

Application of the K-Nearest Neighbor Algorithm for Rainfall Prediction Based on Weather Conditions

Nita Cahyani¹, Rahmat Irsyada²

¹Universitas Padjadjaran

²Politeknik Negeri Subang

Corresponding Author: Nita Cahyani, Nita, e-mail: nita.cahyani@unpad.ac.id

Published: Agustus, 2026

ABSTRACT

Weather and climate patterns in Indonesia are often erratic and difficult to predict. Sometimes, rain fails to fall during the rainy season, while conversely, heavy rains—sometimes leading to flooding—occur during the dry season. Such situations can cause hardship for the community; for instance, during the dry season, farmers may plant crops that require little water—such as tobacco—which are generally unsuitable for cultivation during the rainy season. Research based on BMKG data indicates that rainfall occurs during July and August, while the dry season begins in May and June. To address these issues, scientific prediction can be employed by analyzing actual or historical data; this involves a process of analyzing current data to forecast future events, utilizing data mining for prediction and the K-Nearest Neighbor (K-NN) algorithm for classification. The research model in this thesis employs a development methodology covering software design, modeling, construction, and system delivery to the end-user. The system development follows the Waterfall model. The Weather Prediction System utilizes the K-Nearest Neighbor algorithm—implemented via PHP, MySQL, and RapidMiner—to facilitate future rainfall prediction. Data mining calculations regarding factors influencing rainfall using the K-NN method (with a 70:30 training-to-testing data ratio) yielded the following accuracy results: 98.87% for K=3, 97.30% for K=5, and 97.16% for K=7. Consequently, the use of the K-Nearest Neighbor method for prediction can assist in resolving these issues.

Keywords: Data Mining, Prediction, Weather, Rainfall, K-Nearest Neighbor.

© The Author(s). 2021 Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0>

INTRODUCTION

Weather and climate patterns are often erratic and difficult to predict; for instance, rain may fail to fall during the wet season, while conversely, heavy rains causing floods can occur during the dry season. Such unpredictability creates difficulties for the community—such as when farmers plant crops requiring little water, like tobacco, during the dry season (crops that are generally unsuitable for the wet season), or conversely, when dealing with water-intensive crops like rice. Consequently, the ability to predict the weather is crucial, allowing farmers and the general public to anticipate future conditions and prepare for potential adverse events. These issues can be addressed through scientific prediction methods involving classification processes that utilize historical data to forecast future conditions. Addressing this challenge requires historical rainfall data to generate future weather predictions. Weather refers to the state of the atmosphere at specific times and in relatively small, localized areas; it is short-lived and subject to constant change. According to the World Climate Conference, weather encompasses the overall state of the atmosphere, including the changes, development, and dissipation of atmospheric phenomena. (1)

Indonesia is a tropical country. While the land is highly fertile, the region is prone to natural disasters, particularly flooding. Factors contributing to climate change in Indonesia include ozone

layer depletion, global warming, the greenhouse effect, uncontrolled CFC usage, and industrial emissions. These factors inevitably impact various sectors in Indonesia, notably agriculture and livestock farming, where there is a risk of losses due to crop failure and livestock mortality during periods of extreme cold. Climate changes affecting rainfall patterns and air temperatures have led to a significant decline in agricultural production. Extreme weather events, such as floods and droughts, further increase the likelihood of crop damage. The peak of the dry season is characterized by a drop in temperature; this phenomenon is explained by the BMKG Deputy for Climatology, Drs. Herizal, M.Si: "The clearer the skies during the dry season, the colder the air feels at night and towards dawn." (2)

Previous research conducted in 2019 examined the K-Nearest Neighbors (KNN) algorithm for forecasting rainfall and rice crop temperatures. The use of ICT in agriculture is often referred to as electronic agriculture (e-Agriculture). Agricultural information serves as a key application for promoting and creating opportunities for development and poverty alleviation. The data collection method employed in this study was observation, with the data subsequently processed using the K-NN algorithm. The data utilized originated from Statistics Finland; precipitation data were collected between 2000 and 2010, while temperature data were gathered between 2000 and 2001. Analysis of data from the last 30 years (1981–2010) reveals that Indonesia exhibits 407 distinct climatological climate patterns. Of these, 342 are classified as Seasonal Zone (ZOM) patterns—characterized by clearly defined wet and dry seasons driven by seasonal cycles (typically monsoon patterns)—while the remaining 65 are classified as non-seasonal zones (non-ZOM). The latter typically feature two rainfall peaks per year (equatorial patterns) or consistently high or low rainfall throughout the year. (3)

Data mining offers an alternative solution for data processing and can be utilized across various scientific and research fields, as data mining modeling yields superior and quantifiable data. (4).

