In today's increasingly advanced digital era, increasing operational efficiency and effectiveness is the main goal for various industries. The automotive industry is no

[1] ***Corresponding Author:***
Irendra Lintang Keksi,
Department of Information System,
University of Semarang, Indonesia
Email: irendralintang@students.unnes.ac.id
DOI: https://doi.org/10.52465/josre.v2i2.345

exception and plays an important role in supporting the global economy [1]. In the automotive industry, the engine is the heart of every vehicle. Optimal engine performance not only affects the overall quality of the vehicle, but can also reduce operational and maintenance costs, and increase customer satisfaction [2].

A deep understanding of engine performance can help in various aspects. First, it can help in machine maintenance and repair. By knowing how the machine performs under various conditions, we can better plan and execute routine maintenance, as well as deal with problems that may arise [3]. This in turn can reduce downtime and costs associated with machine repairs. Secondly, an understanding of machine performance can help in the design and development of new machines. By knowing what works and what doesn't, engineers and designers can make improvements to existing machine designs or develop new designs that are more efficient and effective [4]. Third, an understanding of machine performance can also help in strategic decision-making. For example, if a type of machine consistently shows poor performance, it may be better to stop its production and focus on another type of machine that shows better performance [3].

Therefore, monitoring and improving the performance of machines is of utmost importance. One way to achieve this is through machine performance classification. With machine performance classification, we can identify which machines perform well, are feasible and safe to distribute as well as which ones may require further repair or maintenance.

Data Mining is one of the objectives to determine certain patterns that can predict in making a decision in the future [5]. One of the processes in using the technique of coming mining is using the algorithm process with the aim of classification. Data classification is the process of finding a model or function that can explain and distinguish between data classes and concepts [6]. One of the processes in using data mining techniques is using certain algorithms for classification. This research will apply classification using the k-Nearest Neighbor (k-NN) algorithm. k-NN is a classification technique that operates by considering the distance between new data and a number of data or the closest neighbors [7].

The k-NN algorithm has been used for various types of classification. Previous research for motor image classification based on EEG signals showed that the k-NN algorithm is superior to Support Vector Machine (SVM) [8]. In addition, the use of k-NN algorithm for human activity classification also shows high performance results with the average results of precision, recall, F1-score, AUC, and area under the ROC curve are 90.96%, 90.46%, 90.37%, and 96.5%, respectively, while the area under the ROC curve is 100% [9].

The k-Nearest Neighbor (k-NN) algorithm is one of the most popular methods due to its ease of implementation and ability to produce accurate predictions [8]–[10].

This research aims to apply the k-NN algorithm in the classification of engine performance at CAR Car Company using Rapid Miner software.

2. Method

Stages of the Method

This research uses the k-NN algorithm to classify engine performance. In the data generation stage, machine performance data is collected through a generated dataset from the AI Chat GPT bot whose prompts have been adapted to the research needs. Next, the data undergoes pre-processing to remove duplicate data and missing values. The k-NN algorithm is then used to predict the class of a sample with an unspecified class based on the class of its neighboring samples [11]. The classification results are analyzed to draw conclusions about the performance of the machine. The stages of the research method can be explained in the figure, as in the following figure:

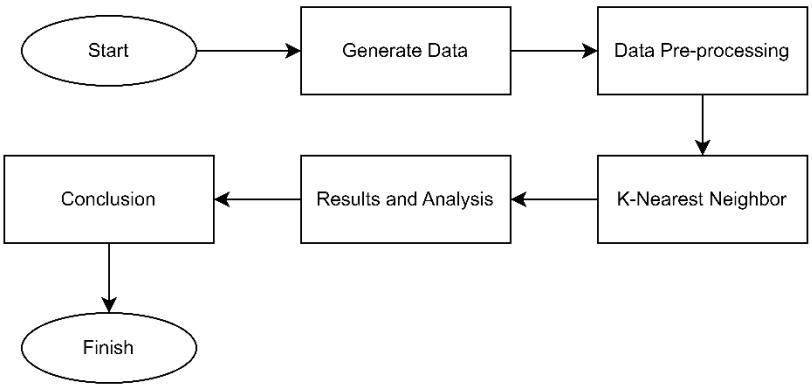

Figure 1. Stages of the research method

Data Source

Data is a series of events obtained from measurements. The right decision is based on a rational analysis of the available information, and also underlines that data reflects a reality that describes various events or entities that have not been fully explored. This data will then be processed to produce valuable information for better decision-making [12]. The dataset used in this research is a dummy dataset. The dummy dataset was created using the AI Chat GPT bot with prompts that have been adjusted to the theme based on research needs. The prompts made are as follows:

