ISSN: 2093-5374 / E-ISSN: 2093-5382

New Framework of Educational Data Mining to Predict Student Learning Performance

Dr. Agung Triayudi[1*], Rima Tamara Aldisa[2], and S. Sumiati[3]

[1*]Department of Information Communications Technology, Universitas National, Indonesia.
agungtriayudi@civitas.unas.ac.id, https://orcid.org/0000-0002-1269-5888

[2]Department of Information Communications Technology, Universitas National, Indonesia.
rima.tamara@civitas.unas.ac.id, https://orcid.org/0009-0005-8071-1252

[3]Department of Information Communications Technology, Universitas Serang Raya, Indonesia.
sumiati82@yahoo.com, https://orcid.org/0000-0002-7347-5379

Received: September 09, 2023; Revised: November 13, 2023; Accepted: January 12, 2024; Published: March 30, 2024

Abstract

Educational systems designed to meet the needs of academic advisors about adaptive learning will always be an essential issue, as this will be the beginning of the development of intelligent learning methods. In an educational institution, such as in a university environment, academic guidance carried out by a teacher to his students significantly affects the student's performance in the lecture stage, where educational guidance that goes poorly is allegedly causing difficulties for the student in carrying out his studies, or worst chance of dropping out of school. Therefore, this study aims to explore the potential and capabilities contained in the features of Educational Data Mining to predict students' learning performance which will later present various recommendations for academic guidance methods based on data analysis related to academic records and social and economic related data. In this study, we will propose data analysis and testing from recorded student data in an information technology class from a private university in Jakarta. The modelling presented in this study uses the Decision Tree, Neural Networks, and Naïve Bayes methods, which then implement these algorithms on academic data from 300 students of the 2017-2019 and 2018-2020 Information Systems and Informatics study program. From the implementation of data mining techniques in this study, performance results were obtained, which stated that the designed framework provided accurate predictions related to student performance.

Keywords: Educational Data Mining, Framework, Prediction.

1 Introduction

Accurate academic guidance plays an essential role in the continuity of a student's study, closely related to talents and interests, academic background, and other factors regarding the student's personality. Several studies that have been conducted state that there are difficulties in supporting a student in making a decision. This action can lead the students to choose a class following their skills, increasing their risk of dropping out (Mathur, G., et al., 2024). Choosing a suitable specialization class or course will greatly determine a person's career path. Therefore, it is essential to get good guidance to find hidden potential

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 15, number: 1 (March), pp. 115-132. DOI: *10.58346/JOWUA.2024.I1.009*

*Corresponding author: Department of Information Communications Technology, Universitas National, Indonesia.

for smooth study. It is now known that many countries, such as Morocco, have offered mentoring methods that can support students well in terms of decision-making. However, sometimes, it is not easy to give the best choice for a student because several parameters and values related to the study are not met. This problem also appears in various classes from several educational institutions, so it can be concluded that a system that can provide recommendations so that the guidance process runs well is needed for the smooth study of various professions (Livingston, K., 2006) (Akçapýnar, G., 2014) (Shahzad, F., 2021) (Babiker, M., 2015).

The knowledge found can be used to support efforts to improve student learning activities, where the decision-making process is carried out through predictions based on data related to student profiles. In addition, this can also help in detecting any difficulties experienced by a student. Moreover, it can be said that the recommendation system that forms the basis of this research allows an academic institution to identify the character of students who will later influence academic orientation and vocational development as early as possible. Through prediction efforts as well as analysis related to the habits of these students, this also allows academic institutions to recognize more deeply the actual abilities of students, along with the risks that may be experienced, so that the available predictions will be beneficial in providing input regarding the suitable guidance method (Dewar, T., 2000) (Triayudi, A., 2018) (Triayudi, A., 2019) (Talakua, M.A., 2023). This action can help reduce the risk of dropping out of school. In addition, this predictive effort can also compare the results of a student's achievement at each level so that the study journey will be very controlled to achieve the best results. In general, making predictions based on in-depth observations is an essential task because it relates to decisions that will be taken in the future, considering that this educational environment re-quires a lot of methods and techniques that need to be implemented in order to develop this field of education towards the best results.

Today, technological developments, such as the internet and software, move quickly and are increasingly advanced. The rapid development of technology also affects the flexibility of the academic world, given the new context surrounding the educational process that involves technology as a medium for storing large amounts of information. Therefore, this study focuses on efforts to implement data mining techniques on data related to education and its development steps, considering that many things can be done related to re-search in this field. In addition, Educational Data Mining, commonly known as EDM, has often appeared in similar studies in recent years. So, the Educational Data Mining technique provides promising prospects in connection with efforts to advance and develop education in line with technological advances (Sharma, S., & Shiwani, S. 2014). In practice, the techniques and features available in the Educational Data Mining method make it possible to find hidden knowledge or information for smooth prediction or the best decision-making in an educational institution (Triayudi, A., 2018) (Miller, L.D., 2015) (Lailiyah, S., 2019).

This paper presents a data mining process framework, which is alleged to provide effective results by referring to predictive models. This data mining process framework aims to develop a predictive model that can provide an accurate guidance model according to the character of each student, as well as detect factors that influence the fluency of the student's studies (Nowakowski et al., 2021). In connection with this framework, the first thing that needs to be done is to collect various datasets related to academic information from relevant agencies. An investigation or classification is carried out based on the suitability of expertise. Later, it can provide choices or recommendations for appropriate and accurate guidance based on student performance analysis results resulting from the process (Sreenivasulu, G. 2024).

The research gap of this study refers to previous studies (Triayudi, A., 2018) (Triayudi, A., 2019), which discussed Stu-dent Performance using Data Mining. Previous research discussed mapping student behaviour and did not discuss predicting student performance. Therefore, this study discuss-es predictions about Student Performance using the Educational Data Mining Framework. The main contribution of this research is a new framework that can be used for recommendations for the best decision-making regarding academic guidance for student learning processes.