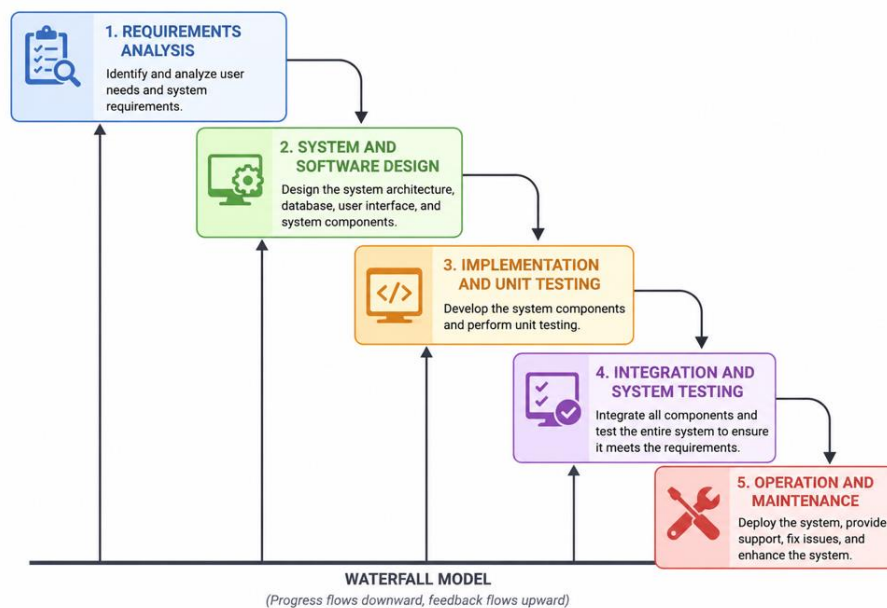
To achieve the best results in this study, the researcher employs the k-nearest neighbor algorithm as a classification technique within data mining. The execution of the k-nearest neighbor algorithm is based on the shortest distance calculated during computation. (5) The calculation process is based on the distance to the new target, determining the distance to the old target using the k-nearest neighbor method. After calculating the distances for the selected K-value, the next stage involves decision-making through the process of determining the K-value.

METHOD

1. System Development Model

The Waterfall method is proposed for the development of the system to be used. It outlines a methodical, sequential approach to software development that begins with communicating user requirements, proceeds through the stages of planning, modeling, and construction (deployment), delivers the system or software to the user, and concludes with ongoing system support. (6)

Figure 1. Waterfall Model



1.1 Requirements Analysis

Regardless of whether the developed application meets the specifications, a system requirements analysis is essential to support the application's performance. Meeting the application's objectives relies on fulfilling system requirements. The following are several analyses to support the system.

1.1.1 Data analysis.

This stage involves data collection. The data is obtained from the publicly accessible BMKG website and covers the four-year period from 3 years. Following collection, the data is split into training and testing sets.

1.1.2 User analysis.

The analysis of system usage aims to provide information on the individuals involved in operating the system, ensure compliance with policies regarding protected assets, and facilitate proper and prudent record-keeping. (6)

1.1.3 System analysis.

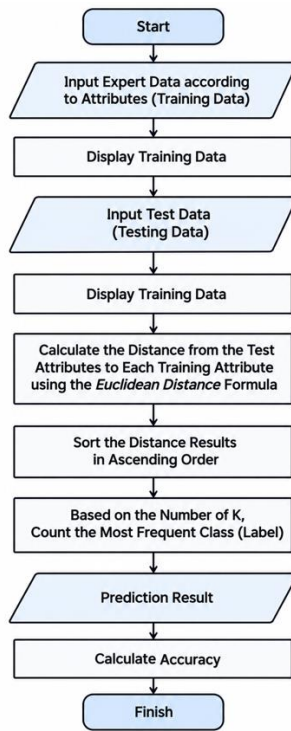
System analysis is a term that collectively describes the initial stages of system development and design. It is a preliminary problem-solving step that characterizes system components by examining how well they function and interact with one another to achieve desired objectives. This stage requires a thorough understanding and familiarity with the existing system, as well as the use of relevant information and data.

