Information Technology and Systems
e-ISSN: 3031-1187
Homepage: https://sanscientific.com/journal/index.php/its
1(1) 45-54 (2023)
https://doi.org/10.58777/its.v1i1.123

Research Article

Implementation of the K-Nearest Neighbor Algorithm to Predict Air Pollution

Claudyana Gabrillia Evitania
Faculty of Engineering and Informatics, Bina Sarana Informatics University, Depok

Received: September 2023; Accepted: November 2023

Abstract
Air pollution is a serious issue that impacts air quality and human health. In this study, the K-Nearest Neighbor (KNN) algorithm is applied using Rapidminer software to predict air pollution levels. The research aims to predict air pollution levels based on various air quality parameters such as particulates, PM10, PM2.5, CO, NO2, SO2, and O3. By implementing the K-Nearest Neighbor algorithm in Rapidminer, the predicted values for air pollution data resulted in an accuracy of 93.94%. This study concludes that employing the K-Nearest Neighbor algorithm using Rapidminer software can be an effective method for predicting air pollution levels. With a strong accuracy rate of 93.94%, this can have a positive impact on both human health and the environment. The predictive model developed can aid decision-making and enhance awareness among the public regarding the importance of maintaining air quality management.

Keywords: Air Pollution, K-Nearest Neighbor (KNN) Algorithm, RapidMiner, Predicts

Abstrak
Polusi udara merupakan masalah serius yang memengaruhi kualitas udara dan kesehatan manusia. Dalam penelitian ini, K-Nearest Neighbor (KNN) diterapkan menggunakan perangkat lunak Rapidminer untuk memprediksi tingkat polusi udara. Penelitian ini bertujuan untuk mendapatkan hasil prediksi pada data tingkat polusi udara menggunakan metode klasifikasi tetangga terdekat, di mana data yang digunakan meliputi parameter kualitas udara seperti partikulat, PM10, PM2.5, CO, NO2, SO2, dan O3. Dengan menerapkan algoritma tetangga terdekat di perangkat lunak Rapidminer, nilai yang berhasil diprediksi dari data polusi udara menghasilkan akurasi sebesar 93,94%. Penelitian ini menyimpulkan bahwa menerapkan algoritma tetangga terdekat menggunakan perangkat lunak Rapidminer dapat digunakan sebagai metode yang efektif untuk memprediksi tingkat polusi udara. Dengan hasil akurasi yang baik, yaitu 93,94%, hal ini dapat berdampak positif bagi manusia dan lingkungan, dan model prediksi yang dikembangkan dapat membantu dalam pengambilan keputusan dan meningkatkan kesadaran di kalangan masyarakat tentang pentingnya menjaga manajemen kualitas udara.

Kata Kunci: Polusi Udara, K-Nearest Neighbor (KNN), RapidMiner, Prediksi

How to cite: Evitania, C. G., (2023). Implementation of the K-Nearest Neighbor Algorithm to Predict Air Pollution, *Information Technology and Systems (ITS) 1(1)*, 45-54

Corresponding author: Claudyana Gabrillia Evitania (claudyanahaumahu9i@gmail.com)

1. Introduction

In this modern era where technology is increasingly advanced, many things are happening because of the times. From human lifestyle to their way of thinking, the condition of the environment around them, and, of course, the air space. We know that population growth is increasing rapidly and, as well as economic needs, human activity is increasing. Starting from morning until night we can see so many community activities. The issue of air pollution is a global concern that significantly impacts human health, the environment, and the climate. Monitoring air pollution levels is crucial for identifying potential risks and taking preventive actions (Wicaksono et al., 2023). Air pollution levels can vary significantly in terms of both time and location. Hence, predictive methods are needed to model this variability and provide more accurate and relevant information. Air pollution data is often complex, consisting of many interconnected parameters. The use of appropriate predictive algorithms can help address this complexity and generate models that offer better predictions (Larasati, 2023).

The K-Nearest Neighbor (KNN) algorithm is one flexible and easily implementable approach for predictive modeling (Setianingsih et al., 2023). KNN excels in handling complex and nonlinear data, providing good results for classification problems (Zulfiansyah et al., 2023). To enhance the effectiveness of air pollution monitoring, developing accurate predictive models is essential. Implementing KNN could be a promising solution to tackle these challenges, particularly in the context of air pollution prediction. Air pollution conditions can change rapidly, making real-time monitoring increasingly crucial. By implementing KNN, known for its ability to swiftly respond to changes, the accuracy of predictions in dynamic environments can be improved. This research aims to make a positive contribution to air pollution management by providing a reliable predictive method. This can assist local authorities, researchers, and the public in taking more effective preventive and mitigation measures. Understanding this background, this research is expected to offer innovative and practical solutions to the challenges of air pollution modeling, specifically through the application of the K-Nearest Neighbor algorithm.

