Journal of Social Science and Business Studies
Volume 2, No. 3, pp. 240-250
E-ISSN: 2987-6079
http://gemapublisher.com/index.php/jssbs
Received: July 2024
Accepted: August 2024
Published: September 2024

Sustainability of Quality Management by Implementing Data Mining to Predict Academic Achievement

Mufidah Karimah*, Fingki Marwati

Universitas Pamulang

Correspondence Email: dosen02829@unpam.ac.id*

Abstract

This research aims to predict the academic achievement of Pamulang University students using Educational Data Mining (EDM) techniques. With the increasing number of students and the complexity of academic data, it is important to apply methods that can analyze and predict learning outcomes to improve learning strategies and academic support. This study collected data from various sources, including course grades, attendance, and participation in extracurricular activities. The collected data is then analyzed using EDM techniques such as decision trees, neural networks, and support vector machines to identify patterns and factors that contribute to student academic achievement. The results of the analysis show that factors such as attendance, involvement in campus activities, and previous test scores have a significant influence on academic achievement. This research provides valuable insights for the development of targeted interventions, with the aim of improving academic outcomes and facilitating more effective learning strategies at Pamulang University. These findings also offer contributions to further research in the field of EDM and its application in higher education contexts.

Keywords: prediction of academic achievement, educational data mining, pamulang university, decision tree technique, neural networks, support vector machines.

INTRODUCTION

Student academic achievement in higher education is one of the main indicators of educational success and individual development. In this digital era, educational institutions face major challenges in managing and utilizing abundant data to support academic and strategic decisions. Pamulang University, as one of the fastest growing higher education institutions in Indonesia, is no exception in facing this challenge. With an ever-increasing student population and greater data complexity, it is important for Pamulang University to apply innovative analytical approaches to understand and predict student academic performance. One approach that offers great potential in this regard is Educational Data Mining (EDM), a data analysis method designed specifically for educational contexts.

Academic achievement not only reflects the results of the learning process but also influences various aspects of student life, including career opportunities, learning motivation, and personal satisfaction. At the university level, academic achievement is often measured through course grades, GPA, and other academic achievements. However, to understand the factors that influence this achievement, a more in-depth analysis involving broader data is needed, such as attendance, involvement in extracurricular activities, and socio-economic conditions.

DOI: https://doi.org/10.61487/jssbs.v2i3.90

Pamulang University faces challenges in managing large and varied academic data. This data includes information about student grades, attendance, extracurricular activities, as well as other demographic data. Although this data has great potential for use in improving learning strategies and academic support, processing and interpreting this data requires sophisticated methods to gain accurate and useful insights. This is where Educational Data Mining plays an important role.

Educational Data Mining (EDM) is a research field that focuses on the use of data mining techniques to analyze educational data with the aim of understanding and improving the teaching and learning process. EDM uses a variety of data analysis techniques, including machine learning, statistics, and mathematical modeling, to unearth patterns and trends in academic data. By using EDM, educational researchers and practitioners can identify factors that influence academic achievement, predict learning outcomes, and design more effective interventions to support students.

Various techniques in EDM can be applied to analyze academic data. Some key techniques include:

a. Decision Trees: This technique is used to create decision models based on the features in the data. Decision trees can help identify the variables that most influence academic achievement.
b. Neural Networks: This model imitates the way the human brain works in processing information. Neural networks can capture complex patterns in data that may not be visible with other methods.
c. Support Vector Machines (SVM): This technique is used for classification and regression by finding the hyperplane that separates data classes by the largest margin.
d. Clustering: This technique groups similar data into more homogeneous groups, which can help in understanding segments of students with similar characteristics.

This research aims to apply the EDM technique to predict the academic achievement of Pamulang University students. By analyzing existing academic data, this research will seek to identify key factors that influence student achievement, as well as develop prediction models that can provide useful insights for improving learning strategies and academic support.