1.2 System and Software Desain.

1.2.1 Flowchart

A flowchart—also known as a process flow diagram—is a diagram that illustrates the steps and decisions involved in completing a programming process. Each step is represented graphically and connected by lines or arrows. Flowcharts play a crucial role in defining the phases or functionality of software projects that involve multiple team members. Furthermore, using these diagrams makes the program flow clearer and more concise, thereby reducing the likelihood of misinterpretation. In the realm of programming, flowcharts also serve as an excellent bridge between technical and non-technical requirements. Below is a system flowchart.

Figure 2. Flowchart System



1.2.2 Application Design Draft

1. Login Page

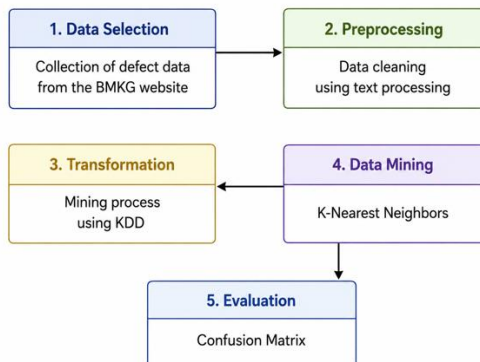
The first step before accessing the system is to log in as an administrator by entering a username and password; this relates to the page design.

2. Halaman Dashboard.

1.3 Implementation and Unit Testing.

In the data mining process, users follow the stages of KDD, ranging from the initial goal-setting phase to the evaluation process.

Figure 3. Datamining Process



1.3.1 Data Selection

The primary data in this study consists of daily rainfall records from the past four years, obtained from the BMKG Meteorological Station in Tuban. Meanwhile, secondary data were collected through documentation techniques and a review of literature including both journals and books pertaining to meteorology and geophysics.

1.3.2 Preprocessing.

Preprocessing is a necessary step prior to data mining. This stage involves data integration to combine data from different databases, followed by data cleaning to produce a clean dataset suitable for the subsequent mining phase. These two processes are explained below.

a) Data Integration

This stage involves merging data from various distinct databases to enable integration. Data integration is performed on specific data attributes to identify unique entities. (7).

b) Data Cleaning

This stage marks the initial step of the KDD process. During this phase, irrelevant data (specifically missing values) and redundant data must be cleaned. A data point is considered to have a "missing value" if an attribute within the dataset contains a null value, whereas "redundant data" refers to the presence of more than one record containing identical values within the same dataset. The data undergoes this cleaning process to better meet the requirements for weather classification.

1.3.3 Transformation.

The transformation stage involves modifying the specified data so that it can be processed for data mining. This transformation process employs KDD methods and depends on the types and patterns of information sought within the database. At this stage, operational data—derived from grouped attributes such as temperature, rainfall, solar radiation, maximum wind speed, and average wind speed—is utilized for the data mining transformation. Pattern calculation is performed using cross-validation techniques, as illustrated in the following figure.

Figure 4. Cross Validation Technique

Experiment	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Experiment 1	Test Set	Training Set	Training Set	Training Set	Training Set
Experiment 2	Training Set	Test Set	Training Set	Training Set	Training Set
Experiment 3	Training Set	Training Set	Test Set	Training Set	Training Set
Experiment 4	Training Set	Training Set	Training Set	Test Set	Training Set
Experiment 5	Training Set	Training Set	Training Set	Training Set	Test Set

1.3.4 Data Mining

This stage involves data modeling using the K-Nearest Neighbor (KNN) algorithm. Prior to modeling, the dataset is divided into training and testing sets (for generalization purposes), utilizing various split ratios for comparison. Specifically, the testing involves the following ratios: 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20, and 90:10.

Note:

Training Data

Training data is used for data training and modeling within a classification process to produce a high-quality data format that is easily applicable to testing data. (8).

Data Testing

Conversely, testing data consists of values obtained from the system and used as a reference to determine whether the resulting output demonstrates a high level of accuracy. (8).

1.3.5 Evaluation

The value derived from the data mining process must be presented in a form that is easily understood by the intended users. This process involves verifying the integrity of the discovered patterns or information against existing facts or hypotheses. Specifically, regarding weather classification patterns derived via the K-Nearest Neighbor method, the resulting patterns or information take the form of rules generated through the K-Nearest Neighbor algorithm's calculations.

1.4 Integration and System Testing

Once the programming code has been written, the next step is to test the resulting program and determine whether the application operates in accordance with the expected procedures. The testing method employed by the author is black-box testing.

