Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12879
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparison Analysis of C4.5 Algorithm and KNN Algorithm for Predicting Data of Non-Active Students at Prima Indonesia University

Jepri Banjarnahor[1)], Ferman Zai[2)*] Janiali Sirait[3)] Dicky Wijaya Nainggolan [4)] Nissi Grace Dian Sihombing[5)]
[1,2,3,4,5)] Universitas Prima Indonesia
jepribanjarnahor@unprimdn.ac.id[1)], fermantigan@gmail.com, [2)] janisirait2906@gmail.com, [3)] dicky.nbisnis@gmail.com [4)] gracediannissi@gmail.com [5)]

Submitted: Aug 14, 2023 | **Accepted**: Aug 18, 2023 | **Published**: Oct 1, 2023

Abstract: Education is important nowadays because universities need to improve their students' skills so they can compete in the globalization era. Education can be obtained through both formal and informal channels, and knowledge is available everywhere, especially in today's world where information tools are rapidly evolving. Inactive students are students who do not participate in a course for a maximum of two consecutive semesters. Students who are not active have the opportunity to drop out of university studies. Students who drop out of college are usually motivated by economic factors, and the cessation of the lecture process can cause inactivity and administrative costs. Therefore, this research was conducted using the C4.5 algorithm method and the K-Nearest Neighbor (KNN) algorithm to compare and predict data on inactive students at Universitas Prima Indonesia. The research continued with the data collection and data preprocessing stages, after which the data mining process was carried out to get the final results of this research. The testing process follows the process of comparing the C4.5 algorithm and the K-Nearest Neighbor (KNN) algorithm with K-fold crossing. This evaluation step is compared by considering the comparison values of the confusion matrix (precision, precision, recall). The accuracy results obtained by each algorithm provide information about the effectiveness of using these techniques in processing the specified dataset. The accuracy of the Decision Tree C4.5 algorithm is 99.12% and the K-Nearest Neighbors algorithm is 99.14%. Based on research conducted using the K-Nearest Neighbors and C4.5 algorithms to predict inactive students, the KNN algorithm is more accurate than the C4.5 algorithm.

Keywords: Analysis; Predicting; Algorithm C4.5; KNN Non active students

INTRODUCTION

Education is important at this time because universities need to improve the ability of their students so they can compete in the era of globalization. Education can be obtained through formal or informal channels, and knowledge is available everywhere, especially in today's world where information tools are growing rapidly according. Education is a process by which students build a learning process that enables them to develop their potential to serve the nation and society. However, formal education recognized by educational institutions is still necessary and complete for Indonesian students. (Anestiviya et al., 2021)

Inactive students are students who do not participate in a course for a maximum of two consecutive

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12879 p-ISSN : 2541-044X

semesters. Students who are not active have the opportunity to drop out of university studies Journal Quotes from Researchers (Noviana et al., 2019). Students who drop out of college are usually motivated by economic factors, and the cessation of the lecture process can cause inactivity and administrative costs. (Karyono, 2016). However, the problem of student inactivity is a problem for tertiary institutions, especially at Prima Indonesia University. Several factors cause student inactivity, including economic factors, academic abilities, and others. (Salam et al., 2020) Prediction of Potentially Inactive Students Using Data Mining in Decision Trees and Algorithms C4.5. Higher education management must anticipate and take action on students with "inactive" status, to find out the factors that cause this problem, it is necessary to take action right in the middle of the student's study period. Students who can be indicated to be inactive will be able to alleviate problems at Universitas Prima Indonesia.

