**JURNAL RISET INFORMATIKA** **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 4, No. 2. March 2022 DOI: https://doi.org/10.34288/jri.v4i2.374
**Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020**

# THE EFFECT OF AMOUNT OF DATA ON RESULTS OF ACCURACY VALUE OF C4.5 ALGORITHM ON STUDENT ACHIEVEMENT INDEX DATA

**Anton Sunardi [1*)], Sienny Rusli [2], Christina Juliane [3]**

Information System Master's Program
STMIK LIKMI
https://likmi.ac.id
antonsunardi@gmail.com[1*)], siennyrusli@gmail.com[2], Christina.juliane@likmi.ac.id[3]
(*) Corresponding Author

**Abstract**

Of the many academic data, data in the form of an achievement index needs to be used in-depth so that it does not become a display of numbers and information only. This achievement index evaluation data reflects the educational process students and teaching staff carries out in an educational process. This study aims to measure the accuracy of data mining processing based on differences in test data by analyzing the C4.5 algorithm using RapidMiner as a data processing tool and determining the decisions students can make and academic institutions in developing study strategies and educational curricula to be maximized. The data processing is carried out by classifying the student achievement index data at a private university using data analysis test equipment. The data source comes from kaggle.com, which consists of 1687 data that have been processed and processed. The conclusion from the results of this study is that the amount of data turns out to have a significant influence on the accuracy value of the C4.5 algorithm, where an accuracy rate of 91.69% is obtained from the test results of 1687 data with four main attributes, namely IPK1, IPK2, IPK3, IPK4 and correctly or not as a label.

Keywords: the amount of data, C4.5, achievement index, data mining

*Abstrak*

*Dari sekian banyak data akademik, data berupa Indeks Prestasi perlu dimanfaatkan secara mendalam agar tidak menjadi tampilan deret angka dan informasi saja. Data evaluasi Indeks Prestasi ini merupakan cerminan dari proses pendidikan yang dilakukan pelajar, mahasiswa, dan tenaga pengajar dalam suatu proses pendidikan. Penelitian ini bertujuan untuk mengukur tingkat akurasi pengolahan data mining berdasarkan perbedaan jumlah data uji dengan menganalisa algoritma C4.5 menggunakan RapidMiner sebagai alat bantu olah data, dan untuk mengetahui keputusan yang dapat diambil oleh pelajar, mahasiswa dan institusi akademis dalam menyusun strategi studi dan kurikulum pendidikan agar lebih maksimal. Proses olah data dilakukan dengan mengklasifikasikan data Indeks Prestasi mahasiswa pada sebuah perguruan tinggi swasta menggunakan alat uji analisis data. Sumber data berasal dari kaggle.com yang terdiri dari 1687 data yang telah diproses dan diolah. Kesimpulan dari hasil penelitian ini adalah jumlah data ternyata memiliki pengaruh signifikan terhadap nilai akurasi algoritma C4.5, dimana tingkat akurasi sebesar 91.69 % didapatkan dari hasil uji terhadap 1687 data dengan 4 atribut utama yaitu IPK1, IPK2, IPK3, IPK4 dengan Tepat atau tidak kelulusan sebagai label.*

*Kata Kunci: jumlah data, C4.5, indeks prestasi, data mining*

## INTRODUCTION

From time to time, along with the rapid development of the world of data and education, a person is required to improve knowledge and skills to have a good and quality thinking pattern to plan strategies in the face of future competition. Performance evaluations are stored and collected by various universities in the form of student achievement index data that universities can use as one of the supports that the management of education organizers can use to determine their educational strategies. The way to use historical data is to process it using data mining methods, one of which is by applying data classification methods so that patterns and rules can produce helpful information to support students and educational institutions in developing learning strategies and educational strategies.