"Bayangkan anda bekerja di sebuah perusahaan mesin mobil lalu anda ingin melihat dan menilai kinerja mesin untuk menentukan apakah mesin tersebut layak digunakan, terdapat 3 kategori mesin berdasarkan kriteria penilaian, pertama "mesin layak" untuk yang kinerjanya bagus dan tidak ada cacat, kedua "mesin kurang layak" untuk yang kinerjanya sedang dan terdapat beberapa

kekurangan atau cacat, dan ketiga "mesin tidak layak" untuk mesin yang kinerjanya rendah dan terdapat banyak kecacatan di dalamnya. Berdasarkan kategori tersebut, buatlah aspek aspek yang mendukung penilaian kinerja mesin pada perusahaan tersebut serta buatlah contoh dataset yang memuat aspek-aspek tersebut. Generate dataset sebanyak sekitar 100-200 rows.″

Data Corpus

The dummy dataset used consists of 10 columns, namely "ID Mesin", "Nama Mesin", "Konsumsi BBM (km/l)", "Daya Mesin (kW)", "Torsi Mesin (Nm)", "Usia Mesin (tahun)", "Perbaikan (jumlah)", "Keausan Komponen (%)", "Emisi Gas Buang (ppm)", and "Kategori Mesin". Table 1 shows the description of each variable.

Table 1. Data Description

Variable	Deskription
ID Mesin	Unique identification for each machine in alphanumeric format with a length of 5 characters
Nama Mesin	Car type and series information
Konsumsi BBM (km/l)	Engine efficiency in using fuel
Daya Mesin (kW)	The power or capacity of the machine to do the job
Torsi Mesin (Nm)	The ability of the engine to produce rotating power
Usia Mesin (tahun)	Length of time the machine has been operating
Perbaikan (jumlah)	Number of repairs or maintenance performed on the machine since the beginning of use
Keausan Komponen (%)	Wear rate of engine components in percentage
Emisi Gas Buang (ppm)	The amount of exhaust emissions produced by the engine
Kategori Mesin	Classification of machine performance based on the assessment criteria of "Mesin Layak", "Mesin Kurang Layak", and "Mesin Tidak Layak"

The variable "Kategori Mesin" is the target variable to be predicted. The data in this attribute will be used as the basis for measuring and predicting whether an engine is worth using, has some deficiencies, or is not even worth using. The evaluation of this performance category will be the basis for decision-making regarding the use of engines in vehicles.

Classification

Classification is the process of placing features into appropriate classes. Training feature vectors whose classes are known are used to design separators. This type of pattern recognition is known as supervised [13]. In particular, the classification process is divided into different categories, which are referred to as decision-based classifiers. A classifier is constructed based on a training dataset whose classes are predefined. In this research, the classification method used is the k-NN algorithm.

K-Nearest Neighbor

The k-NN algorithm is a well-known method for non-numerical data mining, which is then applied for classification or regression. To be able to perform classification with the algorithm, a dataset that includes training data and testing data is required [14]. This algorithm operates on the principle of minimum distance between test data and training data to determine its k-NN. After the k-NN is collected, the majority of the k-NN is taken to be the prediction of the test sample [15]. The distance between neighbors is usually measured using the Euclidean distance.

The value of k in Nearest Neighbor refers to the k-data that is closest to the test data. For example, if K is 3, then the three nearest neighbors of the training data will be selected. Similarly, if k is n, then n nearest neighbors of the training data will be selected [16]. One of the objectives in the nearest neighbor method is to determine the optimal k value so that later the most appropriate decision can be drawn from the results of classifying using the k-NN algorithm. The steps in calculating the classification method with the k-NN algorithm are as follows:

1) Set parameters or k values.
2) Calculate the distance between training data and testing data.
   The distance measurement that is often used in the k-NN algorithm is the Euclidean distance method. The formula is as follows:

$$d_i = \sqrt{\sum_{i=1}^{n}(X_{2i} - X_{1i})^2} \quad (1)$$

   Description:
   d = Distance
   n = Data dimension
   i = Variable data
   x_{2i} = Training data
   x_{1i}= Testing data

3) Set the distance that has been formed.
4) Summarize the smallest distance to the order of k.
5) Connecting the appropriate class.
6) Count the number of nearest neighbor classes and set that class as the data class to be evaluated.