Based on the results of the data, it is predicted that students who are potentially non-active will have an impact on Universitas Prima Indonesia. Therefore, Prima Indonesia University must periodically evaluate the implementation of the learning system and record student activities. This data collection was carried out to improve the quality of a university, especially at Prima Indonesia University. Therefore, this research was conducted using the C4.5 algorithm method and the K-Nearest Neighbor (KNN) algorithm to compare and predict data on inactive students at Universitas Prima Indonesia. The method used to process this data uses the C4.5 algorithm and the K-Nearest Neighbors (KNN) algorithm. In general, the C4.5 algorithm approach shows a high degree of accuracy, (Haryanto et al., 2023). On the other hand, the K-Nearest Neighbors (KNN) algorithm is a powerful algorithm for training data that contains a lot of noise. (Karyono, 2016). This research is expected to provide accurate results, which can be used by Prima Indonesia University, especially the Faculty of Technology and Computer Science, to predict inactive student data

This step was taken so that researchers could find out the problems in compared the C4.5 and K-Nearest Neighbor (KNN) algorithms to predict inactive student data at Universitas Prima Indonesia (UNPRI). The data used includes data on active and non-active students at UNPRI from 2019 – 2022 which were analyzed into a Python-based program. Data taken from Universitas Prima Indonesia (UNPRI) consists of several variables such as Student Name, Name, Gender, Temporary IPS, Absence, Failed Courses, Type of Residence, Class Schedule, Parents' Income, Last UK Status, Student Status, Faculty and Study Program. Further information about the Attribute data used is shown in the following table.

LITERATURE REVIEW

Types of research

The method used in this study is related to observation techniques that identify problems within the framework of literature research. The research continued with the data collection and data preprocessing stages, after which the data mining process was carried out to get the final results of this research. This study then used this data, starting with a dataset divided into training and testing data (Salam et al., 2020). In this study, the researchers used split data and carried out a data mining process using the C4.5 and K-Nearest Neighbor (KNN) algorithms to predict inactive student data. As a result, we get a comparison result for the C4.5 algorithm. In addition, the K-Nearest Neighbors (KNN) algorithm can generate assessment parameters. This research method uses the C4.5 algorithm model and the K-Nearest Neighbors (KNN) algorithm.

Testing this model describes data mining testing techniques to compare the C4.5 and K-Nearest Neighbor (KNN) algorithms used in this study. The procedure for predicting inactive student data using the C4.5 algorithm and the K Nearest Neighbors (KNN) algorithm is based on forward selection. The testing process follows the process of comparing the C4.5 algorithm and the K-Nearest Neighbor (KNN) algorithm with K-fold crossing. This evaluation step is compared by considering the comparison values of the confusion matrix (precision, precision, recall)

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12879
e-ISSN : 2541-2019
p-ISSN : 2541-044X