In previous studies, the level of accuracy in processing student graduation data using several methods can also be tested using machine learning or deep learning techniques, as has been done by (Maryanto, 2017). From the explanation in the research, data processing requires a tool that can measure the level of accuracy. According to 2020 college statistics (Handini et al., 2020) The amount of student data recorded nationally compared to the number of students enrolled in the Ministry of Education and Culture has a significant difference. Therefore, research is needed to apply extensive data collection to support the industrial revolution as a breakthrough in rapid technological progress.

Research that has been done previously by (Budiman & Ramadina, 2015) regarding predictions using the data mining classification algorithm conducted by (Windarti & Suradi, 2019) applies large amounts of data to data analysis concepts that can help readers, especially students and teaching staff and related agencies. Their field determines educational strategies by considering various aspects, especially the influence of GPA and the accuracy of the Student Achievement Index data results.

This research is presupposed to provide benefits for students as an early reminder about the potential for untimely graduation so that students can develop a more effective study plan strategy. For academic institutions to foreknown provide information in the form of patterns and images that can be used to determine policies in minimizing the student's potential untimely graduation, which is not timely in the scientific field, it is hoped that this research can provide changes in data mining testing techniques. The classification method uses the C4.5 algorithm for varying amounts of data so that this research can be used as a reference for parties in need.

## RESEARCH METHODS

The method used in data collection and reference consists of analytical techniques used to classify data and is done by selecting large amounts of data sourced from Kaggle.com, then sorting and cleaning steps to be arranged according to research needs, namely 1687 data divided into five research attributes. Data processing is carried out using the RapidMiner data processing tool to find C4.5 accuracy through careful calculations to find the best level of precision to produce data references to be implemented. The data is applied through the process of extracting patterns from data using training data of 80% and testing data of 20%, then conducted a comparison test of the accuracy rate of the C4.5 algorithm, with the number of data 100, 400, 900, 1250, 1450, and 1687 from a private university in Indonesia, referring to the Student Achievement Index. The research flow can be presented in figure 1 below:

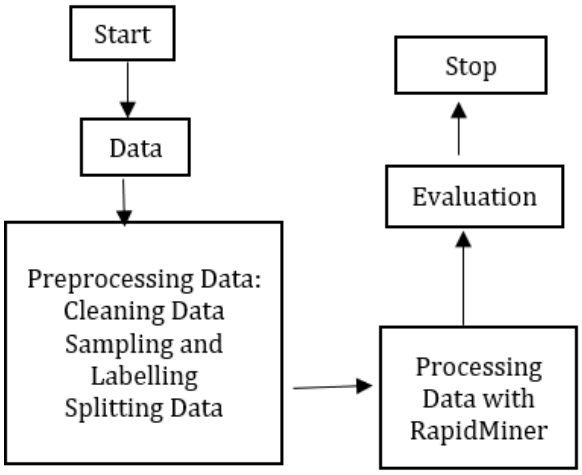


Figure 1. Research Flow

**Research Type**

According to (Chapman et al., 2000), CRISP-DM is a method that connects the background to be achieved to data usage and provides an overview of the data cycle. Research using similar techniques was conducted by (Sabna & Muhardi, 2016) regarding the data cycle of the data wealth of college students. This study uses six stages of the data processing process, namely:

1. Business Understanding

The focus of this stage is more on understanding the objectives and requirements than turning this knowledge into the purpose of extracting data.

2. Data Understanding

Data understanding begins with data collection, then getting to know the data, identifying data quality, looking for insights, and detecting groups of data parts that can generate hypotheses on confidential information.

3. Data Preparation

(Amir & Abijono, 2018) It is posited that the stages in the preparation of data are processed in several locations and can be non-sequential as modelling tools are performed labelling, attribute selection, transformation, and data cleaning to build the final dataset. RapidMiner is a data processing tool used to find patterns, designs, knowledge, and evaluation of large amounts of data. It is an open-source learning machine that contains data tools for pre-processing, classification, rule, and association

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i2.374
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

so that it is easy to visualize stated. (Muis & Affandes, 2015)