With very dense activity in society, this causes changes in environmental conditions, especially the air. Meanwhile, air is a very important factor in living creatures' lives. This condition causes clean air to be needed by living creatures. However, this is very difficult for people living in urban areas, especially Jakarta, to enjoy. On average, air pollution occurs in areas such as large cities in a country.

As defined, the presence of a physical, chemical, or biological substance in the air that threatens human health is known as air pollution(Purwanto & Honggara, 2022). Polluted air will harm all living things, as well as result in a decrease in air quality. Site et al. (2022) define the level of air content determined by the concentration of pollutants in a particular location as air quality.(Ridho & Mahalisa, 2023). Factors that cause this include human factors such as cigarette smoke, industrial factors, transportation factors, and many other factors that cause air pollution. Air pollution is also very dangerous because dirty air can trigger various diseases such as asthma, shortness of breath, etc.

Aini et al. (2022) argue that ISPU is a report of air quality monitoring findings detailing the health effects of clean or polluted air(Amalia et al., 2022). We can use the Air Pollution Standard Index (ISPU) to measure air pollution. The parameters contained in the standard air pollution index include particulates (PM_{10}, $PM_{2.5}$), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and ozone (O_3). The parameters of this standard air pollution index will later be collected and calculated. The resulting value from this calculation will provide information about air quality at certain levels, such as moderate, unhealthy, and good.

In analyzing the air environment, data mining can be used to identify and evaluate the data contained in the ISPU. According to D. Kartini, a method for finding certain rules or patterns in large amounts of data and for searching for new information is known as data mining (Umri

et al., 2021). One way is to apply the K-Nearest Night classification method to identify the quality level of air pollution.
This research is based on several related references, such as the reference with the title "Application of the K-Nearest Neighbor (Knn) Algorithm for Air Pollution Classification in the City of Jakarta" with an accuracy result of 95.78% (Nurjanah et al., 2020). Then the reference entitled "Air Quality Prediction Using the K-Nearest Neighbor Algorithm" with accuracy results of 96%, by(Amalia et al., 2022).

As explained in the introduction above, the K-Nearest Neighbor algorithm can be used to analyze and predict air pollution. Therefore, the results of this research can provide information and help increase public awareness about air pollution and its impact on the environment and health, and also the importance of maintaining the environment and good air quality.

2. Literature Review

A. Air pollution

Air pollution is evil pollution that surrounds us and is very dangerous and detrimental to health. Air pollution can come from various sources, such as burning rubbish, fossil fuels, cigarette smoke, transportation fumes, industrial activities, etc. According to Waluyo (xxxx), EC, The entry or entry of a substance, energy, or other component into the air due to human activities that exceed established air quality standards is known as air pollution(Abidin et al., 2019).

Many factors play a role in polluting the air environment in a polluted air environment. Apart from these factors being caused by natural factors, pollution factors originating from human activities also have a greater impact. Factors that influence air quality include: Transportation factors, namely motor vehicle exhaust fumes. Motor vehicle exhaust is one of the main sources of air pollution in urban areas. Industrial factors: One example of this industrial factor is smoke from industrial factories. Industrial factories in big cities produce various pollutants that can pollute the air environment. The forest fire factor, air pollution caused by forest fires, can be caused by humans or nature. Besides impacting air quality, forest fires, such as asthma or shortness of breath, can also affect human health. Illegal logging factors, illegal logging factors are factors caused by humans. This illegal logging can affect the balance of the ecosystem and reduce air quality, this can cause a decrease in oxygen levels in the air and worsen climate change. Household factors, air pollution factors in the household environment, namely using a cooling system or air conditioner (AC), cigarette smoke, burning rubbish, and so on.

B. Data Mining

According to Ramadina, a process that uses special methods and algorithms with the aim of finding a pattern or information in a number of data is called data mining(Adhi Putra, 2021). According to Siregar and Paspabhuana (xxxx) is defined as a procedure used naturally to fully investigate and get to the surface of complex connections from a very large information index. Meanwhile, according to Aribowo, one technique for finding information in large databases is data mining(Nurjanah et al., 2020). Data mining is a process of extracting valuable knowledge or information from large, complex, and often structured datasets (Jollyta et al., 2020). The primary goal of data mining is to identify patterns, relationships, and trends that may not be discovered through conventional means (Iriadi & Nuraeni, 2016). This process involves utilizing statistical techniques, mathematics, and artificial intelligence to dissect data and reveal insightful information. Data mining aims to extract hidden patterns or information within data, allowing users to gain deeper understanding of observed phenomena.

