Jimi Asmara[1]*
Rusijono[2]
Andi Kristanto[3]

Exploration of Student Performance Through the Application of Data Mining

Journal of Electrical Systems

Abstract: -Using data mining in the education business to estimate the performance of students who are enrolled in universities is the topic of this paper, which presents the findings of a study. Both of these data mining algorithms are utilized. A descriptive assignment that is based on the K-means algorithm is used to pick multiple groups of pupils as the first step. Second, a classification job assists with two classification techniques, namely Decision Tree and Naïve Bayes, which are responsible for predicting students who drop out of college due to poor performance in the first four semesters of their college attendance. Cross-validation techniques were utilized to evaluate the models, which were trained and tested using the academic data collected from students throughout the admissions process. After adding data from past academic enrollment, the findings of the experiment demonstrate that the prediction of students dropping out of school improves, and that student performance is monitored.
Keywords:Data mining, education, student, performance, Academic

I. INTRODUCTION

An important computational development in extracting information from obscure correlations between variables is data mining[1]. Although these links can be found by using mining techniques, the discipline's goal is to extract valuable knowledge from vast volumes of data that are initially unknown[2]. Data mining in education, often known as Educational Data Mining (EDM), is the use of data mining methods and technology in a variety of educational situations[3]. The contributions of data mining in education have been used to increase understanding of the educational process, with the main objective of providing teachers and researchers with recommendations for the improvement of the teaching-learning process[4]. By implementing data mining applications in education, teachers and administrators could organize educational resources more efficiently. The objective of the EDM is to apply data mining to traditional teaching systems – in particular to learning content management systems and intelligent web-based education systems[5]. Each of these systems has different data sources for knowledge discovery. After the pre-processing of the data in each of these systems, the different techniques of data mining are applied: statistics and visualization, grouping and classification, association rules and data mining[6]. The amount of academic information stored in the databases of educational institutions is very useful in the teaching and learning process which is why there has been significant research interest in the analysis of academic information. This research focuses on applying data mining techniques to the academic records of the students that entered the academic periods between July 2020 and June 2024 through the construction of a mining model of descriptive data, which allows to create the different profiles of the admitted students with socioeconomic information. For the development of the research, the CRISP-DM methodology was used to structure the lifecycle of a data mining project in six phases, described in four levels, which interact with each other during the development of the research[7].

II. LITERATURE REVIEW

Data mining is extensively utilized across various interdisciplinary domains, including education. Extensive study has been conducted in the domain of educational data mining. Investigate the determinants influencing school dropout rates by constructing a predictive model. This approach assesses the risk associated with disregarding students' socioeconomic data and academic records by employing decision tree methodologies and logistic regression to identify pupils at elevated risk of dropout[8].

[1*23]Department of Educational Technology, Universitas Negeri Surabaya, Surabaya, East Java, Indonesia
Jimmyasmara26@gmail.com*