The framework related to the proposals in this study makes it possible to provide predictions regarding the results of the analysis of datasets related to student performance, which are then processed using algorithms in EDM (Ahmed, S.T., 2020) (Dutt, A., 2017) (Reddy, B.R., 2019). The first step in this research process is to collect related student data from various classes in the Information Systems and Informatics study program. The dataset contains information regarding 300 students from the 2017-2019 and 2018-2020 classes. This dataset has much personal information related to each student profile, such as student background, social and economic conditions, learning behaviour, learning motivation, and talents and interests of the students. From the available datasets, predictions refer to data mining algorithms, where the analysis process implements the methods contained in various algorithms such as Decision Trees, Neural Networks, and Naïve Bayes. In the end, an aca-demic guidance model is obtained from the analysis process that can meet the needs of the relevant agencies, both for teachers and students.

This research is structured as follows: Section 2 presents an overview of Educational Data Mining, or EDM, with various classification algorithms implemented. In section 3, an architectural framework is presented, which contains the methodology used in making predictions related to student profiles and performance. The results of the analysis obtained are presented in section 4. Conclusions and plans for future development are presented in section 5 of this research paper.

2 Materials and Methods

Educational Data Mining

As shown in Figure 1 below, Educational Data Mining visualizes the intersection of computer science, statistics, and education. From this intersection, a sub-area is formed, which is related to EDM, such as data mining, computer-based education, learning analytics, and machine learning (Ahmed, S.T., 2020) (Ratnapala, I.P., 2014) (Pasarate, S., 2018).

Each method contained in data mining has different capabilities and features depending on the data modelling needs. Generally, data modelling based on the implementation of data mining techniques and methods is usually carried out on descriptive models and predictive models (Varshavardhini, S., 2023). In the descriptive model, a dataset's hidden knowledge or information will go through various stages (grouping, association rules, classification to find sequences and others). In contrast, the predictive model will extract new knowledge or information a smooth data analysis process. The results obtained from the data modelling will be used to analyse student performance and predict how likely it is to drop out of school by deriving the predicted single variable and extracting the model from the available features (Antonenko, P.D., 2012) (HajKacem, M.A.B., 2017) (Ji, J., 2015).

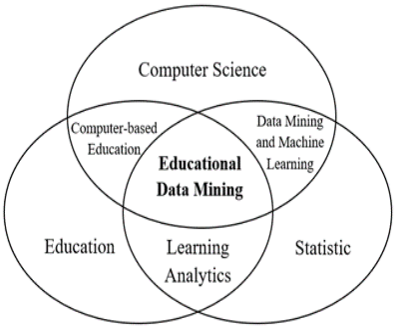

Figure 1: Educational Data Mining Related Works

The research that has been done previously represents a framework that can improve the admissions process and steps in maintaining student performance during the study period. In this case, it can be concluded that implementing the techniques and methods contained in data mining helps predict the success of work programs in the world of education and helps represent students' behaviour and learning styles during the study period. Then, other research suggests that the application of EDM can allegedly improve student performance, both for undergraduate and postgraduate programs, where the results obtained can help overcome students' academic deficiencies so that learning activities can run smoothly. In addition, several studies also describe the effect of implementing data mining techniques and methods on data related to students in universities, where the implementation of this Educational Data Mining technique provides satisfactory results in increasing the effectiveness of performance in an educational institution. Through these previous studies, the application of the techniques that are present in data mining shows significant results related to grouping efforts, implementation of decision tree algorithms and association techniques, ordering process levels in the education system, class selection, improving learning cycles, and student performance in an educational system, such as in online learning classes. This research was conducted by collecting 17,836 log servers' student data that contained 94 student learning behaviours in the Programming class at a private university in Jakarta. The data that has been processed is then carried out in the analysis stage that applies EDM techniques to find hidden patterns and information related to the description of student behaviour so that later conclusions can be drawn regarding predictions of learning outcomes that can be considered in conducting academic guidance.

Data Modelling Algorithm

This study implements techniques and methods from data mining algorithms, including Decision Trees, Neural Networks, and Naïve Bayes, as a classification method commonly used in finding predictive data modelling related to student academic performance.

- **Decision Tree**

A Decision Tree is a data model presented in the form of a decision tree, which is generally the most frequently used (Bienkowski, M., 2012) (Lakshmi, K., 2017) (Wijayanto, A., 2019) (Rosda, R., 2023) (Yusupa, A., 2023). In this form of data modelling, a fast and effective method is available to classify and visualize the data extraction results, where association rules are presented as a decision tree. Concerning the Decision Tree data modelling, the available algorithms include C4.5, ID3 (Iterative Dichotomiser 3), CHAID (Chi-squared Automatic Interaction Detector), and CART (Classification and Regression Tree), which are the basis of all these algorithms to partition the attribute space capable of classifying a data until all conditions have been met.

Various previous studies have been proposed to make predictions to identify association rules that many people can easily understand for the smooth development of knowledge in the field of education. Several available studies state that external factors significantly influence a student's loyalty to continuing his studies at a university level, which is then recorded data related to the profile and behaviour of these students are stored in a database of related educational institutions. In addition, rankings based on the results of each student's achievement are also carried out based on the implemented decision tree algorithm, wherein this case, an application that implements an algorithm such as C4.5 is proposed to reduce prediction rules so that it can assist in making decisions regarding classes or courses and what kind of study plan is right for each student.

The predictions generated later can also be used to consider developing student potential for the smooth running of students' academic journeys with satisfactory achievements. In this regard, decision trees can also be used to make predictions regarding student performance in terms of providing recommendations for the right time for learning with the e-learning system. Decision tree-based data modelling is also commonly used to predict the risk of dropping out of school, even from the first year a student starts the study period, by implementing the CART algorithm.

- **Neural Network**

Neural Network is a data modelling technique based on the Neural Network Model (NNM) consisting of nodes with interconnections (Baker, R.S., 2014) (Hong, J., 2018) (Nirmal, K.R., 2019). This data model contains the nodes and consists of a layer related to all nodes at the following level. Listed below is an illustration of a multilayer perceptron architecture that demonstrates the architectural structure of an artificial neural network (ARN). The Network has three layers of neurons, including an input layer, an output layer, and a hidden layer. The input layer functions to receive input vectors representing all information from the data to be executed according to the available attributes. Then, the output layer represents the classification class based on the analysed dataset. On the other hand, the hidden layer is a layer that serves to re-code related data input that has been done. From these three layers, conclusions can be drawn regarding the results of the classification process from an educational dataset.