Confusion Matrix

Confusion matrix is an evaluation matrix for classification problems in machine learning, where the result can be two or more classes that can be utilized in performance measurement of classification models. Confusion matrix is also often

referred to as error matrix. Its main purpose is to compare the test result data of the classification system with the actual classification target [17].

Table 2. Confusion matrix

Classification	Class Prediction	
	True	False
True	True Positif (TP)	False Negatif (FN)
False	True Negatif (TN)	False Positif (FP)

The calculation of the performance measurement of the classification model with the confusion matrix method is as follows.

a. Accuracy: The accuracy value in measuring the performance of the classification model using the confusion matrix method is calculated by comparing the amount of data that has been classified as correct with the total amount of data. The following is the formula for calculating the accuracy value:

$$Accuracy = \frac{TP + TN}{p + n} \times 100\% \quad (2)$$

b. Precission: Precision in the confusion matrix is a measure that shows how accurate the model is in making positive predictions. Precision measures the ratio of correct positive predictions to positive predictions that are actually negative. In other words, precision measures the percentage of positive predictions that are actually positive. The precision value can be calculated with the following formula:

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (3)$$

c. Recall: Recall in the confusion matrix, also known as sensitivity or True Positive Rate (TPR) indicates a measurement of how well the model identifies all true positive results from the data. Of the actual number of data that are positive, how much of the predicted number of data are positive [18]. The following is the formula for calculating the recall value:

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (4)$$

Description:

$p + n$ = Amount of data

TP = True positif value

TN = True negatif value

FP = False positif value

FN = False negatif value

3. Results and Discussion

Data Preparation

Before applying the k-NN algorithm to the dataset, the first thing to do is to prepare the data. Any missing values contained in the dataset are filled with the appropriate values. Data cleansing is the process of cleaning the data that will be used for data deletion by removing missing values, data duplication, and checking for misalignment in the data and correcting errors in the data [19].

			Min	Max	Average
Perbaikan (jumlah)	Integer	1	1	17	5.836
Keausan Komponen (%)	Integer	1	5	70	30.503

Figure 2. Missing values

Before the dataset was cleansed, there were several missing values. In the "Perbaikan (jumlah)" attribute column there is one missing value, as well as in the "Keausan Komponen (%)" attribute column.

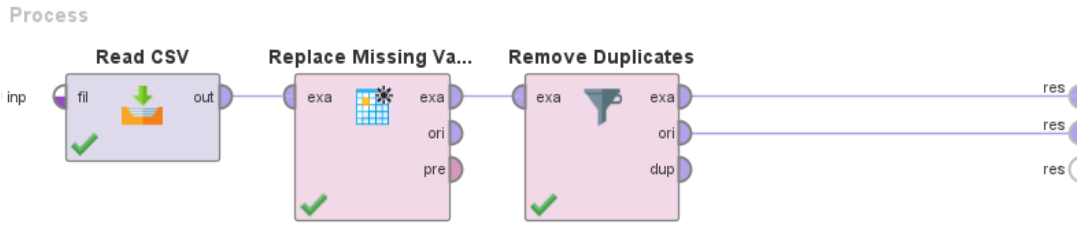

Figure 3. Operators in data cleansing

The operators used in data cleansing are Replace Missing Values and Remove Duplicates. The Replace Missing Values operator in RapidMiner serves to replace missing or absent values in the dataset. This replacement value can be the minimum, maximum, or average value of the attribute, or even a zero value. In this dataset, the missing values are replaced with the average value of each attribute.

Whereas the Remove Duplicates operator is used to compare all the instances against each other based on the specified attributes and remove the duplicate instances. Two examples are considered duplicates if the selected attribute has the same value in it. The following is the result of the dataset that has been applied data cleansing in the statistics view, it can be seen that the missing values in the attribute columns "Perbaikan (jumlah)" and "Keausan Komponen (%)" have changed to blank.