The examination of learning algorithms to forecast student attrition, specifically when a student ceases their studies. Their research was driven by the significant proportion of students who fail to complete courses at universities providing online education[9]. A multitude of tests were performed using academic data, comparing the performance of decision tree algorithms, neural networks, Naive Bayes, logistic regression, and support vector machines to evaluate the suggested system [10], [11]. The analysis of the results indicated that the Naive Bayes algorithm is the most suitable for forecasting student performance in distant education systems[12]. Knowledge acquisition derived from the technique of decision tree induction, employing decision trees to represent data classification. A primary outcome achieved was the characterization of students at elevated risk of discontinuing their university education[13].socioeconomic factors include age, gender, race, handicap, employment position, and remote learning program. The objective of the study was to identify students at elevated risk of school dropout. This study employed data mining techniques, decision trees, and logistic regression[8]. Data mining study aimed at developing predictive models to identify students at elevated risk of dropout by analyzing initial enrollment records[14]. The predictive model's quality was evaluated using the ID3, C4.5, and ADT algorithms of the decision tree methodology. The ADT machine learning method can derive insights from the predictive model utilizing historical student data[3]. A data mining study was conducted employing the same strategy to construct and assess a predictive model for estimating the probability of a certain student's desertion; a decision tree method utilizing the C4.5 algorithm was applied for student classification[4]. proposed the development of a mining course management system utilizing data mining techniques. Upon processing the data within the system, the authors ascertain the attributes of students who fail to succeed in the semester[15]. This study included support vector machines, Naive Bayes, and decision trees. The Naive Bayes algorithm exhibited the highest classification precision, whilst the decision tree had one of the lowest scores.[16].The ruling received one of the lowest evaluations. Assessing critical factors that influence student performance can enhance the quality of the higher education system[17] provides a classification model utilizing the ID3 and C4.5 decision tree algorithms alongside Naive Bayes approaches. The classification accuracy of the three algorithms is suboptimal; to develop a high-quality classification model, it is essential to incorporate sufficient features[1]. The same study, investigated dropout prediction in online academic programs, employing three classification techniques: decision tree, Naive Bayes, and neural network. The algorithms were trained and assessed via cross-validation approaches. [6] Conversely introduced a data mining tool aimed at generating predictive models by analyzing first-period student records. Decision trees were employed for validation and training to identify the optimal classifier for predicting student attrition. [9][13] examined elements influencing student evaluations to enhance performance, employing clustering approaches via K-Means algorithm analysis to delineate the student demographic. [18] elucidated the correlation between admission examinations and success outcomes for students. The study employed group analysis and K-Means algorithm methodologies. [19]elucidated the utilization of data mining within the engineering education context, examining the correlation between universities and student outcomes by K-algorithm approaches analysis.

III. Data Analysis and Modeling

In order to do a preliminary examination of the records and verify the quality of the data, this chapter focuses on the interpretation of the data; visualization tools, such as histograms, are utilized to accomplish this. Following the completion of the analysis, we go on to the data preparation phase, which entails selecting the data to which the modeling techniques will be used within the context of their individual analyses.
The first thing that has to be done is to gather the preliminary data. The data sources of the academic information system of the university are the target of this work, which aims to obtain them. Language, English, Mathematics, and Logic were the subjects of the initial batch of data, which categorized the socioeconomic information and the results of the admission tests. The second set of data consists of the academic and grading history that the students have obtained. This historical information includes the academic year and period of the student's admission, the program in which the student is enrolled, the student's academic situation (academic blocking due to low academic performance or no academic blocking), and the number of academic credits that have been registered, approved, lost, canceled, and or failed. A database management system called PostgreSQL was used to generate the queries that were generated. A flat file containing 55 attributes and 1665 records of students who were admitted and enrolled in the systems and electronics engineering programs was obtained byconcatenating the two data sets of information.

Following that comes the task of data exploration[20]. A task known as exploratory analysis enables a deep investigation of certain variables and the identification of features[21]. To accomplish this, certain visualization tools, such as tables and graphs, were utilized. This endeavor aimed to describe the data mining objectives that were accomplished during the comprehension phase. To complete the process of checking the data quality, it is necessary to specify a revision of the same as the lost or those that have missing values for reasons related to coding problems. In this stage, the quality of the data that corresponds to the socioeconomic information of the student who was accepted into the program is checked[22].