It deals with large, complex, and often multidimensional datasets (Sembiring & Hasibuan, n.d.). It encompasses data from various sources and can involve millions or more entities. The data mining process involves using various algorithms or techniques such as clustering,

classification, regression, association, and others. The selection of algorithms depends on the analysis objectives and data characteristics. One of the primary goals of data mining is to make predictions based on discovered patterns. The results of the analysis can be used to support better decision-making in various fields.

K-Nearest Neighbor Algorithm

According to RP Fitrianti, A. Kurniawati, and D. Agusten, the k-nearest neighbor algorithm is a classification with a core basis that is less solid but relies more on category labels attached to training data that is similar to the test data. In addition, Sari (xxxx) defines the k-nearest neighbor algorithm using training data to classify objects that are closest to each other. K-nearest neighbor uses a direct principle, which is based on the shortest distance between the sample to be tested and the sample to be trained(Adhi Putra, 2021). The k-nearest neighbor algorithm itself is included in the supervised learning algorithm and aims to solve problems in classification methods. With a formula like the following:

$$D_{xy} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Where the stages are:

1.) By determining the value of k.
2.) Then calculate the Euclidean distance on the dataset.
3.) Sort objects based on Euclidean distance and determine the majority of the data.
4.) Then, the data to be classified is given a label or class based on the majority of the closest neighbors.

C. *RapidMiner*

*Rapidminer*or, what was originally called YALE (Yet Another Learning Environment), was developed at the University of Dortmund, Germany in 2001. Rapidminer itself is an open-source software that is used to carry out data mining or data analysis processes. Meanwhile, according to MS Mustafa and IW Simpen, one of the software tools for processing data mining is Rapidminer. Text analysis is the core of rapid-miner text mining, which combines statistical techniques, artificial intelligence, databases, and patterns from large data sets(Nurjanah et al., 2020). Many people use Rapidminer itself for data mining because it is user-friendly, very flexible, has various programming languages , and has good data visualization capabilities, making this software very popular. Due to its advantages, the author uses a rapid miner as a tool to carry out data mining stages of air pollution levels using the k-nearest neighbor classification method.

3. Design/Methods

Research Stages

The stages in this research include steps namely:

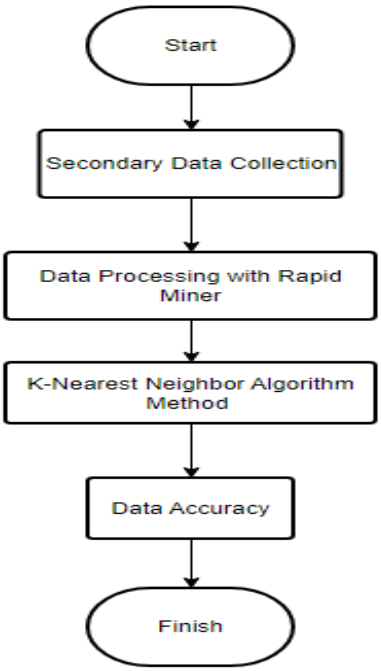


Figure 1. Research Stages

Explanations at this research stage include: 1) Secondary Data Collection. This stage is carried out by collecting various raw data related to air pollution via the kaggle.com website. 2) Data Processing with RapidMiner. After the raw data is collected, the data processing stage is carried out using the rapid-miner application. 3) K-Nearest Neighbor Algorithm Method. The next stage is to choose the right algorithm method to analyze the data. In this case, the k-nearest neighbor method was chosen to carry out the data analysis process. 4) Data Accuracy. The next stage is to carry out a testing process and evaluation process on air pollution data using rapidminer to determine the accuracy of the results.

Research Instrument
According to Sugiyono, research instruments are devices used to measure and view natural and social events that are being considered. Meanwhile, according to DiscoverPhDs, all tools that researchers can use to collect, measure, and analyze data related to research problems or subjects are considered research instruments (Kurniawan, 2021).

Devices, facilities, or tools that play an important role in research and are used to obtain and process data information related to the research object are called research instruments. The instruments used include: 1) Microsoft Office Excel 2019, which was used for initial processing of the dataset.