METHOD

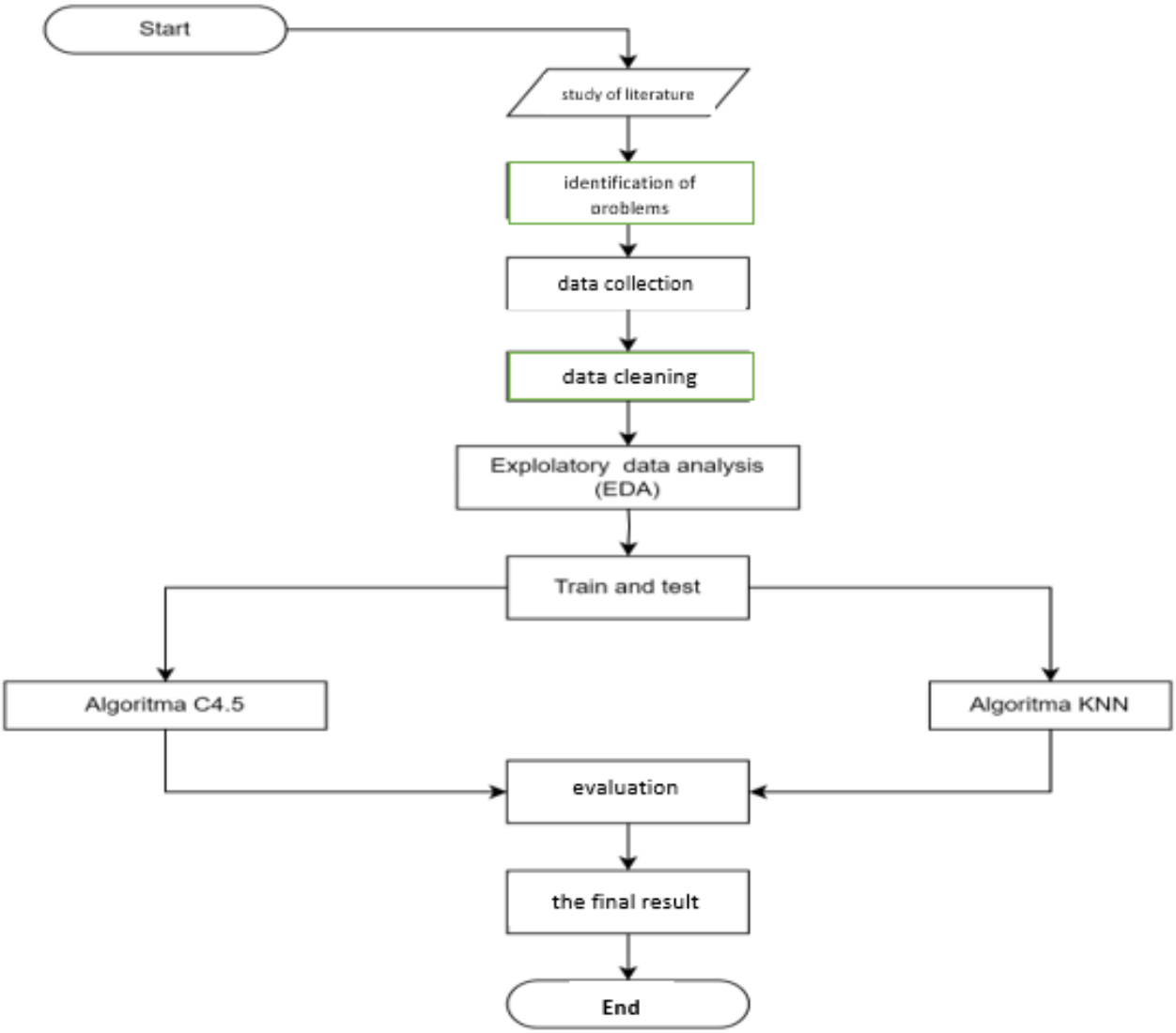

Figure 1. Research Flow

Research attributes

Table 1. research attributes

Attribute Name	Artibut Description Value	Artibut Description Value	Description Value
Nim	Nim Student	Unique Code	223303030246,
Student Name	The Personal Identity Of Lentina Tindaon Student, Ronaldo Harlim.	Personal Identity Of Lentina Tindaon Student, Ronaldo Harlim.	The Personal Identity Of Lentina Tindaon Student, Ronaldo Harlim.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12879
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Gender	Gender Identity Male And Female	Identity Male And Female	Identity Male And Female
Ips	Ips Gpa Provisions Student	Achievement Scores In The Last Semester Between 0 – 4	Student Achievement Scores In The Last Semester Between 0 – 4
Absent Student	Absent Student Attendance	Number Of Student Attendance Between 1 – 100	Number Of Student Attendance Between 1 – 100
Number Of Failed Courses	Number Of Failed Courses	Courses Number Of Students Who Fail To Meet The Minimum Standard 0,8,0,0, And So On	Number Of Failed Courses Failure Courses Number Of Students Who Fail To Meet The Minimum Standard 0,8,0,0, And So On
Type Of Residence	Type Of Residence Type Of Residence Information About The Location Of Student.	Type Of Residence Type Of Residence Information About The Location Of Student Residence With Parents, Boarding Houses, Guardians And So On	Type Of Residence Type Of Residence Information About The Location Of Student Residence With Parents, Boarding Houses, Guardians And So On