What comes next is the selection of the data[23]. Within the scope of this task, the process of picking the pertinent data for the development of the data mining objectives is carried out. When it comes to the final selection of the data, the selection of attributes is the first step in the preprocessing process. It was discovered that there are 55 characteristics or variables that include values that may or may not contribute to the study. This conclusion was reached based on the initial investigation of the data and the description of the fields that are described in the variable dictionary[24]. In the dataset that was chosen for the modeling, there were no errors found in the fields; however, there were differences in the records that were selected, and the errors that were presented in some cases were missing[25]. This was since the processing was not adequate at the time of the typing. These errors included attributes that were not considered to be relevant to the case that was being studied, such as email, residence address, telephone number, date of birth, type of blood, and ethnicity. RapidMiner, an application that enables automatic learning for analysis and data mining, was utilized in the process of developing the model. RapidMiner is a program that enables the development of data analysis processes by linking operators through a graphic environment. The K-Means operator of the grouping and segmentation library was utilized in order to evaluate the quality of the groups that were discovered throughout the implementation of the algorithm. The Euclidean distance was utilized with this operator. The algorithm is accountable for both numerical and category values throughout the process. On the other hand, extra pre-processing was carried out in order to implement the normalize operator in order to normalize all of the numerical attributes that fell between 0 and 1. For the purpose of making a fair comparison between the traits, each one must have the same scale. For the purpose of characterizing the students who were accepted into the program, a grouping model was used to the dataset. This model was used to generate the various profiles of the students who were allocated to the various groups that were discovered. Additionally, the K-Means technique was used to determine what other characteristics define the separation of groups. Interaction was carried out on multiple occasions to ascertain the value of K or the total number of groups. The number K could range anywhere from 2 to 14. To select the group number, the elbow approach was utilized. The results were evaluated based on the quadratic error that was produced by each iteration.

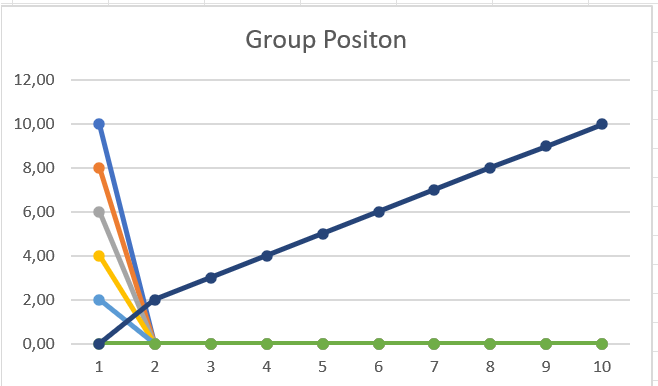

Figure 1. Admitted students' selection of Group Number (K)

The iterations used to determine the value of k on the first dataset of admitted students are depicted in Figure 1. A value of 5 is chosen for k, and the SSE is 7.954. The K Means algorithm creates five groups, and it is anticipated that the descriptions of these groups will describe the characteristics of the admitted students. The distribution of records and the percentage of each group produced are displayed in Table 1. Group 1 has the lowest percentage of records, while Groups 2 and 4 have the most records.

Tabel 1.The K-Means algorithm's implementation of registration number distribution

	Group 0	Group 1	Group 2	Group 3	Group 4
Number Of Record	217	230	319	287	391
Percent%	17%	11%	24%	23%	25%

For each variable value to be returned to its original range, the model has to be "de-normalized." The socioeconomic data and admission exam results were used to analyze the model. Next, the students' academic circumstances were examined, with each group of four enrolled students.

The distribution of records in each group during the first four semesters of academic enrolment is displayed in Table 2. Clustering the most records (28% in each group) distinguishes Groups 2 and 3. In contrast, group 0 has the fewest records (6%). First-semester students made up 47% of the registrants, followed by second-semester attendees (21%), third-semester attendees (19%), and those with four academic registrations (13%).

Tabel 2. Number of students with four enrolments distributed

No of Group	Enrollment 1	Enrollment 2	Enrollment 3	Enrollment 4	total
Group0	22	43	21	20	106
Group1	138	78	45	40	301
Group2	189	45	97	24	355
Group3	157	80	68	60	365
Group4	78	72	54	41	245
Number of Record	584	318	285	185	1372
%	47%	21%	19%	13%	100%

Tabel 3. Number of students with four enrolments distributed

No of Group	Enrollment 1		Enrollment 2		Enrollment 3		Enrollment 4	
	No Block	Block ACAD	No Block	Block ACAD	No Block	Block ACAD	No Block	Block ACAD
Group 0	12	23	45	1	12	4	16	0
Group 1	12	20	18	23	17	6	7	1
Group 2	34	22	28	37	31	2	9	0
Group 3	23	35	11	3	19	1	22	2
Group 4	19	14	11	7	33	0	23	1