In some works of literature, neuronal network architecture is known to be used in predicting student academic performance. One of the previous studies stated that predicting student achievement in a class is known to implement Neural Network-based data modelling. The study involved 200 students enrolled in a methodology and programming class implementing the NNM algorithm. Other research also suggests that artificial neural networks can help smooth the prediction process, which, in this case, involves as many as 165 students with varied learning profiles and behaviours. In addition, the multi-layer perceptron topology is also known to predict a student's possible performance and performance in conducting studies in a class or while in a preparatory class before starting education at the university level.

- **Naïve Bayes**

The Bayesian Naive model (BNM) is a method for classifying data that complies with Bayes' theorem. It is referred to as naive because it simplifies a problem based on two critical assumptions, namely the assumption of conditionally independent prognostic at-tributes and the assumption that there are no hidden attributes that can affect the currently executed prediction process (Baker, R.S., 2014) (Kacem, M.A.B.H., 2015) (Das, M.S., 2019). Therefore, the classification process with this algorithm is carried out by representing an approach that allows finding probabilistic and predictive knowledge regarding students' performance.

The formulation related to Bayesian classification based on posterior probability can be done by referring to the fact that the predictor value of X in a particular class C does not depend on the value of other predictors. Where the determination of the value of X to class C is carried out in such a way that P(C/X) reaches the maximum value:

$$P\left(\frac{C}{X}\right) = P\left(\frac{X}{C}\right) P(C)/P(X) \quad (1)$$

where,

P(C/X) is the posterior probability of class C given the predictor X.

P(C) represents the prior probability of belonging to class C

The probability of predictor X given to class C is denoted by P(X/C).

P(X) is the probability of the predictor X.

Many studies propose applying algorithms in BNM to make predictions related to student performance in units of education level (Baker, R.S., 2016) (Lee, J.E., 2016) (Ji, J., 2013). Data modelling with the Bayesian classification method is usually implemented by dividing the students into various groups before finally carrying out the analysis process to detect the tendency of students to need special attention or guidance from the teacher. By studying students' performance from various levels of education, data classification models such as Naïve Bayes, Neural Networks, C4.5 decision trees, and other algorithms, information is obtained that the data classification process ends in 2 classes: Pass and Fail. The results of the classification process are also based on information such as parent's education level, gender, and many other factors that influence the final prediction results obtained.

3 Results

The framework designed in this research involves a preparatory class for college en-trance that provides a multidisciplinary program to develop potential, talents, and tech-nical and theoretical skills, thus enabling a student to enter the best-desired university, commonly known as the CPGE program. The scientific and technological poles and the economic and commercial poles comprise the CPGE's primary areas. Consequently, the scientific pole comprises three scientific disciplines: mathematics and Programming (MP), Programming and engineering (PSI), and technology and engineering science (TES) (TSI).

After attending preparatory classes for four years of study, students can then take the following exams:

a. Common National Competition (CMC) is a selection test or Competition to enter the best company.
b. Competition to continue to postgraduate

That way, the selection test or Competition will provide students with a choice of where to continue their studies according to the student's profile based on the predictions generated. The prediction was carried out by involving the analysis results related to stu-dent achievement, which were then classified into four classes: GA, ADO, MA, and NA.

In order to build a framework that can provide appropriate recommendations or choices regarding the prediction of student performance on CPGE, as well as to support the implementation of accurate academic guidance, the data mining process is carried out with the following steps: data collection, pre-processing, modeling, and data interpretation. Figure 2 presents an overview of the methodology applied in this study.

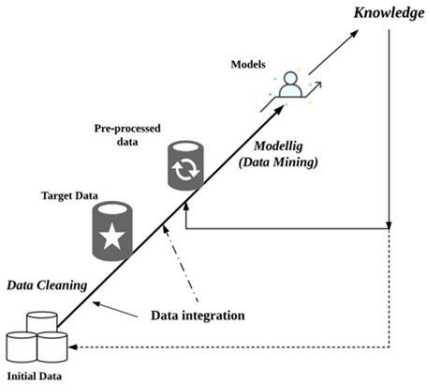

Figure 2: The Methodology Applied in this Research is to Develop a Smart Academic Framework

The visualization of student profiles is carried out through the information that can be used to access preparatory classes and contains data that supports students in making decisions regarding their study plans and academic journey. The description of the student profile is then formulated as follows:

$$M = N_1 + (N_2 - 10) + \frac{170\ x\ N_3}{18} + \frac{Ten\ x\ N_4}{21} \qquad (2)$$

Where:

N1 = 0 if the student repeats the study period in the first 1-2 years.
N1 = 5 if the student repeats the undergraduate education study period,
N1 = ten otherwise.
N2 M1þ2M2 3, where M1 is the average value of students in the initial year of the undergraduate program, and M2 is the average value of students' final grades in the final year of the undergraduate program.
N3 is calculated and formulated through the following equation:

N3 = 4M+3Phy+0.5 LV2+(1Fr+0.5Ar)9 (3)

In the Bachelor class program, M, Phy, LV2, Fr, and Ar are students' ability test scores related to Mathematics, Programming, English, and Indonesian. N4 is the criteria for undergraduate students who have entered the second year of study (from 0 to 21).

Final Result

The visualization of student profiles is carried out through the information that can be used to access preparatory classes and contains data that supports students in making decisions regarding their study plans and academic journey. The description of the student profile is then formulated as follows:

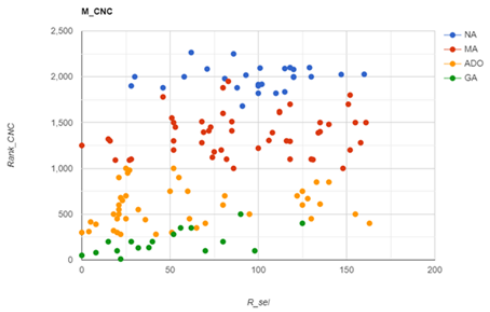

Figure 3: Classification of CPGE Students' based on the Final Result

As represented in Figure 3, related to the CPGE curriculum, it is known that in the student selection criteria with the final results, a correlation indicates the ranking of the analysed dataset. A similar analysis carried out on 300 students makes the results have a low score with an average rating. This phenomenon shows that the formula used at this selection stage does not provide significant results or represent the necessary conditions before concluding effective decision-making to lead to the suggested guidance model. In some conditions, some CPGE students are better classified with the R_sel criteria during the selection process to maintain performance and performance in CNC, and vice versa. Therefore, the data modeling implemented at this stage of data processing must be done based on the characteristics and other more precise models so that the predictions obtained will give good performance results.