Data Preprocessing

Data preprocessing is a series of techniques used to clean, change and prepare raw data so that it can be processed by machine learning algorithms. Preprocessing data is the stage of data processing carried out by researchers before carrying out the classification process. The preprocessing stage generally functions to clean the data so that the next processing step is more structured. In research conducted preprocessing data includes three main stages:

Data selection

The data selection process involves selecting relevant information and can be applied when comparing the C4.5 algorithm and the K-Nearest Neighbor (K-NN) algorithm to predict student activity at Universitas Prima Indonesia.

Data Cleaning

In this study, data cleaning or data cleaning is done by deleting data that has lost value. To eliminate unnecessary data, researchers filter the data obtained during the dataset acquisition stage to produce data that researchers really need. Missing data is blank data (null) which is not used in this study. Missing values can be caused by errors in data collection, errors in data entry, information not provided, hard to find, or non-existent. From all the data collected, 26,838 data were found missing values for several attributes. The number of missing value data for attributes is shown in the following table.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12879
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2. Number of Missing Value data

Attribute To The Amount Of Data Missing Value	Attribute To The Amount Of Data Missing Value
Nim 0	Nim 0
Schedule Of Courses 1415	Schedule Of Courses 1415
Parents Income 1451	Parents Income 1451
Last Uk Status 1451	Last Uk Status 1451
Courses Fail 0	Courses Fail 0
Status Of Students 1451	Status Of Students 1451
Temporary Ips 2636	Temporary Ips 2636
Absent 1451	Absent 1451
Residence 1456	Residence 1456
Faculty 1451	Faculty 1451
Study Program 1451	Study Program 1451

The number of missing values in the dataset represents the number of empty data in each attribute. The total number of missing data (Null/NaN) in the data set is 26,838 because there are multiple attributes with null or empty values in the data rows. Of the 11 attributes used, there are 9 attributes whose data is missing: Class Schedule (1451), Parental Income (1451), Last UK Status (1451), Student Status (1451), Temporary IPS (2636), Absent (1451), residence (1556), faculty (1451) and study program (1451). Researchers delete these values to get records that do not contain zero/blank values for each attribute. Below is the result data with missing values for each attribute.

Table 3. Number of Missing Value data

ATRIBUT	THE NUMBER OF MISSING VALUE DATA AFTER BEING DELETED
Nim	0
Course Schedule	0
Parents Income	0
Latest Uk Status	0
Course Failed	0
Student Status	0
Temporary Ips	0
Roll Call	0
Residence	0
Faculty	0
Study Program	0

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12879 p-ISSN : 2541-044X

Data Transformation

Researchers used two types of data in this study. This type of data consists of categorical and numeric data. Due to the different types of data, changes are needed so that the shape is the same (Prihandoko, 2018). To work around this problem, use the LabelEncoder function from the Python sklearn module. After encoding, the modified data is entered into new columns such as class schedule, absent category, social studies category, parents' income, last UK status, student status, place of residence and number of failed courses. In order to be processed more easily by analytical algorithms or machine learning models, each data is assigned a numeric label. The following details the data labels for each attribute:

Processing Data

Table 4. Class Schedule attribute labeling

Course schedule	Course schedule
Morning Regular Scholar	1
Evening Regular Scholar	2
Bachelor of Night Acceleration	3
Undergraduate Special Program	4
Executive Bachelor	5

Table 5. Labeling of parental income attributes

Parents' income	Parents' income
Nothing	6
Less Than Rp. 500.000	0
Rp.500,000 - Rp. 999,999	5
Rp. 1,000,000- Rp. 1,999,999	2
Rp. 2,000,000- Rp. 4,999,999	3
Rp. 5,000,000- Rp. 20,000,000	4
More Than Rp. 20,000,000	1