METHOD

Research Design

This research uses a quantitative approach with a case study design to explore and apply Educational Data Mining (EDM) techniques in predicting the academic achievement of Pamulang University students. This study focuses on analyzing academic data that has been collected to identify patterns and factors that influence academic achievement.

DOI: https://doi.org/10.61487/jssbs.v2i3.90

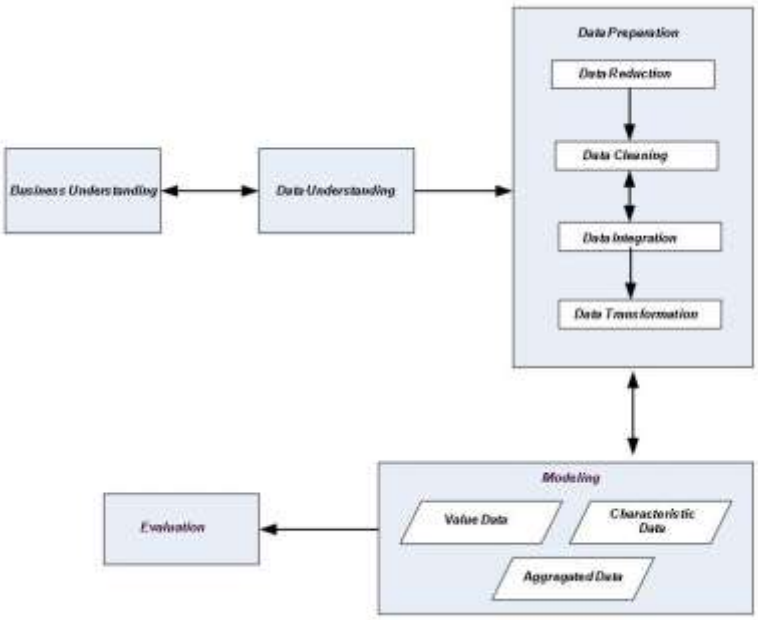

Figure 1. Research Architecture

Sample and Population

The population of this study includes all Pamulang University students from various study programs. Samples will be taken randomly from active students in certain semesters, taking into account sufficient representation from various majors and academic years. Sample size will be determined based on statistical analysis needs and data availability.

Data Collection

Data to be collected includes:

a. Academic Grades: Data about course grades, GPA, and test scores.
b. Attendance: Data on student attendance in lectures and other academic activities.
c. Extracurricular Activities: Information regarding student participation in extracurricular activities and campus organizations.
d. Demographic Data: Information about students' socio-economic background, age, and gender.
e. This data will be obtained from the Pamulang University academic information system and other relevant administrative sources.

Data Processing Techniques

The data collected will go through several processing stages as follows:

a. Data Cleaning: Remove incomplete, duplicate, or irrelevant data.
b. Data Transformation: Transforming data in a format suitable for analysis, such as normalization and encoding categorical variables.
c. Data Splitting: Splitting data into training set and test set for model validation.

Data Analysis Techniques

Several EDM techniques will be applied to predict academic performance:

a. Decision Trees: Create a decision tree model to determine the factors that most influence academic achievement.

DOI: https://doi.org/10.61487/jssbs.v2i3.90

b. Neural Networks: Using neural networks to model non-linear relationships between academic variables and achievement.
c. Support Vector Machines (SVM): Apply SVMs to classification and regression, with a focus on creating optimal hyperplanes to separate data.
d. Clustering: Grouping students based on feature similarity to identify relevant patterns.

Model Evaluation

The developed model will be evaluated using evaluation metrics such as:

a. Accuracy: The percentage of correct predictions compared to actual data.
b. Precision and Recall: To assess the model's accuracy and ability to identify certain categories.
c. F1 Score: A combination of precision and recall to provide a more complete picture of model performance.
d. AUC-ROC Curve: To assess the model's ability to differentiate between different classes.

Analysis of Results

The results of the analysis model will be analyzed to determine the main factors that influence student academic achievement. The research will discuss the implications of the findings in the context of learning strategies and academic support. These findings will also be compared with related literature to identify similarities or differences.