The academic standing of the first four enrolled students in each cohort is displayed in Table 3. In contrast to group 2, which displays the percentage of students with the lowest academic block, group 1 is distinguished by the grouping of the students with the highest academic block. Group 0, which includes 29% of students with blocks at first enrollment, 5% at second enrolment, and 7% at third enrolment, is distinguished by its strong entrance exam results. Similar to group 2, group 1 consisted of students who performed the worst on the admission test. Blocking affected 30% of students during their first enrolment, 12% during their second, and 2% during their third and fourth enrolments. The students who performed the worst on the entrance exam and the fewest number of students with blocks are grouped to create Group 2. At the initial enrollment, 11% of pupils

experienced a block, and at the second enrolment, 5%. Similar to group 0, group 3 is distinguished by the placement of students who performed well on the admission test.With their first enrolment, 29% of pupils had a block, 6% had two, and 1% had four. Lastly, group 4 is distinguished by the fact that it has the fewest number of kids grouped using blocks.

IV. RESULT AND ANALYSIS

Two data mining models for the analysis of student's academic and non-academic data are provided in this section. The models predict the loss of academic status as a result of poor academic performance in their studies using two classification techniques: decision trees and Naïve Bayes. The models are trained using data gathered throughout the admissions process and academic history records, and they are assessed through cross-validation.

Tabel 4. Registration number and academic block per academic period

Academic Condition	Academic Period									
	2020-1	2020-2	2021-1	2021-2	2022-1	2022-2	2023-1	2023-2	2024-1	2024-2
No Block	125	132	167	120	145	112	140	156	165	177
ACAD Block	45	56	78	40	37	11	54	27	27	43
Total Rec	170	188	245	160	182	123	194	183	192	220

The total number of students who enrolled or enrolled for the first time, as well as the number of students who encountered academic obstacles as a result of subpar performance, are shown in Table 4. The number of students who encountered an academic block during each time or registration is displayed in Table 5. The first registration has the greatest number of students with academic blocks. There are fewer pupils suffering blocks in the second, third, and fourth registrations. The greatest number of students with academic blocks is shown in each academic enrollment during the 2020–1 admission year.

Tabel 5. Academic block by period of entry or first enrolment

Period Income	Block Academic									
	2020-1	2020-2	2021-1	2021-2	2022-1	2022-2	2023-1	2023-2	2024-1	2024-2
2020-1	50	23	8	9	0	2	5	0	0	0
2020-2		32	0	0	3	0	0	0	0	0
2021-1			54	6	2	2	0	0	0	0
2021-2				32	8	0	0	0	0	0
2022-1					30	11	0	0	0	0
2022-2						19	8	0	0	0
2023-1							11	10	0	0
2023-2								23	7	0
2024-1									27	1
2024-2										25

This study's classification model makes use of socioeconomic data. Bayesian classifiers and decision trees are two popular methods utilized in the classification model. These algorithms were selected due to their exceptional interpretability and simplicity.The earliest method for classifying data was decision trees; this algorithm creates a recursive decision tree when choosing the characteristic that best classifies the data based on the criterion of

maximum proportion of knowledge gain. This method involves classifying an instance by following a set of criteria that match the designated class from the root to the leaves. It is simple to convert a decision tree into a collection of classification rules. With its ability to handle both continuous and categorical attributes, C4.5 is the most representative algorithm. Considering the criterion of the highest proportion of information obtained, it recursively creates a decision tree. The attribute with the highest gain will be the root node. To increase classification accuracy and remove superfluous branches from the decision tree, the C4.5 method employs pessimistic pruning.