Table 1. Black Box Testing

No.	Module / Feature	Test Scenario	Testing Level	Testing Method
1	Login	Verify that the login page can be accessed successfully.	Unit	Black-Box Testing
2	Dashboard	Verify that the dashboard page is displayed correctly.	Unit	Black-Box Testing
3	K-Nearest Neighbor Classification	Verify that the classification page is displayed correctly.	Unit	Black-Box Testing
4	Dataset Management	Verify that the dataset page can be accessed successfully.	Unit	Black-Box Testing
5	Initial Processing	Verify that the initial processing page is displayed correctly.	Unit	Black-Box Testing
6	Prediction	Verify that the prediction page is displayed correctly.	Unit	Black-Box Testing
7	Performance Evaluation	Verify that the performance evaluation page is displayed correctly.	Unit	Black-Box Testing
8	Add Dataset	Verify that users can add a new dataset successfully.	Unit	Black-Box Testing
9	Logout	Verify that users can log out of the application successfully.	Unit	Black-Box Testing

2. Data collection technique

At this stage, data collection is employed to obtain the information needed to achieve the research objectives. The following are the data collection techniques used.

2.1 Observation

Observation is a data collection process involving direct field observation of the object being studied (9). In the initial stage, the researcher conducted direct observations at the research site—the office of the PPK O&P SDA IV in Bojonegoro Regency. These observations covered the conditions of the research location and the acquisition of weather data from the BMKG.

2.2 Documentation

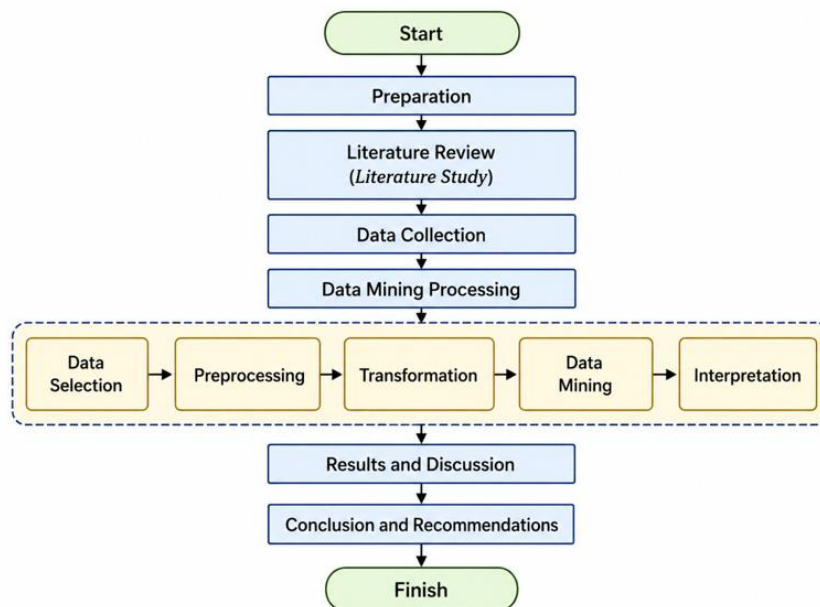
Documentation is a data collection process involving the analysis of documents whether created by the subject themselves or by others regarding the subject (10). At this stage, the researcher requested permission from the BMKG regarding the data to be analyzed; the data was obtained from the BMKG's public website.

2.3 Literature Review

A literature review is the process of studying relevant literature such as scientific articles, books, news reports, or online sources related to the research topic (11). In the initial stage of the research, the researcher studied data mining using existing literature, such as books and journals available online.

3. Research Phase.

Figure 5. Research Stages



The research process outlines the steps to be taken and provides an overview of the study as a whole. The stages involved are as follows:

- Preparation: This stage focuses on the BMKG as the subject, with data serving as the object of study.
- Literature review: This stage involves examining and reviewing existing literature related to rainfall prediction.
- Data collection: Data is gathered by retrieving public information from the BMKG website.
- Data processing: The collected data is processed in accordance with the stages of Knowledge Discovery in Databases (KDD). Hasil dan Pembahasan, pembahasan terkait hasil dari proses data minning yang dilakukan menggunakan metode algoritma K-Nearest Nighbor.