Three types of information describe the profile of CPGE students. The first information contains data on selected students before enrolling in classes using the CPGE curriculum, which are then presented as the results of the selection formula. The second information contains the grades obtained by students during the first year of the preparatory class. The data includes the average score of academic tests that are usually conducted quarterly. The testing material is related to Mathematics, Programming, Engineering Science, Computer Science, Indonesian Language, and English. Then, the third information contains student scores that have been transformed into rankings according to the available preparation classes.

Questionnaires have effectively completed representations related to student profile data to complete missing data, such as students' motivation to choose a study plan, parents' level of study, and the number of hours worked outside of class), where later this collected data will help in increasing accuracy—predictive models. The variables used at this stage are shown in Table 1.

In order to effectively describe the student profile, these data were supplemented from the questionnaire to supplement the missing data (student motivation for the CPGE Choice, level of parental study, number of hours worked outside of class) and to improve the predictions of the model. The introduced variables are described in table 1.

Table 1: Questionnaire Data

Variable	Description
Gender	Gender of the students: Female (F) or Male (M).
MOT	Motivation for entering CPGE (based on parents (pa), friends (am), and personal motivation.
MIT	Average of work at home.
NEP	Parents' education level.
R_s	Selection rank that allowed access to the preparatory classes.
RM11, RM12, RM13	Student's rankings in the first year of Mathematics class.
RP11, RP12, RP13	Student's rankings in the first year of Programming class.
RSI11, RSI12, RSI13	Student's rankings in the first year of Engineering Sciences class.
RINF11, RINF12, RINF13	Student's rankings in the first year of Computer Sciences class.
RE11, RE12, RE13	Student's rankings in the first year of English class.
RB11, RB12, RB13	Student's rankings in the first year of Bahasa class.
M_C	The student's score is based on GA, ADO, MA, and NA.

Proposes Framework in this research can be seen in figure 4.

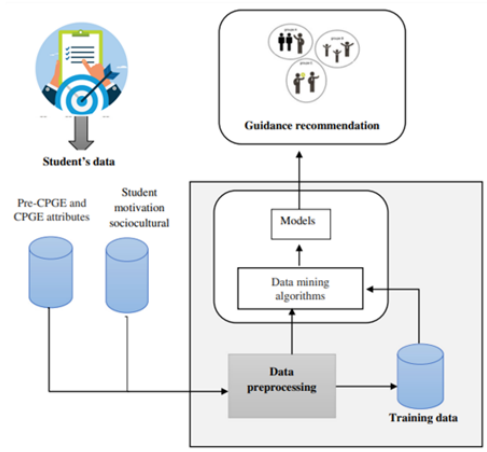

Figure 4: The Proposed Framework is Applied in this Research (Mimis, M., 2019)

The illustration related to the framework proposed in this study is represented in Figure 4, where this activity begins with collecting student data. Then, from the dataset that has been col-lected, a data mining process is implemented based on the Decision Tree, Naïve Bayes, and Neural Networks algorithms. In the end, the evaluation to determine how effective a student's academic performance is will be found by cross-validating, which will later be analysed and interpreted in an easy-to-understand form. The resulting framework can undoubtedly be used as a basis and material for consideration before deciding what guidance methods and models are suitable for each student so that academic performance during the study year will be controlled and improved. In addition, the results obtained can also be used to apply adaptive learning so that overall aca-demic performance will lead to the best achievement.

4 Discussion

This research implements three classification algorithms to predict students' aca-demic performance, including the academic rank of an available preparatory class. The algorithms applied are Decision Tree C4.5, Naïve Bayes Classification, and Neural Network. All data extraction processes are carried out using the Rapid Miner application.

From the CPGE database sample that has been carried out, data modeling validation is carried out into two groups, including Mathematics, Programming, and Engineering Sciences, which lasts for 2017-2019 and 2018-2020. The sample data comprises four classes (GA, ADO, MA, NA).

Every possibility from every available feature and information is collected for later analysis. The ranking of each influence of academic performance on the success of stu-dent's studies will be determined based on statistical order. The acquisition of information (IG) indicates how well an attribute can distinguish between data categories based on the notion of the chosen target. In information theory, often known as entropy, the determination of the idea of the target is a frequent application.

Entropy is defined as follows H(s)= H\left(s\right)=\ \sum{x\in d p\left(x\right)\log{2\left(p\left(x\right)\right)}}, where P(x) is the probability of class x chosen randomly from the set of S. Obtaining information IG(A) is the entropy reduction obtained by studying variable A.

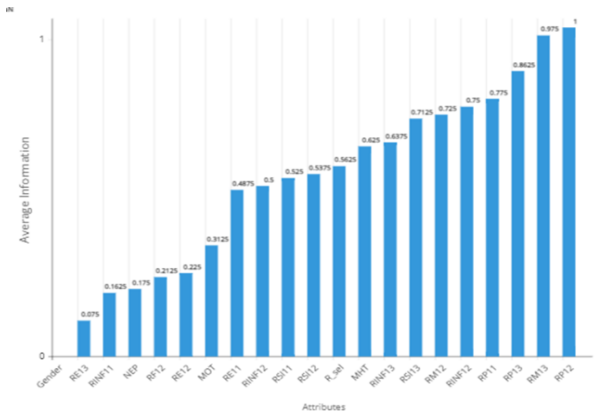

Figure 5: Gain and Variance on Each Feature Implemented on the Dataset

According to the data presented in Figure 5, it can be seen that the average value of information gain and Variance of each feature implemented in this study, where the more significant the information gain that affects the data modelling process, the greater the in-fluence of these attributes on the whole process. In this case, features, or attributes such as gender and language do not significantly influence. They have lower weights than other variables related to acquiring predictive results. Clearly, it is shown that the students' rankings related to the fields of Mathematics, Programming, Engineering Science, and Computer Science proved to have higher levels when compared to their social attributes.