The Bayesian classifier is the second method to consider while creating a model. Among the best classification models is this one. Bayesian networks, which are probabilistic graphical models that enable the straightforward and accurate modeling of the underlying probability distribution to a data collection, are the foundation of Bayesian classifiers. The dependency and independence relationships between variables in a data set are represented graphically by a Bayesian network, which makes the model easier to comprehend and interpret. To estimate these probability, numerous techniques have been developed. Naive Bayes is one of the most popular practical learning algorithms because of its ease of use, high predictive capacity, short processing time, and noise resilience. Multiple models were developed and evaluated to ascertain the likelihood of a student being blocked in a specific enrollment. The initial model examined the decline of academic standing based on socio-economic data and test scores obtained during the entrance process. The second model examined the preliminary data from the enrollment process and academic records from the first four enrollments. Table 6 depicts the quantity of enrolments in the initial four enrolments that possess academic status (excluding Block and Academic Block).

Tabel 6. Academic situation in the first four enrolments

situations academic	Enrollment-1	Enrollment-2	Enrollment-3	Enrollment-4
No Block	309	190	214	145
ACAD BLOCK	255	66	9	10

The model was designed using the RapidMiner application, a tool for automated learning and data mining that employs a modular architecture, enabling the creation of learning models through the use of chained operators for diverse challenges. The Stratified Sampling approach was employed for the validation of the classification model. The operator for partitioning the dataset is termed 'split data'; this operator generates partitions of the dataset into subsets based on the specified size and chosen approach. The Decision Tree operator and the Naive Bayes algorithm were employed for the implementation of the decision tree algorithm. Table 7 presents the record count for the initial four enrollments, indicating that 70% of the records were allocated to the training set with 10-fold cross-validation, while 30% of the sample was designated as the test set.

Tabel 7. Testing Training Validation data

Number Enrollment	Total Rec	Training and Validation Data 80%		Test Data 20%	
		No Block	ACAD Block	No Block	ACAD Block
Enrolment 1	564	312	213	42	61
Enrolment 1	256	211	67	78	14
Enrolment 1	223	117	5	93	1
Enrolment 1	155	123	9	29	1

The X-validation operation was employed to assess the model' sperformance. This operator facilitates the implementation of 10-fold cross-validation on the input dataset to assess the learning algorithm. The model's

performance was evaluated using the operator Performance Binomial Classification. This operation displays the algorithm's performance metrics, including accuracy, precision, recall, error, and ROC curve. The confusion matrix is employed to analyze the errors produced by a classification model. It is a visualization instrument employed in supervised learning. Each column of the matrix denotes the quantity of predictions for each class, whereas each row signifies the instances of the actual class. The experiment yields the following measurements: accuracy, classification error, exhaustiveness (Recall), precision, F-measure, specificity, sensitivity, false negative rate, false positive rate, and area under the curve (AUC). During this phase, many models were developed and evaluated to categorize children experiencing academic block in their initial four academic enrollments, utilizing socio-economic data. We employed 10-fold cross-validation for model training and utilized the test dataset for model evaluation. The model's performance was assessed using 80% of the training and validation data, while 20% of the sample served as the test set. In the decision tree methodology utilizing training and validation data, the tree depth was adjusted from 1 to 20; the minimum classification error occurred at depth 3, where the error exhibited a degree of stability across the four academic periods. The training and validation models were ultimately assessed using the test dataset. The findings of a preconditioned model of academic state loss using training and validation data are shown in Table 8, which also compares several classification methods concerning various performance metrics.

Table 8. Academic condition loss prediction model with Training and Validation datasets

Prediction	Decision Tree			Naïve Bayes		
	Enrollment-2	Enrollment-3	Enrollment-4	Enrollment-2	Enrollment-3	Enrollment-4
Measure-F	0	0	30.77%	0	53.41%	38.51%
Precision	0	0	33.33%	0.00%	56.26%	44.17%
Exhaustive	0.00%	0.00%	30.27%	0.00%	61.23%	26.23%
Accuracy	60.66%	81,24%	88,98%	95,67%	48.90%	75.34%
Error	56.44%	25.86%	5.07%	7.24%	40.57%	30.12%
Curva	0.8	0.8	0	0	0.608	0.65
Kappa	0.1	0.0	0.282	-0,15	0.177	0.23
Specificity	100%	100%	97.61%	99.23%	66.01%	76.77%
Sevsitivity	0.00%	0.00%	28.57%	0.00%	51.45%	36.23%
False Positive	0%	0%	2%	1%	19%	12%
FALSE	43%	15%	6%	8%	31%	11%
Negative	-	-	-	-	-	-