Name	Type	Missing
Perbaikan (jumlah)	Integer	0
Keausan Komponen (%)	Integer	0

Figure 4. Data cleansing result

Implementation of k-NN Algorithm

Data that has been cleansed is ready to be implemented by the k-NN algorithm. After data preparation, data processing will then be carried out using the k-NN algorithm stages [20]. The application of the k-NN algorithm still uses RapidMiner tools with the following operators.

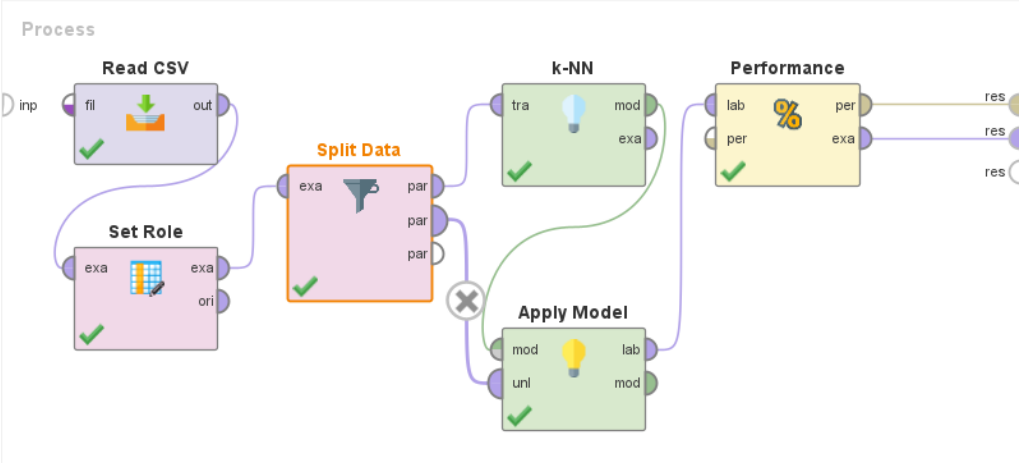

Figure 5. Operators in k-NN algorithm implementation

A total of 160 data will be applied to the k-NN algorithm, with 80% training data and 20% testing data [21]. This data division is done using the Split Data operator. The attribute classified as a label in this study is the "Kategori Mesin" column with three criteria namely; "Mesin Layak", "Mesin Kurang Layak", and "Mesin Tidak Layak".

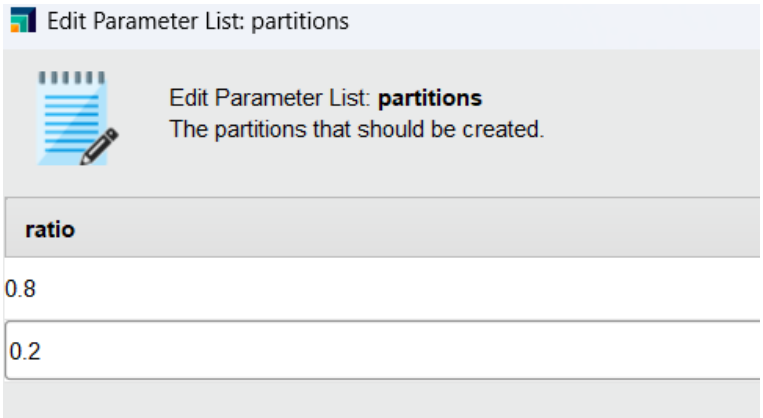

Figure 6. Data splitting

The next step is to apply the k-NN algorithm using the k-NN operator. The determination of the k value used in this study is the value of k = 2, 4, and 7. From the choice of the k value, it is determined which value has the most optimal accuracy value. The following are the experimental results of applying the k-NN algorithm with the values of k = 2, 4, and 7.

accuracy: 90.62%

	true Mesin Layak	true Mesin Kurang Layak	true Mesin Tidak Layak	class precision
pred. Mesin Layak	15	1	0	93.75%
pred. Mesin Kurang Layak	1	7	1	77.78%
pred. Mesin Tidak Layak	0	0	7	100.00%
class recall	93.75%	87.50%	87.50%	