Of the three data mining approaches applied to the datasets in this study, a comparison was done by deleting some variables that had minimal impact on the prediction findings. A 10-fold cross-validation procedure is also employed to validate the results. In this instance, the dataset is separated into ten subsets before the holdout approach is ap-plied, a process that is repeated ten times. Each of the ten subgroups serves as a test set, while the remaining nine subsets comprise the training set. Prediction results are obtained by sorting the classification level from the average value of the ten experiments. Regarding the prediction results related to the data modelling, two features are implemented: Accu-racy and Kappa Cohen.

In order to get the best information and results related to the influence of input variables on the final prediction ranking in this study, a comparison is made on the performance of data modeling using attributes in pre-CPGE, such as Gender, R_sel, MOT, MHT, NEP, and other related attributes. As shown in Table 2 below, the classification level to-wards acquiring predictive results increases sharply when students' scores in the first three quarters of the study period are included. This phenomenon shows that all the at-tributes a student possesses in the first year of study play a significant role in predicting final results and fluency to achieve good academic performance. In this case, the acquisition of classification results with the Decision Tree C4.5 algorithm implementation is 45.28% using only the pre-CPGE at-tribute. However, it increases sharply to 57.43% when new variables, such as student scores, are added in the first three-quarters of the study period.

Table 2: Method and Algorithm Applied

Method and algorithm applied.	Attributes of Pre-CPGE		All the attributes	
	Accuracy	Kappa	Accuracy	Kappa
Decision Tree (C4.5)	45.28%	0.257	57.43%	0.476
Naïve Bayes	51.60%	0.321	61.18%	0.524
Neural Networks (21 neurons of an input layer, one hidden layer, and four neurons of an output layer)	54.88%	0.458	59.35%	0.493

Then, as shown in Table 2, it can be stated that the accuracy of predictions generated by the Naïve Bayes algorithm and Neural Networks are known to be more critical than just interpreting the model. However, the model generated by the Decision Tree algorithm is easier to understand.

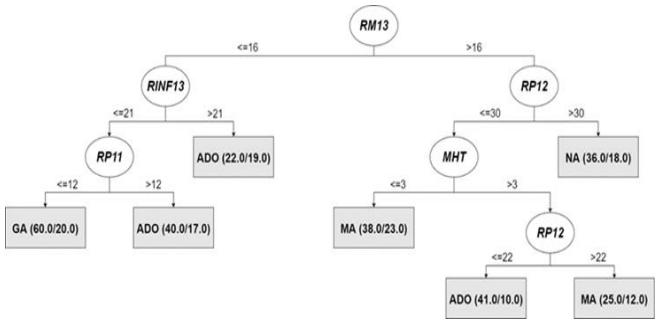

Figure 6: Visualization of Decision Tree Algorithm Applied

Figure 6 depicts the Decision Tree model generated by applying the C4.5 algorithm with the following parameters: Minimum number of instances: 8.0. Pruning confidence threshold: c = 0.25. It can be utilized and comprehended by a teacher when providing academic counseling for the effectiveness and efficiency of the learning process.

True:	NA	MA	ADO	GA
NA:	18	11	4	0
MA:	13	26	34	5
ADO:	8	35	58	28
GA:	0	6	19	29

True:	NA	MA	ADO	GA
NA:	13	13	7	1
MA:	17	35	19	1
ADO:	9	28	68	23
GA:	1	3	23	33

Figure 7: Confusion, Matrix of the Final Result, based on Decision Tree C4.5 Algorithm Applied

Figure 7a presents a confusion matrix regarding the final prediction results with implementing the Decision Tree C4.5 algorithm on pre-CPGE attributes. In this Matrix can be concluded that the closest class generally causes errors in a prediction. It can be seen that the errors that occurred in the predictions made occurred in the MA, ADO, and GA classes. For example, as many as 35 students from the ADO profile were identified as belonging to the profile classification following the MA. Conversely, as many as 34 students from the MA profile were included in the profile classification according to the ADO because the two variables were close to each other. On the other hand, there were no errors related to the prediction results on two variables, such as NA and GA classes.

Figure 7b presents a confusion matrix related to the system obtained if there is a condition where all student scores in the first year are entered. In this case, we can see the fix for the error mentioned earlier. For example, the prediction results based on the represen-ation in Figure 7b now have 28 students (instead of 35) and 19 students (instead of 34). This phenomenon shows that attributes related to a student's performance during the first year have a significant role in acquiring predictive academic performance results. Furthermore, the experiments can be expanded with more attributes to produce a more accu-rate final prediction value. This experiment can be carried out by implementing other data mining algorithms to obtain a broader approach and more significant and accurate results.

In this discussion section, the research tries to explore research results using the Agglomerative Hierarchical Clustering (AHC) algorithm by dividing the dataset into 5 different time sections of student activities, namely:

1. Dataset in submission
2. Dataset in course modul

3. Dataset in discussion
4. Dataset in course view
5. Dataset in observation

The aim of processing data using this clustering algorithm is to divide student characteristics according to a psychological perspective, where psychological aspects are an important aspect in improving student performance.

Clustering process on dataset in submission

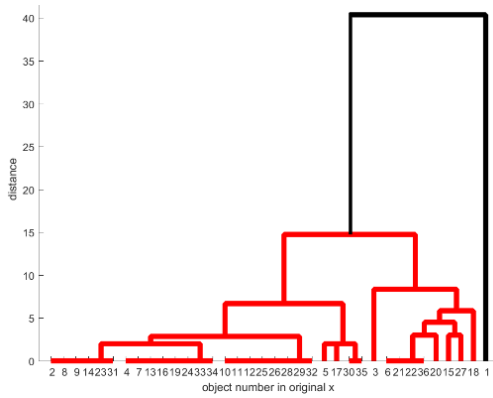

Figure 8: Dendrogram Submission Dataset

Evaluation

- The number of clusters obtained by two clusters, according to the optimal cluster analysis and the silhouette coefficient value is 0.91.
- The cophenetic correlation coefficient value is quite high at 0.96
- Of the 36 objects, only 10 series had inconsistent relationships, with an average value of 0.25

Clustering process on dataset in course module (forum)