Tools RapidMiner Studio Version 10.1, which functions to carry out data processing using the k-nearest neighbor algorithm for data mining.

4. Results

Manual Calculation
Manual calculations begin by cleaning the data, dividing training data, and testing data using rapidminer. Data that has been divided and cleaned will be calculated manually using Microsoft Excel. The following is the training data table:

Table 1. Training Data

No	PM_{10}	$PM_{2,5}$	SO_2	CO	O_3	NO_2	Category
1	73	126	38	26	46	34	Not Good
2	53	70	40	14	55	25	Keep
3	32	53	40	11	42	19	Keep
4	36	59	40	14	47	24	Keep
5	29	51	40	14	45	35	Keep
6	34	53	40	8	57	15	Keep
7	33	55	40	10	57	13	Keep
8	26	44	39	10	54	17	Keep
9	33	57	40	13	47	22	Keep
10	50	64	40	13	49	16	Keep
.							
.							
.							
301	64	90	52	44	37	33	Keep

Source: Data Processing Results (2023)

After determining the training data, the next step is to determine the testing data. The following is the testing data table:

Table 2. Testing Data

No	PM_{10}	$PM_{2,5}$	SO_2	CO	O_3	NO_2	Category
1	50	80	55	20	65	40	?

Source: Data Processing Results (2023)

The next stage is to calculate the Euclidean distance as in the calculation formula below:

$$Dxy = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$$Dxy = \sqrt{(73-50)^2 + (126-80)^2 + (38-55)^2 + (26-20)^2 + (46-65)^2 + (34-40)^2}$$

$$Dxy = \sqrt{529 + 2116 + 289 + 36 + 361 + 36}$$

$$Dxy = \sqrt{3367}$$

$$Dxy = 58.02585631$$

The calculation will continue until the 301st calculation. The calculation results that have been obtained can be seen in the following table of Euclidean distance calculation results:

Table 3. Table of Euclidean Distance Calculation Results

No	PM_{10}	$PM_{2,5}$	SO_2	CO	O_3	NO_2	Category	Distance
1	73	126	38	26	46	34	Not Good	58,02
2	53	70	40	14	55	25	Keep	26,36
3	32	53	40	11	42	19	Keep	48,25
4	36	59	40	14	47	24	Keep	38,44
5	29	51	40	14	45	35	Keep	44,36
6	34	53	40	8	57	15	Keep	45,19
7	33	55	40	10	57	13	Keep	45,07
8	26	44	39	10	54	17	Keep	53,64
9	33	57	40	13	47	22	Keep	41,71
10	50	64	40	13	49	16	Keep	36,90
.								
.								
.								
301	64	90	52	44	37	33	Keep	42,82

Source: Data Processing Results (2023)

After calculating the Euclidean distance, the next thing to do is calculate the rank or order of the distance. In this process, the data will be ranked or ordered based on the value from the smallest data to the largest data.

Table 4. Ranking Calculation Results

No	Category	Distance	Ranking
1	Not Good	58,02	271
2	Keep	26,36	62
3	Keep	48,25	238
4	Keep	38,44	171
5	Keep	44,36	212
6	Keep	45,19	221
7	Keep	45,07	219
8	Keep	53,64	255
9	Keep	41,71	194
10	Keep	36,90	152
.			
.			
.			
301	Keep	42,82	200

Source: Data Processing Results (2023)

Data whose Euclidean distance has been calculated and ranked, then the classification will be determined based on testing data with a value of K=7.

Table 5. Manual Calculation Result K=7

Distance	Ranking	Category	K=7
9,11	1	Keep	Yes
14	3	Keep	Yes
16,76	6	Keep	Yes
16,91	7	Keep	Yes
16,18	5	Keep	Yes
12,08	2	Keep	Yes
16,03	4	Keep	Yes

Source: Data Processing Results (2023)

Based on the results of the manual calculations above, using testing data and determining the value K=7, it can be concluded that the resulting category is included in the MEDIUM category because most of the K=7 classification results produce the MEDIUM category.

Calculations using RapidMiner

1.) Pre-Processing (Cleaning) Stage

Pre-processing or data cleaning is carried out with the aim of normalizing data and cleaning empty or invalid data. In rapidminer, the data will be imported, and 1 attribute will be selected, which will be used as a target or class. The next stage is data selection, where it will be determined which attributes will be used and which attributes will not be used. In this case, the attributes used are the PM_{10}, PM2.5, SO2, CO, O3, NO2, and Category attributes.