Validation and Testing

To ensure the validity of the results, this research will carry out cross-validation and model testing on different data subsets. This process will help measure the consistency and reliability of the model in predicting academic achievement.

RESULT AND DISCUSSION

Modeling

There are three types of modeling carried out at this stage: Data modeling of student characteristics. Modeling student grade data. Modeling combined data of student characteristics and grades. In general, the classification carried out in each modeling uses cross-validation method because the number of data records is too small to be divided into training data and test data. Classification is carried out by using the C4.5 and CART algorithms. This modeling was carried out using training data described.

A. Data Modeling of Student Characteristics

In modeling data on student characteristics, there are two types of modeling carried out. The first modeling was carried out with classify all student characteristic data attributes combined with three types of class attributes using the cross-validation method. The second modeling is carried out on the main attributes of the data characteristics and attributes of the best class to find out how good the model is obtained from data from previous generations is able to predict the academic achievement of new students.

B. Modeling Using the Cross-Validation Method

In this modeling, classification of Basic Data and Training Kar is carried out (subsection III.4.5) using the ross-validation method. For each training data, classification is carried out using three different class attributes namely two manual class attributes (K2, K3) and two class attributes resulting from clustering value data (KlusterValue, KlusterGab).

DOI: https://doi.org/10.61487/jssbs.v2i3.90

Table 1. Classification Results of Student Characteristics Data (Cross validation)

Data	Basic Data (K2)		Basic Data (K3)		Basic Data (Value Cluster)		Basic Data (Gab Cluster)	
Algorithm	C4.5	CART	C4.5	CART	C4.5	CART	C4.5	CART
Accuracy	64.88%	67.56%	49.16%	49.50%	39.13%	39.78%	38.46%	36.78%
Recall	0.702	0.735	0.627	0.669	0.373	0.461	0.475	0.505
Precision	0.639	0.661	0.592	0.556	0.355	0.359	0.416	0.439
F-measure	0.669	0.696	0.609	0.608	0.364	0.403	0.443	0.469
Attributes	School	School	School	School	School	RTAUN	School	School
	Choice	Options	Choice	Options	RTSTTB	Options	Jekel	RTSTTB
	Jekel	RTSTTB	RTUAN		Father's Education	RTSTTB	Father's Education	Choice
	Father's PKJ		Father's PKJ		Jekel		Mother's PKJ	Mother's PKJ
	Post Child		Mother's PKJ		Choice		RTUAN	Father's Education
	Mother's Education		Father's Education		Region		Post Child	
	Age				Mother's Education			
Data	**TrainingKar (K2)**		**TrainingKar (K3)**		**TrainingKar (Value Cluster)**		**TrainingKar (Gab Cluster)**	
Algorithm	C4.5	CART	C4.5	CART	C4.5	CART	C4.5	CART
Accuracy	67.22%	61.53%	44.81%	47.49%	38.46%	43.81%	36.45%	42.14%
Recall	0.742	0.748	0.661	0.627	0.611	0.685	0.515	0.657
Precision	0.655	0.595	0.538	0.561	0.433	0.477	0.425	0.520
F-measure	0.696	0.663	0.593	0.592	0.514	0.563	0.466	0.580
Attributes	School	School	School	School	School	Options	School	School
	Choice	Choice	Options	Options	Options		Jekel	RTSTTB
	Jekel	Father's PKJ	Mother's PKJ		Father's Education		Father's Education	Options
	Father's PKJ		Father's Education		Jekel		Father's PKJ	Mother's PKJ
	Post Child		Jekel		Mother's PKJ		Mother's PKJ	Father's Educat

DOI: https://doi.org/10.61487/jssbs.v2i3.90

								ion
	Mother's Education		RTUAN		Regional		Child's Post	Region
	RTUAN		RTSTTB		Mother's Education		RTSTTB	RTUAN