Table 6. Last UK Status attribute labeling

Last UK Status	Last UK Status
Not yet Paid 0	Not yet Paid 0
Paid 1	Paid 1

Table 7. Labeling of student status attributes

Student status	Student status
Active 0	Active 0
Inactive 1	Inactive 1

Exploratory Data Analysis (EDA)

Researchers conducted Exploratory data analysis (EDA) to analyze data on non-active students at Universitas Prima Indonesia (UNPRI). Exploratory data analysis (EDA) plays a very important role in processing data into useful information and providing value to users. Exploratory data analysis (EDA) was performed using visualization techniques and descriptive statistics to understand the data set used. Exploratory Data Analysis (EDA) is a process of analyzing a set of data to summarize its main characteristics in order to understand the condition of the dataset. From the EDA analysis the researchers used the outliers, histogram, and heatmap methods, the researchers concluded that the data the researchers observed did not have significant anomalies. The distribution of the data was relatively normal and homogeneous, and there were no strong linear relationships or obvious anomalies between the observed variables.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12879 p-ISSN : 2541-044X

Data Sharing

After completing the exploratory data analysis (EDA) stage, the next step is to separate the data into two parts: training data and test data. The percentage of data separation is 70% training data and 30% test data. The sum of the two parts depends on the total number of initial records available after the pre-processing process is complete.Tabel 10. Pembagian data

Table 8. Data sharing

predictions	Amount of data	Data Training (70%)	Data Testing (30%)
Aktif	13916	9.741	4.175
Non aktif	1327	928	399
Total	**15.243**	**10.669**	**4.574**

System planning

Next is to design a decision support system which is the initial stage in building a decision support system.after building the system we will enter the stage of testing the system whether it is feasible or not used by users. **Results and conclusions**

The conclusions obtained after going through the testing stages, the conclusions stages are also useful forfurther researchers.

RESULT

In analyzing and designing a good system, data and information are needed that are appropriate and in accordance with system requirements. This can be obtained by analyzing the system in advance or that is currentlyrunning.

Comparison Results of the C4.5 Algorithm and the KNN Algorithm

The accuracy results obtained by each algorithm are presented below to provide information about the effectiveness of using these techniques in processing the specified dataset.

Table 9. Accuracy

Algoritma	Akurasi (%)
Decision Tree C4.5	99.12%
K-Nearest Neighbors	**99.14%**

From the results of the comparison of the algorithms above, it can be seen that the predicted results using the KNN algorithm are better than the C4.5 algorithm.

Implementation of Results Based on the Best Algorithm

The implementation of the results in this study was carried out using the knn algorithm. This is because this algorithm provides the best prediction of inactive students with the highest accuracy compared to the C4.5 algorithm. Implementation of K-Nearest Neighbor Algorithm Results Using k-NN with Varying Number of Neighbors (k-NN Varying number of neighbors) and ROC Curve.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12879 p-ISSN : 2541-044X

Visualisasi *(k-NN Varying number of neighbors)*

Visualization using (k-NN Varying number of neighbors) to predict inactive students at Prima Indonesia University can be seen in the image below.