FINDING AND DISCUSSION

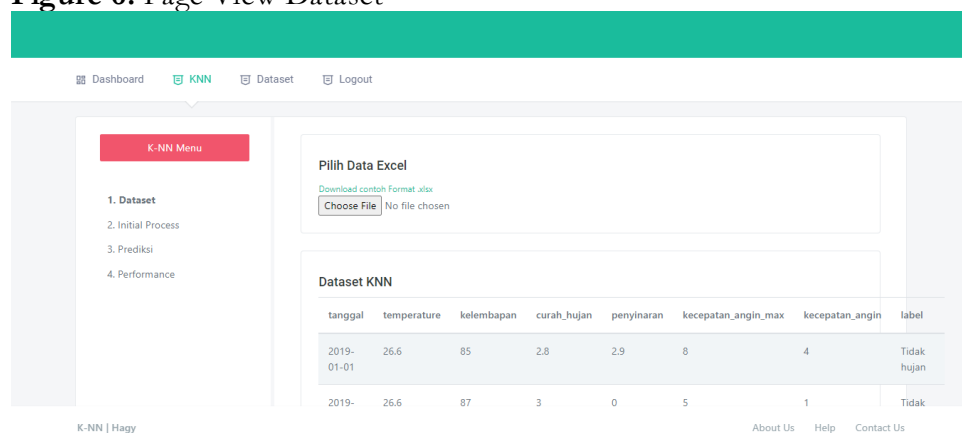
1. System Implementation

System implementation was carried out by translating the previously created system design into program code. During this process, the system design was realized through coding. The programming languages and technologies employed include PHP (Hypertext Preprocessor), HTML (HyperText Markup Language), and the MySQL database.

Dataset Page

The Dataset page includes all previously obtained measurement and expert data. The layout of the Dataset page is shown in the following figure.

Figure 6. Page View Dataset



Prediction Page

The prediction page facilitates the process of calculating the squared Euclidean distance between a query instance and its K-nearest neighbors; this involves inputting data samples for processing by the system application. The prediction process page is shown in the following figure.

Figure 7. Prediction Page

Performance Page

The performance page features accuracy testing methods and includes both training and testing data, with testing-to-training data ratios ranging from 10% to 90%. The following shows the performance page layout.

Figure 8. Performance Page

On the performance page, following the comparison of testing data and training data, the results of the data testing expressed as a percentage are displayed. The following shows the performance results derived from the testing and training data.

Figure 9. Accuracy display on the Performance page

2. Black-Box Test Results

The results of the software testing conducted using the black-box method aligned with expectations.

3. Validation Test Results

Based on the feasibility test questionnaire, there are six key aspects: language presentation, operational workflow, ease of understanding, ease of access, responsiveness, and test result accuracy. The feasibility test results can be found in the appendix. The questionnaire for the Rainfall Prediction System was distributed to 20 test engineers; the following details the average scores derived from their responses.

$$\begin{aligned}\text{Average score} &= \frac{\text{Average score}}{\text{Number of items}} \\ &= \frac{1340}{320} = 3.72\end{aligned}$$

Selanjutnya dihitung dalam bentuk presentase dengan perhitungan sebagai berikut.

$$\begin{aligned}\text{Score percentage} &= \frac{\text{Average score}}{\text{ideal score}} \times 100\% \\ &= \frac{3.72}{4} \times 100\% \\ &= 93.05 \%\end{aligned}$$

The calculation above yielded a score percentage of 93.05%, according to Khairunnisa as cited by Mimin Ninawati, Burhendi, and Wulandari (12), states that the results are categorized according to the following eligibility criteria.

Table 2. Eligibility Criteria

Percentage Range	Criteria
86% - 100%	Very Good
76% - 85%	Good
60% - 75%	Fair
≤55% - 59%	Poor

4. Testing Using RapidMiner

The following are the stages of rainfall data testing using RapidMiner tools:

a) Dataset Input

This stage involves importing data into RapidMiner. A CSV file format is used for the data implementation. The "Read CSV" operator is employed, utilizing the import configuration wizard for the data input process.

b) Labeling

This stage involves assigning a class to the label. The "Set Role" operator is used in RapidMiner; the process involves selecting the attribute name and then designating the target role as "label."

c) Split Data

This stage involves partitioning the data into training and testing sets. The "Split Data" operator is used in RapidMiner. The process involves selecting a sampling type to divide the dataset; specifically, stratified sampling is used to create random subsets that maintain approximately equal proportions of the two class label values.

d) Algorithm Method

k-NN is an operator that generates a k-Nearest Neighbor model for regression or classification tasks. The k-Nearest Neighbor algorithm works by identifying the *k* nearest training instances based on *n* attributes within the training ExampleSet. Subsequently, the k-NN algorithm classifies unknown instances based on the majority vote of the identified neighbors.

e) Prediction

In this stage, the "Apply Model" operator is used to apply the previously trained model (derived from the training data) to the unlabeled testing data. The objective is to generate predictions for the data that lacks class labels.

f) Classification

This stage involves a process that benchmarks the accuracy of the classification pattern in predicting class labels for future data. The process utilizes a primary criterion for comparison and performance evaluation; specifically, accuracy is measured as the relative proportion of correctly predicted classifications.