The Bayesian classifier demonstrated the highest accuracy of accurately identified academic block records when the results of the training and validation datasets were analyzed using the admission information from the admissions procedure. Regarding the decision tree, enrollment rose by 11% in the third group. The decision tree in the first and second enrollments also performed poorly below 0.8 when the area under the curve (AUC) was examined. The greatest proportion of cases without academic block that were incorrectly categorized as having academic block was shown by the Naive Bayes algorithm. The decision tree displayed the largest percentage of academic block classes that were incorrectly categorized as non-academic block classes. The academic condition loss preconditioning model's results are shown in Table 9 along with training and validation data from the previous semester's academic records and admission information. Various classification methods are contrasted based on various performance metrics.

Tabel 9. Model for academic condition loss prediction based on training and validation data

Prediction	Decision Tree			Naïve Bayes		
	Enrollment-2	Enrollment-3	Enrollment-4	Enrollment-2	Enrollment-3	Enrollment-4
Measure-F	76.23%	1	27.99%	0	56.61%	45.56%
Precision	55.78%	1	43.67%	0.90%	67.44%	46.57%
Exhaustive	90.78%	0.00%	34.56%	0.90%	51.23%	36.23%

Accuracy	87.45%	90.76%	89.70%	90.34%	69.00%	69.34%
Error	14.55%	7.89%	6.80%	7.24%	45.90%	30.12%
Curva	0.879	0.56	0	0	0.608	0.65
Kappa	0.541	-0.435%	0.450	-0,15	0.177	0.23
Specificity	87.88%	100%	98.70%	90.78%	78.12%	81.65%
Sevsitivity	88.76%	0.00%	30.28%	0.00%	61.60%	36.23%
False Positive	10%	0%	2%	1%	20%	10%
FALSE	2%	34%	3%	9%	22%	14%
Negative	-	-	-	-	-	-

We saw how the decision tree increased its accuracy rate in the second and fourth registrations by analyzing the outcomes of the training and validation datasets. The records with correctly categorized academic blocks have higher accuracy thanks to the Bayesian classifier. In the same way, both algorithms in the second enrollment performed much better than 0.9 when looking at the area under the curve (AUC).The findings of the academic condition loss preconditioning model are shown in Table 10 along with admission data from the admissions process and exam data from the previous semester. Performance metrics are used to compare various categorization methods.

Table 10. Test data-based prediction model for decline in academic status

Prediction	Decision Tree			Naïve Bayes		
	Enrollment-2	Enrollment-3	Enrollment-4	Enrollment-2	Enrollment-3	Enrollment-4
Measure-F	79.10%	0	67.32%	0	45.61%	0.20%
Precision	65.70%	0	34.67%	0.90%	89.70%	0.00%
Exhaustive	100%	0.00%	45.56%	0.90%	65.12%	0.00%
Accuracy	54.12%	98.78%	96.70%	90.34%	78.34%	80.90%
Error	50.12%	9.89&	3.90%	7.24%	78.32%	78.54%
Curva	0.890	0.78	2	0	0.980%	0.756
Kappa	0.766	-0.760%	0.342	-0,15	0.222%	-342%
Specificity	89.70%	100%	90.90%	90.78%	90.41%	97.65%
Sevsitivity	99.76%	0.70%	45.28%	0.00%	78.50%	0,11%
False Positive	20%	0%	5%	1%	23%	9%
FALSE	2%	23%	1%	9%	20%	14%
Negative	0%	1%	2%	7%	3%	-5

The decision tree displayed the most predictions with correctly classified academic blocks on the second enrollment, according to our analysis of the testing dataset's results. Comparing the Naive Bayes method to the decision tree approach, the former performs well, with an area greater than 0.9 when the area under the curve (AUC) is examined.