Table 1. displays a comparison of the classification results carried out on the data student characteristics. In this table it can be seen that the algorithm CART provides the best accuracy (67.56%) for Basic Data, meanwhile The C4.5 algorithm gave the best results (67.22%) on TrainingKar. This matter This occurs because the CART algorithm with the binary tree concept is more suitable to use for numeric data, where in Basic Data, most of the values The characteristic data attributes still use numerical form. Algorithm C4.5 build a tree with the number of branches per node corresponding to the number of values the node. Because the values of TrainingKar attributes are grouped into several groups (>2), the C4.5 algorithm is more suitable to use and provides better results than the CART algorithm.

C. Modeling for Predicting New Student Academic Achievement

This second characteristic data modeling was carried out using characteristic data new (TrainingKar) which consists of the attributes School, Choice, Jekel, RTUAN, RTSTTB which are directly related to personal background and student academics. The K2 class attribute is used as the class attribute. For In this second classification, TrainingKar data is divided into training data and test data with composition: Training data consists of data on the characteristics of students from the class 2006-2009 with number of records = 247. Test data uses data on the characteristics of students from the class of 2010 with number of records = 52.

There are two methods used in this modeling, first modeling carried out using classification alone. In the second method, before classification is carried out, the data is clustered to group the data based on the same attribute values.

Table 2. TrainingKar2 Classification Results Using the Open Method Classification

Algorithm	**C4.5**		**CART**	
Data	Training	Test	Training	Test
Accuracy	72.47%	78.85%	68.82%	76.92%
Recall	0.775	0.773	0.775	0.818
Precision	0.719	0.739	0.676	0.692
F-measure	0.746	0.756	0.722	0.750
Attributes	School		School	
	Options		Options	
	Jekel		Jekel	
	RTUAN			
	RTSTTB			

DOI: https://doi.org/10.61487/jssbs.v2i3.90

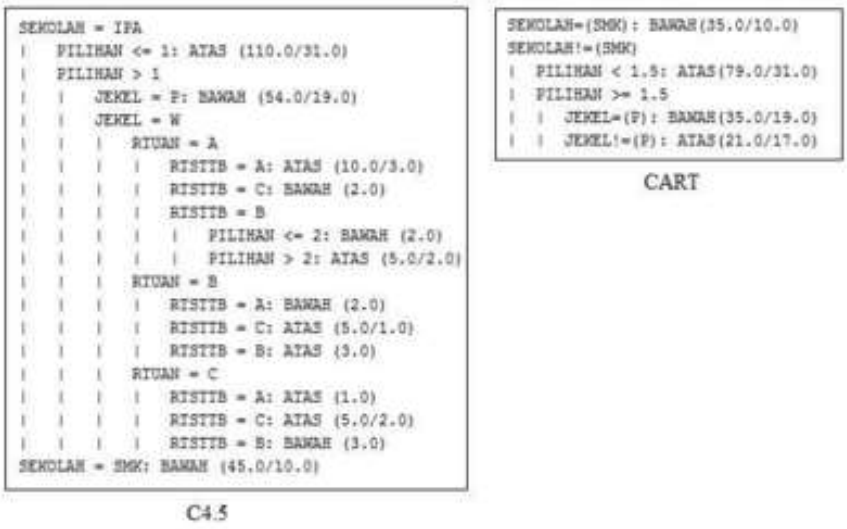

Figure 2. The Cart

TrainingKar2 classification results using the C4.5 and CART algorithms shown in Table IV-2. The best model is produced during training data is classified using the C4.5 algorithm with values of accuracy, recall, precision, and f-measure above the results of the CART algorithm. Use of test data on the model produced by the C4.5 algorithm provides very good accuracy, namely 78.85%. As many as 11 of the 52 records in the test data had errors predictions. From the attributes used in the decision tree (Figure 2.), the CART decision tree uses only the attributes School, Choice, and Just Jekel, in contrast to the C4.5 decision tree which uses all five training data attributes.