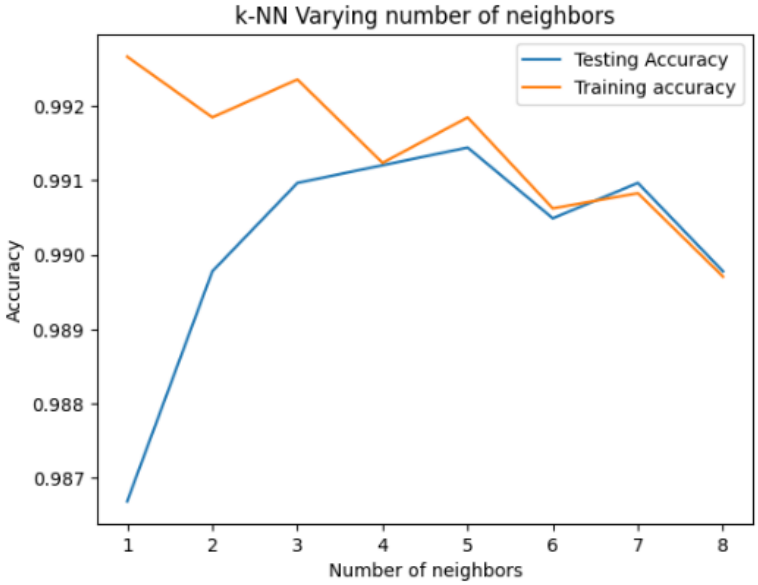

Figure 2. Visualisasi *(k-NN Varying number of neighbors)*

DISCUSSIONS

In this study, the authors wanted to examine a decision regarding the selection The researchers analyzed the data to compare the C4.5 and K-Nearest Neighbor (KNN) algorithms to predict inactive student data at Universitas Prima Indonesia (UNPRI). The data used includes data on active and non-active students at UNPRI from 2019 – 2022 which were analyzed into a Python-based program. Data taken from Universitas Prima Indonesia (UNPRI) consists of several variables such as Student Name, Name, Gender, Temporary IPS, Absence, Failed Courses, Type of Residence, Class Schedule, Parents' Income, Last UK Status, Student Status, Faculty and Study Program.

As for some suggestions submitted by researchers, namel Dig further by comparing the C4.5 and KNN algorithms to other machine learning algorithms, such as Random Forest, Naive Bayes, or Support Vector Machines. Expanding this research by using a larger and more diverse dataset to assess the reliability and generalizability of the prediction results of the C4.5 and KNN algorithms. Applying the comparison of the C4.5 and KNN algorithms to more specific use cases, such as classifying other case studies or predicting student failure.

CONCLUSION

Based on the results of research, the system needs to addtest data in order to get maximum results with SPK calculations. The system needs to be added to determine subcriteria in order to get maximum results with MOORA. Based on research conducted using the K-Nearest Neighbors and C4.5 algorithms to predict inactive students at Prima Indonesia University, the KNN algorithm is more accurate with a value of 99.14%, while the C4.5 algorithm provides an accuracy of 99.12%.

*name of corresponding author

REFERENCES

Anestiviya, V., Ferico, A., & Pasaribu, O. (2021). Analisis Pola Menggunakan Metode C4.5 Untuk Peminatan Jurusan Siswa Berdasarkan Kurikulum (Studi Kasus : Sman 1 Natar). *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, *2*(1), 80–85. http://jim.teknokrat.ac.id/index.php/JTSI

Atma, Y. D., & Setyanto, A. (2018). Perbandingan Algoritma C4.5 dan K-NN dalam Identifikasi Mahasiswa Berpotensi Drop Out. *Metik Jurnal*, *2*(2), 31–37.

Dewi, N. A. K., Zukhri, A., & Dunia, I. K. (2014). Analisis Faktor-Faktor Penyebab Anak Putus Sekolah Usia Pendidikan Dasar di Kecamatan Gerokgak Tahun 2012 / 2013. *Jurnal Pendidikan Ekonomi Undiksha*, *4*(1), 1–12. https://ejournal.undiksha.ac.id/index.php/JJPE/article/view/1898

Gaol, N. Y. L. (2020). Prediksi Mahasiswa Berpotensi Non Aktif Menggunakan Data Mining dalam Decision Tree dan Algoritma C4.5. *Jurnal Informasi & Teknologi*, *2*, 23–29. https://doi.org/10.37034/jidt.v2i1.22

Haryanto, C., Rahaningsih, N., & Muhammad Basysyar, F. (2023). Komparasi Algoritma Machine Learning Dalam Memprediksi Harga Rumah. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 533–539. https://doi.org/10.36040/jati.v7i1.6343