Open in Turbo Prep Auto Model Filter (365 / 365 examples): all

Row No.	categori	pm10	pm25	so2	co	o3	no2
1	SEDANG	43	?	58	29	35	65
2	SEDANG	58	?	86	38	64	80
3	SEDANG	64	?	93	25	62	86
4	SEDANG	50	?	67	24	31	77
5	SEDANG	59	?	89	24	35	77
6	SEDANG	73	?	81	29	66	85
7	SEDANG	36	?	52	22	55	72
8	SEDANG	38	?	68	26	51	71
9	SEDANG	60	?	77	34	42	80
10	SEDANG	24	?	39	16	38	59
11	SEDANG	51	?	72	17	57	68
12	SEDANG	29	?	58	20	44	77
13	SEDANG	36	?	47	17	32	68
14	SEDANG	36	?	78	20	38	65
15	SEDANG	52	?	82	20	56	65

ExampleSet (365 examples, 1 special attribute, 6 regular attributes)

Source: Data Processing Results (2023)

Figure 2. Selected Data

After the data is selected, the data will go through a cleaning process for empty data or invalid data.

Open in Turbo Prep Auto Model Filter (334 / 334 examples): all

Row No.	categori	pm10	pm25	so2	co	o3	no2
1	TIDAK SEHAT	73	126	38	26	46	34
2	SEDANG	53	70	40	14	55	25
3	SEDANG	32	53	40	11	42	19
4	SEDANG	36	59	40	14	47	24
5	SEDANG	29	51	40	14	45	35
6	SEDANG	34	53	40	8	57	15
7	SEDANG	33	55	40	10	57	13
8	SEDANG	26	44	39	10	54	17
9	SEDANG	33	57	40	13	47	22
10	SEDANG	50	64	40	13	49	16
11	SEDANG	38	57	43	13	35	17
12	SEDANG	63	98	43	16	33	42
13	SEDANG	59	89	40	12	40	16
14	SEDANG	55	73	40	11	42	19
15	SEDANG	42	66	40	13	37	25

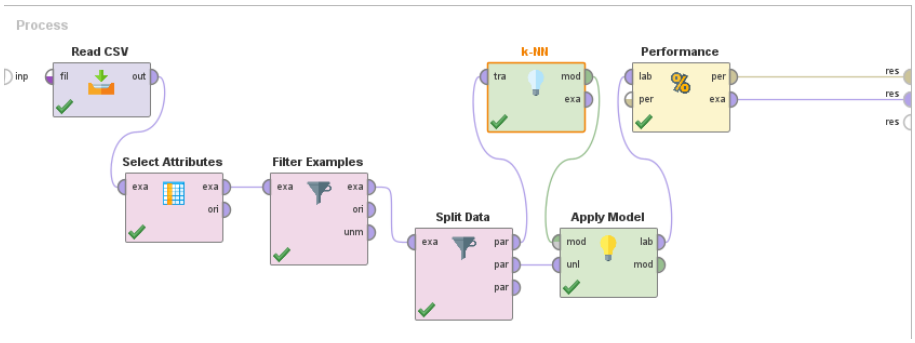
ExampleSet (334 examples, 1 special attribute, 6 regular attributes)

Source: Data Processing Results (2023)

Figure 3. Cleaned Data

2.) Testing Phase

In this section, the data will be tested using the k-nearest neighbor algorithm method. The initial step taken for this testing stage is the data sharing stage, by dividing the data into training data and testing data. This data division uses a ratio of 10% for testing data and 90% for training data. The data that has been shared will go through the k-nearest neighbor algorithm selection process using the classification method. The following is the process of testing the k-nearest neighbor algorithm classification method using rapidminer software.



Source: Data Processing Results (2023)

Figure 4. Data Testing Process View

3.) RapidMiner Accuracy and Precision Results

The value obtained with the rapidminer software using the k-nearest neighbor classification method by selecting the value K=7 results in an accuracy of 93.94%. Where the precision value is 94.74% for the medium class, the value is 92.86% for the unhealthy class, and 0.00% for the healthy class.

5. Conclusion

Based on the results of the analysis that has been carried out in the implementation of data mining using the k-nearest neighbor algorithm with secondary data classification methods to predict the level of air pollution, a conclusion can be obtained, namely: Air pollutants have a huge impact on the environment and human health. The classification method with the k-nearest neighbor algorithm can be applied to identify, analyze and predict air pollution levels with good accuracy results. The results of the level of accuracy resulting from predicting air pollution levels using rapidminer software with the k-nearest neighbor classification method show good results, namely 93.94%. With these high accuracy results, it can have a positive impact on humans and the environment, such as increasing public awareness of the importance of protection and health related to air pollution.