The accuracy results obtained from the data processing are presented in the following table.

Tabel 3. K-Value Result

Value K	Comparison Data	Accuracy Value
3	70:30	98.87%
5		97.30%
7		97.16%

CONCLUSIONS

Based on the research conducted—which addressed the application of the k-Nearest Neighbor algorithm to predict weather conditions influencing rainfall—it is concluded that this study resulted in the development of a rainfall prediction system using the k-Nearest Neighbor method. The process involved utilizing historical data to perform predictions—analyzing past events to forecast future occurrences—through k-Nearest Neighbor calculations. Black-box testing demonstrated that the system functions correctly and meets expectations. The system was evaluated using a 70:30 data split (70% training data and 30% testing data) with RapidMiner serving as the computational tool. Testing with the k-Nearest Neighbor algorithm yielded an accuracy of 98.87% for K=3, 97.30% for K=5, and 97.16% for K=7.

REFERENCES

- Hairunisa H. Analisis Dampak Musim Hujan Terhadap Hasil Panen Tomat di Desa Ciloto , Kabupaten Cianjur , Provinsi Jawa Barat. 2020;101.
- Dasanto BD, Sulistiyanti S, Anria A, Boer R. Dampak Perubahan Iklim Terhadap Kenaikan Muka Air Laut Di Wilayah Pesisir Pangandaran. Risal Kebijakan Pertan DAN Lingkungan Rumusan Kaji Strateg Bid Pertan dan Lingkungan. 2022;9(2):82–94.
- Ruslana ZN, Giarno G. Edukasi Petani Dalam Memanfaatkan Informasi dan Prakiraan Iklim/Musim Melalui Sekolah Lapang Iklim di Tegalsari, Kedu, Temanggung. JEPKM J Edukasi dan Pengabdian 2022;1(2):41–51.
- Zai C. Implementasi Data Mining Sebagai Pengolahan Data. Portal Data. 2022;2(3):1–12.
- Salsabila A, Yunita R, Rozikin C. Identifikasi Citra Jenis Bunga menggunakan Algoritma KNN dengan Ekstraksi Warna HSV dan Tekstur GLCM. Technomedia J. 2021;6(1):124–37.
- Praja AB, Darmansah D, Wijayanto S. Sistem Informasi Pencatatan Surat Masuk dan Surat Keluar Berbasis Website Menggunakan Metode Waterfall. J Sist Komput dan Inform. 2022;3(3):273.
- Eprilianto DF, Sari YEK, Saputra B. Mewujudkan Integrasi Data Melalui Implementasi Inovasi Pelayanan Kesehatan Berbasis Teknologi Digital. JPSI (Journal Public Sect Innov. 2019;4(1):30.
- Simbolon PH. Implementasi Data Mining Pada Sistem Persediaan Barang Menggunakan Algoritma Apriori (Studi Kasus : Srikandi Cash Credit Elektronik dan Furniture). J Ris Komput. 2019;6(4):401–6.
- Sudrajat RT, Firmansyah D. Peningkatan Kemampuan Menulis Teks Laporan Hasil Observasi Dengan Menggunakan Pendekatan Discovery Di Kelas X Sma Xix Kartika 1 Bandung Semantik. 2020;9(2):157–62.
- Juliyanto F, Parjito. Rekayasa Aplikasi Manajemen E-Filing Dokumen Surat Pada PT ALP (Atoism Lampung Pelayaran). J Teknol dan Sist Inf. 2021;2(1):43–9.
- Surahman E, Satrio A, Sofyan H. Kajian Teori Dalam Penelitian. JKTP J Kaji Teknol Pendidik.

2020;3(1):49–58.
Mimin Ninawati, Burhendi FCA, Wulandari W. Pengembangan E-Modul Berbasis Software iSpring Suite 9. J Educ FKIP UNMA. 2021;7(1):47–54.