Figure 7. K-NN result with k value = 2

Figure 7 shows the results of the performance of the application of the k-NN algorithm at the value of k = 2 with the performance calculated is precision, recall, and accuracy.

accuracy: 87.50%

	true Mesin Layak	true Mesin Kurang Layak	true Mesin Tidak Layak	class precision
pred. Mesin Layak	15	1	0	93.75%
pred. Mesin Kurang Layak	1	7	2	70.00%
pred. Mesin Tidak Layak	0	0	6	100.00%
class recall	93.75%	87.50%	75.00%	

Figure 8. K-NN result with k value = 4

Figure 8 shows the results of the performance of the application of the k-NN algorithm at a value of k = 4 with the performance calculated is precision, recall, and accuracy.

accuracy: 84.38%

	true Mesin Layak	true Mesin Kurang Layak	true Mesin Tidak Layak	class precision
pred. Mesin Layak	15	1	0	93.75%
pred. Mesin Kurang Layak	1	7	3	63.64%
pred. Mesin Tidak Layak	0	0	5	100.00%
class recall	93.75%	87.50%	62.50%	

Figure 9. K-NN result with k value = 7

Figure 9 shows the results of the performance of the application of the k-NN algorithm at a value of k = 7 with the performance calculated is precision, recall, and accuracy.

The experimental results of applying the k-NN algorithm with three different k values, namely 2, 4, and 7 show the highest accuracy of 90.62% when the value of k = 2. The highest precision is in the prediction of the "Mesin Tidak Layak" category which is 100%, this result is the same for all three values of k. This means that all data of the "Mesin Tidak Layak" category can be predicted correctly (True Positive). The highest Recall is in the "Mesin Layak" category prediction which is 93.75%, this result is the same for all three values of k. This result shows the good performance of the k-NN algorithm for the classification of engine eligibility.

4. Conclusion

Based on the research that has been done, it can be concluded that the k-Nearest Neighbor (k-NN) algorithm is suitable for engine performance classification in car companies. The highest accuracy of 90.62% is achieved at a value of k = 2. The highest Precision is 100% in the "Mesin Tidak Layak" category for values of k = 2, k = 4, and k = 7. The highest Recall is 93.75% in the "Mesin Layak" category for values of k = 2, k = 4, and k = 7. Although this research shows excellent performance, research with similar objects can be developed through the application of other machine learning algorithms to obtain superior classification results and on larger datasets.

REFERENCES

[1] Z. He, L. Sun, Y. Hijioka, K. Nakajima, and M. Fujii, "Systematic review of circular economy strategy outcomes in the automobile industry," *Resour. Conserv. Recycl.*, vol. 198, p. 107203, Nov. 2023, doi: 10.1016/j.resconrec.2023.107203.

[2] D. Dhabliya, A. H. Alkkhayat, J. Sivakumar, R. Bhokde, and M. B, "Design and Analysis of Four-Wheeler Chassis for Improved Performance," in *2023 4th International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, Dec. 2023, pp. 1–8. doi: 10.1109/ICCAKM58659.2023.10449564.

[3] A. P. Lubis, "Analisis Keandalan dan Pemeliharaan Mesin Industri," 2024. [Online]. Available: https://coursework.uma.ac.id/index.php/mesin/article/view/768

[4] K. Kudelina, B. Asad, T. Vaimann, A. Rassõlkin, A. Kallaste, and H. Van Khang, "Methods of Condition Monitoring and Fault Detection for Electrical Machines," *Energies*, vol. 14, no. 22, p. 7459, Nov. 2021, doi: 10.3390/en14227459.

[5] Q. A. A' yuniyah and M. Reza, "Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Di Sma Negeri 15 Pekanbaru," *Indones. J. Inform. Res. Softw. Eng.*, vol. 3, no. 1, pp. 39–45, 2023, doi: 10.57152/ijirse.v3i1.484.

[6] N. B. Putri and A. W. Wijayanto, "Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing," *Komputika J. Sist. Komput.*, vol. 11, no. 1, pp. 59–66, 2022, doi: 10.34010/komputika.v11i1.4350.