4. Modelling

This research refers to previous research by (Romadhona, Suprapedi, & Himawan, 2017). In his discussion of data modelling using the C4.5 algorithm decision tree, he explains that there is a higher level of accuracy when compared to ID3 and Chaid algorithms. Therefore, modelling is determined and adjusted to achieve optimal values at the modelling stage. At this stage, the authors tested the amount of data against the level of accuracy by comparing data < 1000, namely 100, 400, and 900, and data >1000 data, namely 1250, 1400, and 1687. (Hermawanti, Asriyanik, & Sunarto, 2019) found an accuracy rate of 68.42% using 145 test data, comparing < 1000 data i.e. 100, 400, 900 and data > 1000 data namely 1250, 1400 and 1687. This study used training data of 80% test data of 20% to produce optimal accuracy values referring to discussions done previously by (Musu, Ibrahim, & Heriadi, 2021)

5. Evaluation

At the evaluation stage, an assessment of the data that has been generated from accuracy values based on Confusion Matrix, under curve area values (AUC) and execution time (ET). (Olson & Shi, 2007) his book entitled Introduction to the Science of Business Data Excavation explains that a confusion matrix produces four types of classifications, namely true Positive (TP), true negative (TN), false positive (FP), and false-negative (FN). The formulation of the confusion matrix in table form is described in table 1 below:

Table 1. Confusion Matrix Table (Olson and Shi 2007)

| | | True Value | |
|---|---|---|---|
| | | TRUE | FALSE |
| Predicted Value | True | TP (True Positive) Correct result | FP (False Positive) Unexpected result |
| | False | FN (False Negative) Missing result | TN (True Negative) Correct absence of result |

Information:
True Positive (TP): The amount of positive and predicted data to be accurate as Positive.
False Positive (FP): The amount of harmful data but predicted as Positive.
False Negative (FN): The amount of data that is positive but predicted as Negative.
True Negative (TN): The amount of negative and predicted data to be accurate as Negative.

Referring to research conducted by (Azhari, Situmorang, & Rosnelly, 2021), it is mentioned that precision is the accuracy of getting information from accurate positive and negative class data. In addition to the accuracy value, there is also a value to recall specific information obtained from the recall value. The comparison value in testing against research data is obtained from the system's predictive value and the tester's prediction value called accuracy. Once the results of the Confusion Matrix calculation are obtained, estimations for precision, recall, and accuracy values can be calculated as in the measures in Formula 1 below:

$$\text{Precision} = \frac{TP}{TP+FP} \quad \ldots\ldots (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad \ldots\ldots (2)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots\ldots (3)$$

Formula1. Table Precision, Recall, and Accuracy (Olson & Shi, 2007) Formula with the following description:
TP: True Positive
FP: False Positive
TN: True Negative
FN: False Negative

6. Deployment

According to (Saefulloh & Moedjiono, 2013), after the formation of the model, further analysis and measurements are carried out at the previous stage. At this stage, the most accurate model or rule is applied to predict timely graduation and can then be used to evaluate new data. The concept of data deployment refers to the application of a model to predict the accuracy results of 91.69% of 1687 data that students, educators, and readers can use to monitor data processing strategy plans with the implementation of data mining if the purpose of the model is to increase knowledge about data, knowledge. What is obtained needs to be arranged and presented so that students, educators, and related institutions can use it? The implementation stage produces research reports that can be used for repetitive data mining. Students or students, educators, and educational institutions that carry out deployment steps need to be understood to redevelop necessary knowledge so that the results have use value.

**Procedure**

This research was conducted by referring to various literature studies in the form of research results in journals and books seized according to related research needs as a reference for writing. The steps of the data excavation and processing process procedure using RapidMiner that we do can be presented in Figure 2 below:

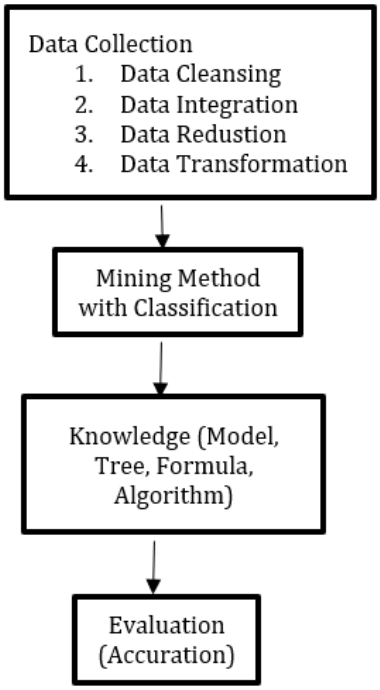


Figure 2. Process of processing data, methods, knowledge, evaluation using RapidMiner (author source)

**RESULTS AND DISCUSSIONS**

(Megna, 2021) In his research in 2021, he believed that large amounts of data are enormous data sets in volume but grow exponentially with the time that traditional data does not have, so it requires management tools that can store or process it to be efficient. Based on this, the diversity of data formats that are increasingly complex and grow over time requires data governance using techniques and technology.

Here are the results obtained based on the crisp-dm methodology theory

1. Business Understanding

The purpose of the study was to determine the amount of data in the dataset that affects the degree of accuracy in the decision tree classification method with the C4.5 algorithm.

2. Data Understanding

The student achievement index dataset is in the form of public data taken from kaggle.com, with 1687 data and four attributes containing student achievement index values and 1 point of appropriate or inappropriate graduation used as a label. The initial data is presented in table 2 and table 3 below:

Table 2. Initial dataset table

| No | SAI1 | SAI2 | SAI3 | SAI4 | Yes / No |
|---|---|---|---|---|---|
| 1 | 2.30 | 1.97 | 1.80 | 1.56 | No |
| 2 | 1.81 | 1.68 | 1.57 | 1.86 | No |
| 3 | 3.07 | 3.00 | 2.75 | 3.21 | No |
| 4 | 2.71 | 2.33 | 2.61 | 1.98 | No |
| 5 | 3.17 | 3.02 | 3.28 | 2.96 | No |
| 6 | 3.16 | 3.45 | 3.02 | 3.06 | No |
| 7 | 2.72 | 2.50 | 2.92 | 3.00 | No |
| . | . | . | . | . | . |
| 1687 | 3.18 | 3.05 | 3.05 | 3.27 | Yes |

Table 3. Dataset description table

| Attribute Name | Information |
|---|---|
| SAI1 | Student Achievement Index 1 |
| SAI2 | Student Achievement Index 2 |
| SAI3 | Student Achievement Index 3 |
| SAI4 | Student Achievement Index 4 |
| Yes / No | Right or Not Graduation |

3. Data Preparation

In the preparatory stage, the dataset was adjusted to the data mining processing process using RapidMiner, and the decision tree classification method was carried out with the C4.5 algorithm and then obtained 1687 data with four attributes. Student Achievement Index and one attribute are appropriate or not timely as labels as presented in table 4 below:

Table 4. Pre Processing data table

| No | SAI1 | SAI2 | SAI3 | SAI4 | Yes / No |
|---|---|---|---|---|---|
| 1 | 2.30 | 1.97 | 1.80 | 1.56 | No |
| 2 | 1.81 | 1.68 | 1.57 | 1.86 | No |
| 3 | 3.07 | 3.00 | 2.75 | 3.21 | No |
| 4 | 2.71 | 2.33 | 2.61 | 1.98 | No |
| 5 | 3.17 | 3.02 | 3.28 | 2.96 | No |
| 6 | 3.16 | 3.45 | 3.02 | 3.06 | No |
| 7 | 2.72 | 2.50 | 2.92 | 3.00 | No |
| . | . | . | . | . | . |
| 1687 | 3.18 | 3.05 | 3.05 | 3.27 | Yes |

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i2.374
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

According to (Dengen, Kusrini, & Luthfi, 2020) The calculation steps in the C4.5 algorithm decision tree using manual calculation methods are as follows: The calculation steps in the C4.5 algorithm decision tree using manual calculation methods are as follows:

1. Prepare dataset samples
2. Calculate the *entropy* value using the formula::

$$\text{Entropy (S)} = \sum_{i=1}^{n} - pi * \log 2\, pi \quad \text{.....(4)}$$

Information:
S: set
N: number of partitions S
Pi: proportion of Si to SS

3. Calculate the gain value of each attribute, followed by selecting the highest gain value

$$\text{Gain (S,A)} = \text{Entropy(S)} - \sum_{i=1}^{n} * \text{Entropy(Si)} \text{ ..... (5)}$$

Information:
S: Set
A: Attributes
n: Number of attribute partitions A
| Si |: The amount of data on the ith partition
| S |: Number of cases in S

4. Modelling

This stage progressed by testing <1000 data with 100, 400, and 900 samples and testing data >1000 with examples taken from 1250, 1400, and 1687. This study processed the composition of training data by 80% and testing data by 20%. Using Rapid Miner, data testing was executed to get performance vector results through the presentation in Table 5 and Table 6.

Table 5. In performance, Vector data < 1000 obtained calculation results for 100 data with an accuracy result of 70% with confusion matrix on prediction "Yes" and prediction result "Yes" of 1, and on prediction "Yes" and prediction result "No" by 5. The results of the Performance Vector 400 data calculation were obtained with an accuracy result of 85% with confusion matrix on prediction "Yes" and prediction result "Yes" of 0, and on projection "Yes" and prediction result "No" of 5. Performance Vector calculation results from 900 data obtained an accuracy result of 86.11% with Confusion Matrix on prediction "Yes" and prediction result "Yes" of 1, and on prediction "Yes" and prediction result "No" of 12.

Table 5. RapidMiner Vector Performance Table (Data < 1000)

| Data < 1000 | | | |
|---|---|---|---|
| Amount of Data | *Performance Vector* | | |
| 100 | Accuracy 70.00 % | | |
| | Confusion Matrix | | |
| | True | No | Yes |
| | No | 1 | 5 |
| | Yes | 1 | 13 |
| 400 | Accuracy 85.00 % | | |
| | Confusion Matrix | | |
| | True | No | Yes |
| | No | 0 | 5 |
| | Yes | 7 | 68 |
| 900 | Accuracy 86.11 % | | |
| | Confusion Matrix | | |
| | True | No | Yes |
| | No | 1 | 12 |
| | Yes | 13 | 154 |

Table 6. RapidMiner Vector Performance Table (Data > 1000)

| Data > 1000 | | | |
|---|---|---|---|
| Amount of Data | *Performance Vector* | | |
| 1250 | Accuracy 86.60 % | | |
| | Confusion Matrix | | |
| | True | No | Yes |
| | No | 0 | 6 |
| | Yes | 20 | 224 |
| 1400 | Accuracy 90.36 % | | |
| | Confusion Matrix | | |
| | True | No | Yes |
| | No | 0 | 3 |
| | Yes | 24 | 253 |
| 1687 | Accuracy 91.69 % | | |
| | Confusion Matrix | | |
| | True | No | Yes |
| | No | 2 | 3 |
| | Yes | 25 | 307 |

Table 6. In performance, Vector data >1000 obtained calculation results for 1250 data with an accuracy result of 86.6% with Confusion Matrix on prediction "Yes" and prediction result "Yes" of 0, and on prediction "Yes" and prediction result "No" of 6. The results of the Performance Vector 1400 data calculation were obtained an accuracy result of 90.36% with confusion matrix on prediction "Yes" and prediction result "Yes" of 0, and on prediction "Yes" and prediction result "No" of 3. Performance Vector 1687 data obtained an accuracy result of 91.69% with Confusion Matrix on the forecast "Yes" and the prediction result "Yes" of

2, and on the prediction "Yes" and the development of the prophecy. "No" of 3.