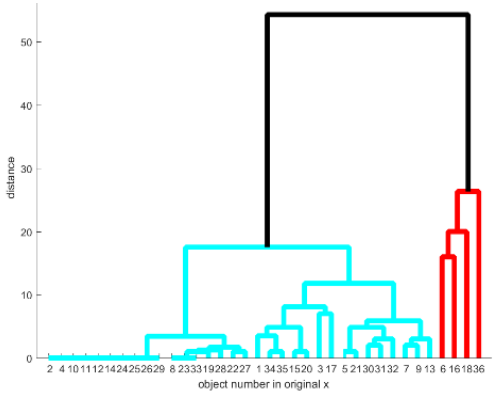

Figure 9: Dendrogram Dataset Course Module

Evaluation

- The number of clusters obtained by two clusters, according to the optimal cluster analysis and the silhouette coefficient value is 0.915.
- The cophenetic correlation coefficient value is quite high at 0.92
- Of the 36 objects, only 16 series had inconsistent relationships, with an average value of 0.39

Clustering process in dataset discussion (forum)

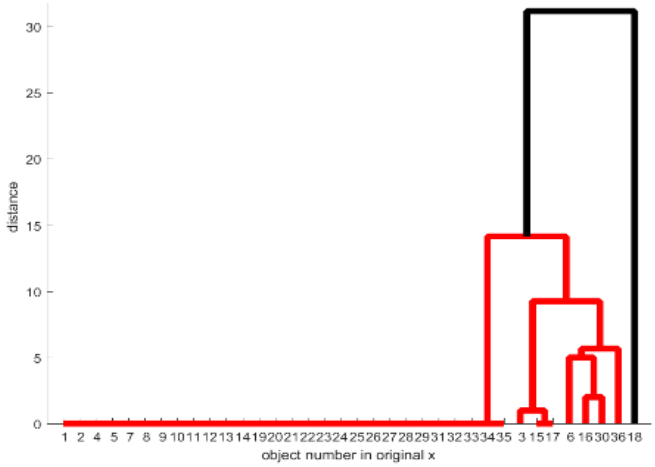

Figure 10: Dendrogram Dataset Discussion (Forum)

Evaluation

- The number of clusters obtained by two clusters, according to the optimal cluster analysis and the silhouette coefficient value is 0.89.
- The cophenetic correlation coefficient value is quite high at 0.93
- Of the 36 objects, only 6 series have inconsistent relationships, with an average value of 0.13

Clustering process on course view dataset

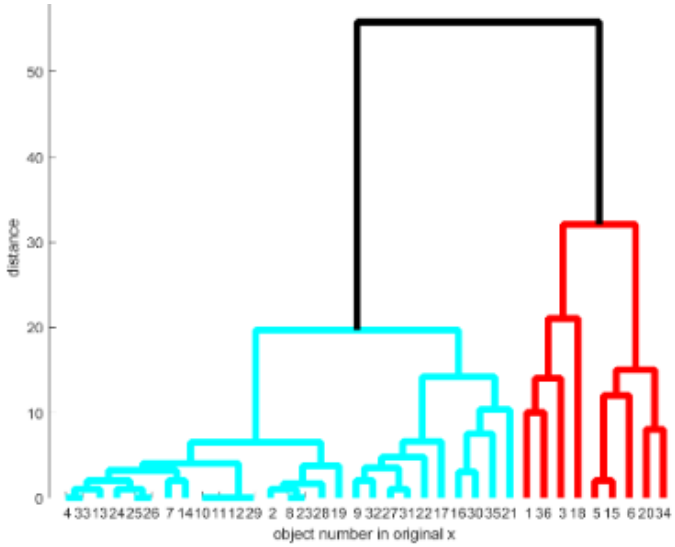

Figure 11: Dendrogram Dataset Course View

Evaluation

- The number of clusters obtained by two clusters, according to the optimal cluster analysis and the silhouette coefficient value is 0.845.
- The cophenetic correlation coefficient value is quite high at 0.87
- Of the 36 objects, there are 23 series that have inconsistent relationships, the most among other datasets with an average value of 0.54

Clustering process on the observe dataset

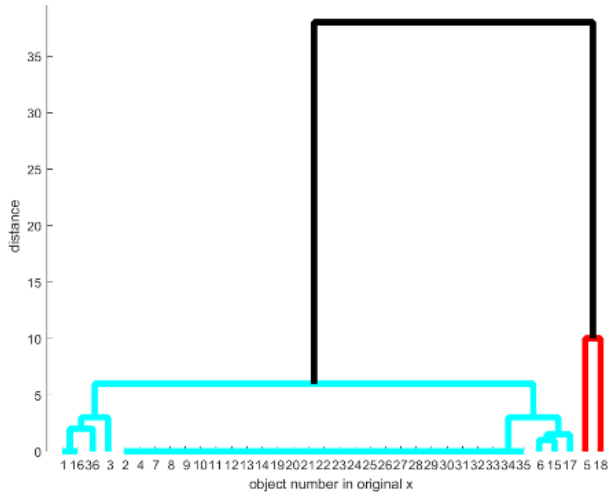

Figure 12: Dendrogram of the Observed Dataset

Evaluation

- The number of clusters obtained by two clusters, from the silhouette coefficient and the silhouette coefficient value is 0.99.
- The cophenetic correlation coefficient value is quite high at 0.99
- Of the 36 objects, only 6 series have inconsistent relationships, this too with an average value of 0.15

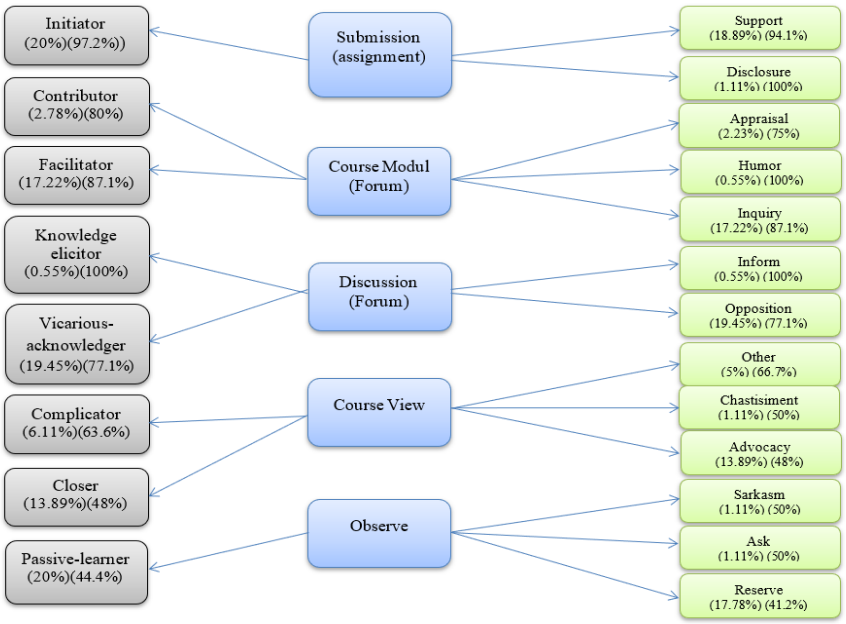

Figure 13: Model of Characteristics from the Psychological Side of Students

Explanation of Figure 13, namely the characteristics of the model from the psychological side of students, divided into 5 large datasets, namely the "submission" dataset which is mapped as the "initiator", "support", "disclosure" models where in the first dataset this is a collection of behavioral models and The best and most appropriate interpersonality for students to apply or follow can be seen from the very high graduation rate.