Husen, A. H., Nur Afiah, A. S., Soesanti, S., & Tempola, F. (2022). Deteksi Dini Resiko Tuberkulosis di Kota Ternate: Pelacakan dan Implementasi Algoritma Klasifikasi. *Jurnal CoSciTech (Computer Science and Information Technology)*, *3*(2), 217–225. https://doi.org/10.37859/coscitech.v3i2.3986

Jurnal, H., Mambang, M., Hidayat, A., Dona Marleny, F., Wahyudi, J., Informasi, T., & Sari Mulia, U. (2022). Jurnal Informatika Dan Teknologi Komputer Explanatory Data Analisis Untuk Mengevaluasi Penelusuran Kata Kunci Video Pembelajaran Di Youtube Dengan Pendekatan Machine Learning. *Juli*, *2*(2), 181–189.

Karyono, G. (2016). ANALISIS TEKNIK DATA MINING " ALGORITMA C4.5 DAN K-NEAREST NEIGHBOR " UNTUK MENDIAGNOSA PENYAKIT DIABETES MELLITUS. *Seminar Nasional Teknologi Informasi*, 77–82. http://news.palcomtech.com/wp-content/uploads/downloads/2016/06/IT13_Giat-Karyono.pdf

Latifah, R., Wulandari, E. S., & Kreshna, P. E. (2019). Model Decision Tree Untuk Prediksi Jadwal Kerja Menggunakan Scikit-Learn. *Jurnal Universitas Muhammadiyah Jakarta*, 1–6. https://jurnal.umj.ac.id/index.php/semnastek/article/download/5239/3517

Nasrullah, A. H. (2018). Penerapan Metode C4.5 untuk Klasifikasi Mahasiswa Berpotensi Drop Out. *ILKOM Jurnal Ilmiah*, *10*(2), 244–250. https://doi.org/10.33096/ilkom.v10i2.300.244-250

Nikmatun, I. A., & Waspada, I. (2019). Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. *Jurnal SIMETRIS*, *10*(2), 421–432.

Noviana, D., Susanti, Y., & Susanto, I. (2019). Analisis Rekomendasi Penerima Beasiswa Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Algoritma C4.5. *Seminar Nasional Penelitian Pendidikan Matematika (SNP2M) 2019 UMT*, 79–87.

Novianti, B., Rismawan, T., & Bahri, S. (2016). Implementasi Data Mining Dengan Algoritma C4.5 Untuk Penjurusan Siswa (Studi Kasus: Sma Negeri 1 Pontianak). *Jurnal Coding, Sistem Komputer Untan*, *04*(3), 75–84.

Prihandoko, P. (2018). Perbandingan Kinerja Algoritma C4. 5, Naïve Bayes, K-Nearest Neighbor, Logistic Regression, Dan Support Vector Machines Untuk Mendeteksi Penyakit Kanker Payudara. *Jurnal Teknologi Informasi Dan Komunikasi*, *7*(2), 1–10.

Rosandy, T. (2016). Perbandingan Metode Naive Bayes Classifier dengan Metode Decision Tree Untuk Menganalisa Kelancaran Pembiayaan. *Jurnal TIM Darmajaya*, *02*(01), 52–62.

Salam, A., Nugroho, F. B., & Zeniarja, J. (2020). Implementasi Algoritma K-Nearest Neighbor Berbasis Forward Selection Untuk Prediksi Mahasiswa Non Aktif Universitas Dian Nuswantoro Semarang. *JOINS (Journal of Information System)*, *5*(1), 69–76. https://doi.org/10.33633/joins.v5i1.3351

Wanto, A. (2016). Analisis Penerapan Fuzzy Inference System (FIS) Dengan Metode Mamdani Pada Sistem Prediksi Mahasiswa Non Aktif (Studi Kasus : AMIK Tunas Bangsa Pematangsiantar). *Seminar Nasional Inovasi Dan Teknologi Informasi (SNITI) 3*, *3*, 393–400.

*name of corresponding author