[7] E. W. Jumadi, "Penggunaan K-NN (K-Nearest Neighbor) Untuk Klasifikasi Teks Berita yang Tak-Terkelompokkan pada Saat Pengklasteran Oleh STC (Suffix Tree Clustering)," *Istek*, vol. 9, no. 1, pp. 50–81, 2015.

[8] N. E. Md Isa, A. Amir, M. Z. Ilyas, and M. S. Razalli, "The Performance Analysis of K-Nearest Neighbors (K-NN) Algorithm for Motor Imagery Classification Based on EEG Signal," *MATEC Web Conf.*, vol. 140, p. 01024, Dec. 2017, doi: 10.1051/matecconf/201714001024.

[9] S. Mohsen, A. Elkaseer, and S. G. Scholz, "Human Activity Recognition Using K-Nearest Neighbor Machine Learning Algorithm," 2022, pp. 304–313. doi: 10.1007/978-981-16-6128-0_29.
[10] K. Samruddhi and R. Ashok Kumar, "Used Car Price Prediction using K-Nearest Neighbor Based Model," *Int. J. Innov. Res. Appl. Sci. Eng.*, vol. 4, no. 2, pp. 629–632, Aug. 2020, doi: 10.29027/IJIRASE.v4.i2.2020.629-632.
[11] M. R. Alghifari and A. P. Wibowo, "K-NN 14," *J. Teknol. Manaj. Inform.*, vol. 5, no. 1, 2019.
[12] A. Oluwaseun and M. S. Chaubey, "Data Mining Classification Techniques on the analysis of student performance," *Glob. Sci. J.*, vol. 7, no. April, pp. 79–95, 2019, doi: 10.11216/gsj.2019.04.19671.
[13] A. P. Wibawa, M. G. A. Purnama, M. F. Akbar, and F. A. Dwiyanto, "Metode-metode Klasifikasi," *Pros. Semin. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 1, p. 134, 2018.
[14] N. Nuraeni, "Klasifikasi Data Mining untuk Prediksi Potensi Nasabah dalam Membuat Deposito Berjangka Data Mining Classification for Predicting Customer Potential in Making Term Deposits," *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 3, no. 01, pp. 65–75, 2021.
[15] 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.
[16] H. P. Herlambang, F. Saputra, M. H. Prasetiyo, D. Puspitasari, and D. Nurlaela, "Perbandingan Klasifikasi Tingkat Penjualan Buah di Supermarket dengan Pendekatan Algoritma Decision Tree, Naive Bayes dan K-Nearest Neighbor," *J. Insa. - J. Inf. Syst. Manag. Innov.*, vol. 3, no. 1, pp. 21–28, 2023, doi: 10.31294/jinsan.v3i1.2097.
[17] S. Prayogo, A. A. Chamid, and A. C. Murti, "Perancangan Sistem Klasifikasi Jenis Bunga Mawar Menggunakan Metode K-Nearest Neighbor (K-NN)," *Indones. J. Technol. Informatics Sci.*, vol. 3, no. 2, pp. 52–56, 2022, doi: 10.24176/ijtis.v3i2.7881.
[18] Nikmatun, I. Alvi, Waspada, and Indra, "Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," *J. SIMETRIS*, vol. 10, no. 2, pp. 421–432, 2019.
[19] I. Nawangsih and J. Sahar, "Penerapan Data Mining Untuk Analisa Kualitas Produk Welding Dengan Algoritma Naïve Bayes Dan C4.5 Pada Pt. Karya Bahana Unigam," *Sigma J. Teknol. Pelita Bangsa*, vol. 13, no. 1, pp. 21–26, 2022.
[20] H. Paul, A. Sartika Wiguna, and H. Santoso, "Penerapan Algoritma Support Vector Machine Dan Naive Bayes Untuk Klasifikasi Jenis Mobil Terlaris Berdasarkan Produksi Di Indonesia," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 39–44, 2023, doi: 10.36040/jati.v7i1.5555.
[21] H. Mubarok, S. Murni, and M. M. Santoni, "Penerapan Algoritma K-Nearest Neighbor untuk Klasifikasi Tingkat Kematangan Buah Tomat Berdasarkan Fitur Warna," *Semin. Nas. Mhs. Ilmu Komput. dan Apl. Jakarta-Indonesia*, no. April, pp. 773–782, 2021.