The second dataset, namely "course module" which is mapped as a "contributor", "facilitator", "appraisal", "humor", "inquiry" model is a collection of behavioral and interpersonal models that are ranked second, these models are most likely to be able to applied to various types of students, because it is not too highly oriented and can be achieved with the right learning strategies.

The third dataset, namely "discussion" which is mapped as a "knowledge elicitor", "vicarious acknowledger", "inform", "opposition" model is a collection of behavioral and interpersonal models that teachers must be aware of, teachers must be able to direct their students with appropriate strategies and true because of the variety of models in this type.

The fourth dataset, namely "course view" which is mapped as a "complicator", "closer", "other", "chastisiment", "advocacy" model is a collection of models of behavior and interpersonality that will lead to negative, models in this type are carried by bad traits In their relationships, teachers must pay extra attention to students with this model.

Finally, the fifth dataset, namely "observe" which is mapped as a model of "passive learner", "sarcasm", "ask", "reserve" is a collection of models that are the worst in terms of the type of student behavior and interpersonality models, teachers must work hard to change the type of students like This type of pass rate is also very poor.

5 Conclusion

This study focuses on the proposed framework of recommendations to make the best decisions regarding the most appropriate academic guidance for the smooth learning process and student academic performance. This research was conducted based on pre-dictions of student performance to create appropriate educational techniques and not only based on assumptions, as is still often done, including by teachers to their students. The framework proposed in this study helps to provide predictions of performance and what kind of guidance is appropriate for the smooth running of the study, where the achievement of the performance results depends on the variables before and after the preparatory class and the type of classification used. The rules generated by the Decision Tree provide crucial information and knowledge for implementing the counselling and guidance process for students and teachers based on the best decisions regarding the predicted results. A comparative study needs to be done by implementing different decision-making techniques and structures, which requires much exploration in the future. Therefore, the generalization of studies to a set of big data needs to be done by applying data mining algorithms and other machine learning so that this can help improve prediction accuracy on student performance.

This research also discusses the psychological characteristics of students using an agglomerative hierarchical clustering algorithm, by dividing them into 5 large datasets and then grouping them into several psychological aspects according to the cluster where the students are located. This division and psychological analysis are important as a reference for stakeholders on how to form students with the best performance and best character.

For further research, datasets can be developed from various ethnicities or unique characteristics of students, such as hobbies, study environment, or place of residence. From an algorithm standpoint, it can be developed using the ensemble method and the expectation-maximization algorithm to obtain perfect data.

References

[1] Ahmed, S.T., Al-Hamdani, R., & Croock, M.S. (2020). Developed third iterative dichotomizer based on feature decisive values for educational data mining. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(1), 209-217.

[2] Akçapýnar, G., Altun, A., & Cosgun, E. (2014). Investigating students' interaction profile in an online learning environment with clustering. *In IEEE 14th International Conference on Advanced Learning Technologies*, 109-111.

[3] Antonenko, P.D., Toy, S., & Niederhauser, D.S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, *60*, 383-398.

[4] Babiker, M., & Elmagzoub, A. (2015). For Effective Use of Multimedia in Education, Teachers Must Develop their Own Educational Multimedia Applications. *Turkish Online Journal of Educational Technology-TOJET*, *14*(4), 62-68.

[5] Baker, R.S. (2014). Educational data mining: An advance for intelligent systems in education. *IEEE Intelligent systems*, *29*(3), 78-82.

[6] Baker, R.S., & Inventado, P.S. (2014). *Educational Data Mining and Learning Analytics*. En JA Larusson & B. White (Eds.), Learning Analytics: From Research to Practice, 61-75.

[7] Baker, R.S., Martin, T., & Rossi, L.M. (2016). Educational data mining and learning analytics. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, 379-396.

[8] Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief. *Office of Educational Technology, US Department of Education*.

[9] Das, M.S., Govardhan, A., & Lakshmi, D.V. (2019). Classification of web services using data mining algorithms and improved learning model. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *17*(6), 3191-3202.

[10] Dewar, T., & Whittington, D. (2000). Online learners and their learning strategies. *Journal of Educational Computing Research*, *23*(4), 385-403.

[11] Dutt, A., Ismail, M.A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, *5*, 15991-16005.

[12] HajKacem, M.A.B., N'Cir, C.E.B., & Essoussi, N. (2017). KP-S: a spark-based design of the K-prototypes clustering for big data. *In IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, 557-563.

[13] Hong, J., Xiang, Y., Liu, Y., Liu, J., Li, R., Li, F., & Gou, J. (2018). Development of EV charging templates: an improved K-prototypes method. *IET Generation, Transmission & Distribution*, *12*(20), 4361-4367.

[14] Ji, J., Bai, T., Zhou, C., Ma, C., & Wang, Z. (2013). An improved k-prototypes clustering algorithm for mixed numeric and categorical data. *Neurocomputing*, *120*, 590-596.

[15] Ji, J., Pang, W., Zheng, Y., Wang, Z., Ma, Z., & Zhang, L. (2015). A novel cluster center initialization method for the k-prototypes algorithms using centrality and distance. *Applied Mathematics & Information Sciences*, *9*(6), 2933-2942.

[16] Kacem, M.A.B.H., N'cir, C.E.B., & Essoussi, N. (2015). MapReduce-based k-prototypes clustering method for big data. *In IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 1-7.