V. CONCLUSION

Given the new methods and resources that make it possible to comprehend the data, educational institutions—which produce vast volumes of data—have shown a significant deal of interest in data analysis in recent years. To train and validate descriptive and predictive models for this study, a data set containing socioeconomic data and academic records of prior enrollments was gathered from the "X" University database.The K-Means algorithm's descriptive model is used to analyze a university student population to find commonalities in group characteristics. It's intriguing to learn that certain basic socioeconomic traits enable the definition of particular groupings or profiles. The model's review revealed that socioeconomic information has an impact on kids' academic achievement, with the groups with the best knowledge test scores attending schools with lower socioeconomic standing. This paper's classification approach examines socioeconomic data and prior academic records of students who have been enrolled. When the academic records from the previous semester are added, the decision tree algorithm using test data performs better than the Naive Bayes approach. The data analysis can

reveal that there are several performance kinds based on students' academic records and socioeconomic profiles, suggesting that predictions can be made and that this research can be a very helpful tool for decision-making. The University's permanent and graduating programs may use this research to inform their decisions, and it may serve as a springboard for further data mining studies in the field of education. An additional crucial suggestion is that additional data sources, such the records of high school students before they enroll in college, should be incorporated into the model to enhance its effectiveness.

REFERENCES

[1] W. Zhou and T. Yang, "Application analysis of data mining technology in ideological and political education management," *J. Phys. Conf. Ser.*, vol. 1915, no. 4, 2021, doi: 10.1088/1742-6596/1915/4/042040.

[2] S. Zhang, J. Chen, W. Zhang, Q. Xu, and J. Shi, "Education Data Mining Application for Predicting Students' Achievements of Portuguese Using Ensemble Model," *Sci. J. Educ.*, vol. 9, no. 2, p. 58, 2021, doi: 10.11648/j.sjedu.20210902.16.

[3] J. Hu and H. Li, "Composition and Optimization of Higher Education Management System Based on Data Mining Technology," *Sci. Program.*, vol. 2021, 2021, doi: 10.1155/2021/5631685.

[4] S. MP and G. Lumacad, "Role of Data Mining in Education Sector," *TechnoareteTransactions Intell. Data Min. Knowl. Discov.*, vol. 2, no. 3, pp. 8–12, 2022, doi: 10.36647/ttidmkd/02.03.a002.

[5] L. Wang and S. J. Chung, "Sustainable Development of College and University Education by Use of Data Mining Methods," *Int. J. Emerg. Technol. Learn.*, vol. 16, no. 5, pp. 102–115, 2021, doi: 10.3991/ijet.v16i05.20303.

[6] B. Chen, Y. Liu, and J. Zheng, "Using Data Mining Approach for Student Satisfaction With Teaching Quality in High Vocation Education," *Front. Psychol.*, vol. 12, no. January, pp. 1–9, 2022, doi: 10.3389/fpsyg.2021.746558.

[7] E. Okewu, P. Adewole, S. Misra, R. Maskeliunas, and R. Damasevicius, "Artificial Neural Networks for Educational Data Mining in Higher Education: A Systematic Literature Review," *Appl. Artif. Intell.*, vol. 35, no. 13, pp. 983–1021, 2021, doi: 10.1080/08839514.2021.1922847.

[8] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electron.*, vol. 9, no. 8, pp. 1–12, 2020, doi: 10.3390/electronics9081295.

[9] H. E. Abdelkader, A. G. Gad, A. A. Abohany, and S. E. Sorour, "An Efficient Data Mining Technique for Assessing Satisfaction Level With Online Learning for Higher Education Students during the COVID-19," *IEEE Access*, vol. 10, pp. 6286–6303, 2022, doi: 10.1109/ACCESS.2022.3143035.