D. Modeling Student Grade Data

In this modeling, classification is carried out to obtain a comparison model the results of the characteristic data modeling carried out previously. Classification process carried out using the cross-validation method. Classification is carried out on three types of grouping of course attribute values in student value data which is represented by Basic Data, Training1, and Training2. Attribute The class used in this classification is K2 which is an attribute best class results from classification of student characteristic data. Table 3.3 displays the classification results for each training data. From Table 3 it is known that the classification uses the C4.5 algorithm in Training2 where student scores are grouped into 4 groups (Best, Good, Pass, and Fail) is able to provide accuracy, recall, precision, and a better f-measure than other classifications. From the whole Classification of value data carried out can be seen that Calculus courses, AlPro, Physics, and P.Aplikom are attributes that are at the level on the decision tree.

DOI: https://doi.org/10.61487/jssbs.v2i3.90

Table 3. Classification Results of Student Grade Data

Data	Basic Data (K2)		Training1 (K2)		Trainig2 (K2)	
Algorithm	C4.5	CART	C4.5	CART	C4.5	CART
Accuracy	83.72%	78.59%	80.60%	79.26%	85.61%	85.95%
Recall	0.834	0.768	0.828	0.828	0.848	0.848
Precision	0.834	0.800	0.796	0.776	0.865	0.853
F-measure	0.834	0.812	0.812	0.801	0.856	0.850
Attributes	Calculus	Calculus	Calculus	Calculus	Calculus	Alpro
	Physics	Physics	Computer Applications	Physics	Physics	Physics
	Alpro	Alpro	English	Computer Applications	Alpro	EL Basics
	Computer Applications	Computer Applications	EL Lab	Introduction to IT	Introduction to IT	Calculus
			EL Basics	EL Lab	Introduction to Program Basics	MatDis
			Physics	English		Introduction to IT
			Alpro			

DOI: https://doi.org/10.61487/jssbs.v2i3.90

E. Aggregate Data Modeling

Table 4. Classification Results of Combined Data on Student Characteristics and Values

Data	Basic Data (K2)		Training1 (K2)		Trainig2 (K2)	
Algorithm	C4.5	CART	C4.5	CART	C4.5	CART
Accuracy	83.72%	78.59%	80.60%	79.26%	85.61%	85.95%
Recall	0.834	0.768	0.828	0.828	0.848	0.848
Precision	0.834	0.800	0.796	0.776	0.865	0.853
F-measure	0.834	0.812	0.812	0.801	0.856	0.850
Attributes	Calculus	Calculus	Calculus	Calculus	Calculus	Alpro
	Physics	Physics	Computer Applications	Physics	Physics	Physics
	Alpro	Computer Applications	Alpro	Computer Applications	Alpro	EL Basics
	RTSTTB	Alpro	English	EL Lab	Introduction to IT	Calculus
	Jekel		RTSTTB	Father's PKJ	Introduction to Program Basics	MatDis
			EL Lab	Mother's Education	RTUAN	Introduction to IT
			EL Basics	English	Options	Father's PKJ

Classification is carried out to find out whether the data combines characteristics and student score data can provide better classification results than on separate data modeling. Classification is carried out using the cross-validation method. Classification is carried out on Basic Data, Training 1, and Training 2. The class attribute used in this classification is K2 which is best class attributes resulting from classification of student characteristic data. Table 4 displays classification results for each training data. Compared with the results of data classification of student scores in data modeling student scores, there was a decrease in the accuracy values of the models generated. For each model produced, it is compared with the data attributes student characteristics, the value data attribute is still the most important attribute dominant in the modeling results. Same as student score data, accuracy best obtained when classification is done in Training2 using C4.5 algorithm. In this Training2 classification, student characteristic data attributes used in decision trees are the RTUAN and Choice attributes.