5. Evaluation

The evaluation is seen from the results of calculating accuracy values based on confusion matrix using RapidMiner obtained results as table 7 below:

Table 7. C4.5 algorithm data evaluation table

| Amount of data | Accuracy | Algorithm of C4.5 | | | |
|---|---|---|---|---|---|
| 100 | 70.00 % | | True no | True Yes | Class Precision |
| | | Prediction No | 1 | 5 | 16.67 % |
| | | Prediction Yes | 1 | 13 | 92.86 % |
| | | Class Recall | 50.00 % | 72.22 % | |
| 400 | 85.00 % | | True no | True Yes | Class Precision |
| | | Prediction No | 0 | 5 | 0.00 % |
| | | Prediction Yes | 7 | 68 | 90.67 % |
| | | Class Recall | 0.00 % | 93.15 % | |
| 900 | 86.11 % | | True no | True Yes | Class Precision |
| | | Prediction No | 1 | 12 | 7.69 % |
| | | Prediction Yes | 13 | 154 | 92.22 % |
| | | Class Recall | 7.14 % | 92.77 % | |
| 1250 | 89.60 % | | True no | True Yes | Class Precision |
| | | Prediction No | 0 | 6 | 0.00 % |
| | | Prediction Yes | 20 | 224 | 91.80 % |
| | | Class Recall | 0.00 % | 97.39 % | |
| 1400 | 90.36 % | | True no | True Yes | Class Precision |
| | | Prediction No | 0 | 3 | 0.00 % |
| | | Prediction Yes | 24 | 253 | 91.34 % |
| | | Class Recall | 0.00 % | 98.83 % | |
| 1687 | 91.69 % | | True no | True Yes | Class Precision |
| | | Prediction No | 2 | 3 | 40.00 % |
| | | Prediction Yes | 25 | 307 | 92.47 % |
| | | Class Recall | 7.41 % | 99.03 % | |

Table 7 shows the percentage value of C4.5 algorithm calculation accuracy against the number of data > 1000, i.e., 1250, 1400, and 1687 and the amount of information < 1000, i.e., 100, 400, and 900. The C4.5 algorithm showed higher accuracy in > 1000 data, namely 1250, 1400, and 1687. It indicates that data processing using RapidMiner from several data samples has a high degree of accuracy on data amounting to > 1000. The highest accuracy obtained is 91.69%, with 1687 data.

6. Deployment

The results of this research are enclosed conceive writing comprise information on differences in the results of calculating the accuracy value of the C4.5 algorithm, which is affected by the amount of data. This paper can be functioned as an information and reference for research by using the following data classification methods.

## CONCLUSIONS AND SUGGESTIONS

### Conclusion

The study concluded that the Achievement Index data processing test using RapidMiner against <1000 data showed a lower graduation accuracy rate than the graduation accuracy rate with the number of data > 1000. The difference in the amount of data tested causes a difference in the results of the accuracy value. The highest accuracy percentage is shown in data 1687 at 91.69%, while the test data amounting to 100 produced the lowest accuracy value of 70%.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i2.374
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

**Suggestion**

To obtain the outcome of the achievement index grouping, appropriate data support is needed. To aim for a more accurate potential value in the C4.5 algorithm technique, the suitability of the type of data and the amount of data is essential because it has a significant effect. The more data tested, the accuracy level will be determined, therefore it is concluded that the amount of data influences the result of processing the accuracy value.

**REFERENCE**

Amir, S., & Abijono, H. (2018). Penerapan Data Mining untuk Mendukung Pemasaran Produk. *CAHAYAtech*, *7*(2), 161–182. Retrieved from https://ojs.cahayasurya.ac.id/index.php/CT/article/view/102

Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4. 5, Random Forest, SVM dan Naive Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(2), 640–651. https://doi.org/10.30865/mib.v5i2.2937

Budiman, I., & Ramadina, R. (2015). Penerapan Fungsi Data Mining Klasifikasi untuk Prediksi Masa Studi Mahasiswa Tepat Waktu pada Sistem Informasi Akademik Perguruan Tinggi. *JUPITER (Jurnal Penelitian Ilmu Dan Teknologi Komputer)*, *7*(1), 39–50. Retrieved from https://jurnal.polsri.ac.id/index.php/jupiter/article/view/709