[17] Lailiyah, S., Yulsilviana, E., & Andrea, R. (2019). Clustering analysis of learning style on anggana high school student. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *17*(3), 1409-1416.

[18] Lakshmi, K., Visalakshi, N.K., & Shanthi, S. (2017). Cuckoo search based K-prototype clustering algorithm. *Asian Journal of Research in Social Sciences and Humanities*, *7*(2), 300-309.

[19] Lee, J.E., Recker, M., Bowers, A.J., & Yuan, M. (2016). Hierarchical Cluster Analysis Heatmaps and Pattern Analysis: An Approach for Visualizing Learning Management System Interaction Data. *In EDM*, 603-604.
[20] Livingston, K., & Condie, R. (2006). The impact of an online learning program on teaching and learning strategies. *Theory into Practice*, *45*(2), 150-158.
[21] Mathur, G., Nathani, N., Chauhan, A. S., Kushwah, S. V., & Quttainah, M. A. (2024). Students' Satisfaction and Learning: Assessment of Teaching-Learning Process in Knowledge Organization. *Indian Journal of Information Sources and Services, 14*(1), 1–8.
[22] Miller, L.D., Soh, L.K., Samal, A., Kupzyk, K., & Nugent, G. (2015). A Comparison of Educational Statistics and Data Mining Approaches to Identify Characteristics That Impact Online Learning. *Journal of Educational Data Mining, 7*(3), 117-150.
[23] Mimis, M., El Hajji, M., Es-saady, Y., Oueld Guejdi, A., Douzi, H., & Mammass, D. (2019). A framework for smart academic guidance using educational data mining. *Education and Information Technologies*, *24*, 1379-1393.
[24] Nirmal, K.R., & Satyanarayana, K. (2019). Redic k-prototype clustering algorithm for mixed data (numerical and categorical data). *International Journal of Recent Technology and Engineering*, *7*(6), 1-6.
[25] Nowakowski, P., Zórawski, P., Cabaj, K., & Mazurczyk, W. (2021). Detecting Network Covert Channels using Machine Learning, Data Mining and Hierarchical Organisation of Frequent Sets. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 12*(1), 20-43.
[26] Pasarate, S., & Shedge, R. (2018). Concept based document clustering using K prototype Algorithm. *In IEEE International Conference on Control, Power, Communication and Computing Technologies (ICCPCCT)*, 579-583.
[27] Ratnapala, I.P., Ragel, R.G., & Deegalla, S. (2014). Students behavioural analysis in an online learning environment using data mining. *In IEEE 7th International Conference on Information and Automation for Sustainability*, 1-7.
[28] Reddy, B.R. (2019). A Brief Analysis of the Key Technologies and Applications of Educational Data Mining on Online Learning Platform. *International Journal of Research, 8*(5), 2398-2402.
[29] Rosda, R., Wali, M., & Imilda, I. (2023). Evaluation of the Successful Implementation of the SIMARDI Using the Technology Acceptance Model (TAM). *SAGA: Journal of Technology and Information System*, *1*(2), 56-63.
[30] Shahzad, F., Saeed, A., Asim, G.A., Qureshi, F., Rehman, I.U., & Qureshi, S. (2021). Political connections and firm performance: Further evidence using a generalised quantile regression approach. *IIMB Management Review*, *33*(3), 205-213.
[31] Sharma, S., & Shiwani, S. (2014). Data mining based accuracy enhancement of ANN using Swarm intelligence. *International Journal of Communication and Computer Technologies (IJCCTS)*, *2*(1), 43-46.
[32] Sreenivasulu, G. (2024). A Hybrid Optical-Acoustic Modem Based on Mimo Ofdm for Reliable Data Transmission in Green Underwater Wireless Communication. *Journal of VLSI Circuits and Systems, 6*(1), 36-42.
[33] Talakua, M.A. (2023). The Role of Christian Religious Education Teachers in Shaping Student Character Through Peace Education and Conflicts Resolution Among Students. *Athena: Journal of Social, Culture and Society*, *1*(4), 257-261.
[34] Triayudi, A., & Fitri, I. (2018). ALG clustering to analyze the behavioural patterns of online learning students. *Journal of Theoretical & Applied Information Technology*, *96*(16), 5327-5337.
[35] Triayudi, A., & Fitri, I. (2019). A new agglomerative hierarchical clustering to model student activity in online learning. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *17*(3), 1226-1235.
[36] Varshavardhini, S., & Rajesh, A. (2023). An Efficient Feature Subset Selection with Fuzzy

Wavelet Neural Network for Data Mining in Big Data Environment. *Journal of Internet Services and Information Security (JISIS), 13*(2), 233-248.

[37] Wijayanto, A., Suprapto, Y.K., & Wulandari, D.P. (2019). Clustering on Multidimensional Poverty Data using PAM and K-prototypes Algorithm: Case Study: Jambi Province 2017. *In IEEE International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 210-215.

[38] Yusupa, A., Manullang, J., Marbun, N., & Ginting, S.B.F. (2023). Decision Support System for Determining the Best PAUD Teacher Using the MOORA Method. *SAGA: Journal of Technology and Information System*, *1*(2), 50-55.

Authors Biography

Dr. Agung Triayudi, S.Kom., M.Kom, Ph.D, was born in Jakarta, on Juni 19, 1986. As for my Educational History, S1 at the STMIK Tunas Bangsa graduated in 2010. S2 at Budi Luhur University Jakarta graduated in 2012. S3 Informatic & Communication Technology at Asia e University Malaysia. Currently I am serving as a Permanent Lecturer at National University Jakarta. My research study in the field of Data Mining, Machine Learning.

Rima Tamara Aldisa, S.Kom., M.Kom was born in Jakarta, on Januari 18, 1994. As for my Educational History, S1 at the National University Jakarta graduated in 2016. S2 at Budi Luhur University Jakarta graduated in 2019. On Going S3, Information System at Diponegoro University Semarang. Currently I am serving as a Permanent Lecturer at National University Jakarta. My research study in the field of Data Mining, Artificial Intelligence, Decision Support System.

S. Sumiati, ST., MM., Ph.D is a lecturer and researcher at the Department of Informatics, Serang Raya University, Indonesia. Strong educational background in the field of Artificial Intelligence, Data Mining. Currently, research is developing in the field of medicine combined with Artificial Intelligence.