References

Abidin, J., Artauli Hasibuan, F., kunci, K., Udara, P., & Gauss, D. (2019). PENGARUH DAMPAK PENCEMARAN UDARA TERHADAP KESEHATAN UNTUK MENAMBAH PEMAHAMAN MASYARAKAT AWAM TENTANG BAHAYA DARI POLUSI UDARA. *Prosiding SNFUR-4*.

Adhi Putra, A. D. (2021). Analisis Sentimen pada Ulasan pengguna Aplikasi Bibit Dan Bareksa dengan Algoritma KNN. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *8*(2), 636–646. https://doi.org/10.35957/jatisi.v8i2.962

Amalia, A., Zaidiah, A., Isnainiyah, I. N., Program,), S1, S., Informasi, S., Komputer, I., Veteran, U., Jl, J. R., Fatmawati, P., & Labu, J. (2022). *PREDIKSI KUALITAS UDARA MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR*.

Iriadi, N., & Nuraeni, N. (2016). kajian penerapan metode klasifikasi data mining algoritma C4. 5 untuk prediksi kelayakan kredit pada bank mayapada jakarta. *Jurnal Teknik Komputer Amik BSI*, *2*(1), 132–137.

Jollyta, D., Ramdhan, W., & Zarlis, M. (2020). *Konsep Data Mining Dan Penerapan*. Deepublish.

Kurniawan, H. (2021). *Pengantar Praktis Penyusunan Instrumen Penelitian*. Deepublish.

Larasati, D. A. (2023). Penerapan Kecerdasan Buatan dalam Analisis Data untuk Pengambilan Keputusan yang Lebih Baik. *Paradoksal Article*, *1*(1).

Nurjanah, S., Siregar, A. M., & Kusumaningrum, D. S. (2020). Penerapan Algoritma K – Nearest Neighbor (KNN) Untuk Klasifikasi Pencemaran Udara Di Kota Jakarta. *Scientific Student Journal for Information, Technology and Science*, *1*(2), 71–76.

Purwanto, D. D., & Honggara, E. S. (2022). *Klasifikasi Kategori Hasil Perhitungan Indeks Standar Pencemaran Udara dengan Gausian Naïve Bayes (Studi Kasus : ISPU DKI Jakarta 2020)*. *04*(2), 102–108. https://doi.org/10.52985/insyst.v4i2.259

Ridho, I. I., & Mahalisa, G. (2023). ANALISIS KLASIFIKASI DATASET INDEKS STANDAR PENCEMARAN UDARA (ISPU) DI MASA PANDEMI MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM). *Technologia : Jurnal Ilmiah*, *14*(1), 38. https://doi.org/10.31602/tji.v14i1.8005

Sembiring, M. T., & Hasibuan, C. F. (n.d.). *Data Science: Strategi UMKM dalam Pengambilan Keputusan*.

Setianingsih, S., Chasanah, M. U., Kurniawan, Y. I., & Afuan, L. (2023). IMPLEMENTATION OF PARTICLE SWARM OPTIMIZATION IN K-NEAREST NEIGHBOR ALGORITHM AS OPTIMIZATION HEPATITIS C CLASSIFICATION. *Jurnal Teknik Informatika (Jutif)*, *4*(2), 457–465.

Umri, S. S. A., Firdaus, M. S., Primajaya, A., Studi, P., Informatika, T., Komputer, F. I., Karawang, U. S., Karawang, K., Machine, V., Neighbors, K., Bayes, N., Backpropagation, N. N., Neighbors, K., & Machine, S. V. (2021). *Analysis and Comparison of Classification Algorithm in Air*. *4*(2), 98–104. https://doi.org/10.33387/jiko

Wicaksono, R. R., KM, S., KKK, M., Putri, M. S. A., ST, S., Sulistiono, E., Ismarina, S. S. T., Hanif, M., ST, S., & KL, M. (2023). *Manajemen kesehatan lingkungan*. Cendikia Mulia Mandiri.

Zulfiansyah, A. D. K., Kusuma, H., & Attamimi, M. (2023). Rancang Bangun Sistem Pendeteksi Keaslian Uang Kertas Rupiah Menggunakan Sinar UV dengan Metode Machine Learning. *Jurnal Teknik ITS*, *12*(2), A166-173.