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T. P., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. In *SPSS inc*. CRISP-DM consortium. Retrieved from https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

Dengen, C. N., Kusrini, K., & Luthfi, E. T. (2020). Implementasi Decision Tree Untuk Prediksi Kelulusan Mahasiswa Tepat Waktu. *SISFOTENIKA*, *10*(1), 1–11. Retrieved from https://www.stmikpontianak.ac.id/ojs/index.php/ST/article/view/484

Handini, D., Hidayat, F., Attamimi, A. N. R., Putri, D. A. V., Rouf, M. F., & Anjani, N. R. (2020). *Statistik Pendidikan Tinggi Tahun 2020*. Jakarta: Sekretaris Direktorat Jenderal Pendidikan Tinggi. Retrieved from Sekretaris Direktorat Jenderal Pendidikan Tinggi website: https://pddikti.kemdikbud.go.id/asset/data/publikasi/Statistik Pendidikan Tinggi 2020.pdf

Hermawanti, S. N., Asriyanik, A., & Sunarto, A. A. (2019). Implementasi Algoritma C4.5 untuk Prediksi Kelulusan Tepat Waktu ( Studi Kasus : Program Studi Teknik Informatika ). *Jurnal Ilmiah SANTIKA*, *9*(1), 853–864. https://doi.org/10.37150/jsa.v9i1.552

Maryanto, B. (2017). Big Data dan Pemanfaatannya dalam Berbagai Sektor. *Media Informatika*, *16*(2), 14-19. Retrieved from https://jurnal.likmi.ac.id/Jurnal/7_2017/0717_02_BudiMaryanto.pdf

Megna, A. A. K. (2021). *Big Data: Development of Revolutionary technologies in Business*. Istanbul.

Muis, I. A., & Affandes, M. (2015). Penerapan Metode Support Vector Machine (SVM) Menggunakan Kernel Radial Basis Function (RBF) Pada Klasifikasi Tweet. *Jurnal Sains, Teknologi Dan Industri*, *12*(2), 189–197. Retrieved from http://ejournal.uin-suska.ac.id/index.php/sitekin/article/view/1010

Musu, W., Ibrahim, A., & Heriadi, H. (2021). Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4.5. *Seminar Sistem Informasi Dan Teknologi Informasi (SISITI)*, 186–195. Makasar: STMIK Dipanegara Makassar. Retrieved from https://www.ejurnal.dipanegara.ac.id/index.php/sisiti/article/view/802

Olson, D., & Shi, Y. (2007). *Pengantar Ilmu Penggalian Data Bisnis - Introduction to Business Data Mining*. Jakarta: Salemba Empat.

Romadhona, A., Suprapedi, S., & Himawan, H. (2017). Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Usia, Jenis Kelamin, Dan Indeks Prestasi Menggunakan Algoritma Decision Tree. *Jurnal Teknologi Informasi CyberKU*, *13*(1), 69–83. Retrieved from http://research.pps.dinus.ac.id/index.php/Cyberku/article/view/10

Sabna, E., & Muhardi, M. (2016). Penerapan Data Mining Untuk Memprediksi Prestasi Akademik Mahasiswa Berdasarkan Dosen, Motivasi, Kedisiplinan, Ekonomi, dan Hasil Belajar. *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer Dan Teknologi Informasi*, *2*(2), 41. https://doi.org/10.24014/coreit.v2i2.2392

Saefulloh, A., & Moedjiono, M. (2013). Penerapan Metode Klasifikasi Data Mining Untuk Prediksi Kelulusan Tepat Waktu. *InfoSys*

*Journal*, *2*(1), 41–54.

Windarti, M., & Suradi, A. (2019). Perbandingan Kinerja 6 Algoritme Klasifikasi Data Mining untuk Prediksi Masa Studi Mahasiswa. *Telematika*, *12*(1), 14–30. Retrieved from https://ejournal.amikompurwokerto.ac.id/index.php/telematika/article/view/778