DOI: https://doi.org/10.61487/jssbs.v2i3.90

F. Recommended Models

After going through the analysis process of quality values and decision trees resulting from the modeling process, there are two models that can be used for Initial prediction of academic achievement in the first semester of new study program students PNP Computer Engineering uses student characteristic data:

1. If the characteristic attributes are related to the conditions social and economic issues students want to use in predictions, models used is a model resulting from data modeling TrainingKar uses the C4.5 algorithm.
2. If the process of predicting student academic achievement is based on attributes that reflect the student's academic background, then the model resulting from the TrainingKar2 classification can used as a prediction model. From the characteristic data modeling carried out, it is known that the use of student characteristic data attributes that reflect their background Student academics can provide good predictive results. Addition other attributes that more clearly reflect academic background students such as UAN scores per subject, Polytechnic entrance exam scores Padang State both basic and natural science abilities, as well as TOEFL scores students may be able to further improve the quality of prediction models. In terms of courses, basic courses such as Calculus, Physics, Algorithms and Programming, Computer Application Practicum and Basic Electronics become the subject lectures that most determine student achievement in the first semester. No means other subjects are not important, but the modeling carried out, students who have good grades in these courses tend to are in the class of students with the above first semester academic achievements average.

CONCLUSION

Conclusions that can be drawn from this research include: Modeling for initial predictions of student academic achievement can be done carried out using data on student characteristics, but accuracy the resulting model is lower compared to usage student grade data. Using student characteristic data, there are two model's Initial prediction of academic achievement in the first semester of new Engineering students PNP computers that can be used a model that uses all characteristic data attributes both related to the student's personal and academic background and students' social background and a model that only uses background-related attributes student personal and academic. School, Choice, Jekel, RTSTTB, and RTUAN attributes are the main attributes for modeling student characteristic data. Basic courses such as Calculus, Physics, Algorithms and Programming, Application Practicum Computers and Basic Electronics are the most crucial subjects in the first semester.

REFERENCES

Husain, I., & Sari, R. P. (2021). Predictive Modeling of Student Academic Performance Using Data Mining Techniques: A Case Study of Indonesian Universities. *Journal of Educational Data Mining, 13*(1), 45-62.

Khan, M. S., & Hussain, M. S. (2022). Educational Data Mining Techniques for Predicting Student Success in Higher Education. *International Journal of Educational Technology, 19*(3), 198-215.

Lee, H. K., & Kim, Y. J. (2023). Analyzing Academic Performance Using Data Mining: A Comprehensive Review. *Computers & Education, 181*, 104383.

DOI: https://doi.org/10.61487/jssbs.v2i3.90

Martínez, A., & García, E. (2021). Applying Machine Learning Techniques for Predicting Student Performance in Online Learning Environments. *Journal of Learning Analytics, 8*(2), 32-50.

Nguyen, T. T., & Zhang, X. (2022). Data Mining Approaches for Academic Achievement Prediction in Higher Education. *Educational Data Science, 5*(4), 67-84.

Pérez, J. L., & García, I. (2023). Predicting Students' Academic Outcomes Using Educational Data Mining Techniques: A Case Study in Southeast Asia. I*nternational Journal of Advanced Computer Science and Applications, 14*(1), 123-136.

Reddy, S., & Ali, S. (2021). Exploring the Use of Data Mining Techniques for Academic Success Prediction: Evidence from a University Setting. *Educational Research Review, 16*(4), 289-305.

Sari, R., & Nugroho, S. (2022). A Comparative Study of Classification Algorithms for Predicting Student Performance in Higher Education. *International Conference on Machine Learning and Data Mining*, 114-121.

Yılmaz, R., & Çelik, E. (2021). Predictive Analytics in Higher Education: Techniques and Applications. *Data Science & Engineering, 7*(2), 204-222.

Zhang, L., & Xu, J. (2023). Harnessing Educational Data Mining to Forecast Student Performance: A Systematic Review. *Educational Technology Research and Development, 71*(1), 87-103.

DOI: https://doi.org/10.61487/jssbs.v2i3.90