[10] A. S. Al-Gahmi, K. Feuz, and Y. Zhang, "On Time-based Exploration of LMS Data and Prediction of Student Performance," *ASEE Annu. Conf. Expo. Conf. Proc.*, 2022.

[11] A. Al-Gahmi, K. D. Feuz, and Y. Zhang, "On Time-based Exploration of Student Performance Prediction," *ASEE Annu. Conf. Expo. Conf. Proc.*, 2023, doi: 10.18260/1-2--43772.

[12] P. Subarkah, S. A. Solikhatin, I. Darmayanti, A. N. Ikhsan, D. U. Hidayah, and R. M. Anjani, "Prediction of Education Level in Population Data Using Naïve Bayes Algorithm," *TIERS Inf. Technol. J.*, vol. 3, no. 2, pp. 69–75, 2022, doi: 10.38043/tiers.v3i2.3865.

[13] X. Zheng, Q. Lei, R. Yao, Y. Gong, and Q. Yin, "Image segmentation based on adaptive K-means algorithm," *Eurasip J. Image Video Process.*, vol. 2018, no. 1, 2018, doi: 10.1186/s13640-018-0309-3.

[14] A. Alshanqiti and A. Namoun, "Predicting student performance and its influential factors using hybrid regression and multi-label classification," *IEEE Access*, vol. 8, pp. 203827–203844, 2020, doi: 10.1109/ACCESS.2020.3036572.

[15] Y. Su *et al.*, "Exercise-enhanced sequential modeling for student performance prediction," *32nd AAAI Conf. Artif. Intell. AAAI 2018*, pp. 2435–2443, 2018, doi: 10.1609/aaai.v32i1.11864.

[16] A. Supriyadi, "Perbandingan Algoritma Naive Bayes dan Decision Tree(C4.5) dalam Klasifikasi Dosen Berprestasi," *Gener. J.*, vol. 7, no. 1, pp. 39–49, 2023, doi: 10.29407/gj.v7i1.19797.

[17] B. Macfarlane, "Student performativity in higher education: converting learning as a private space into a public performance," *High. Educ. Res. Dev.*, vol. 34, no. 2, pp. 338–350, 2015, doi: 10.1080/07294360.2014.956697.

[18] A. P. Windarto *et al.*, "Analysis of the K-Means Algorithm on Clean Water Customers Based on the Province," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, 2019, doi: 10.1088/1742-6596/1255/1/012001.

[19] M. Capo, A. Perez, and J. A. Lozano, "A Cheap Feature Selection Approach for the K-Means Algorithm," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 5, pp. 2195–2208, 2021, doi: 10.1109/TNNLS.2020.3002576.

[20] J. Goyal and S. Sharma, "Education Data Mining: A Review," *Education*, pp. 13515–13518, 2017, doi: 10.15680/IJIRSET.2017.0607230.

[21] P. Blikstein and M. Worsley, "Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks," *J. Learn. Anal.*, vol. 3, no. 2, pp. 220–238, 2016, doi: 10.18608/jla.2016.32.11.

[22] C. A. Palacios, J. A. Reyes-Suárez, L. A. Bearzotti, V. Leiva, and C. Marchant, "Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in chile," *Entropy*, vol. 23, no. 4, pp. 1–23, 2021, doi: 10.3390/e23040485.

[23] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, "Predicting student performance from LMS data," *IEEE Trans. Learn. Technol.*, vol. 10, no. 1, pp. 17–29, 2017.

[24] A. J. Boevé, R. R. Meijer, R. J. Bosker, J. Vugteveen, R. Hoekstra, and C. J. Albers, "Implementing the flipped classroom: an exploration of study behaviour and student performance," *High. Educ.*, vol. 74, no. 6, pp. 1015–1032, 2017, doi: 10.1007/s10734-016-0104-y.

[25] S. Freeman *et al.*, "Active learning increases student performance in science, engineering, and mathematics," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 23, pp. 8410–8415, 2014, doi: 10.1073/pnas.1319030111.