BULLETIN OF ENGINEERING SCIENCE, TECHNOLOGY AND INDUSTRY
(BESTI Journal) | ISSN(e): 3025-5821
Website: https://bestijournal.org/index.php/go/index | Volume 4 No. 3 (2026)

IMPLEMENTATION OF MACHINE LEARNING WITH SVM ALGORITHM FOR EARLY DETECTION OF STUDENT DROPOUT RISK

Sri Mulyati[1]*, Endar Nirmala [2]
Universitas Pamulang / Tangerang Selatan
Universitas Pamulang / Tangerang Selatan
E-mail: dosen00391@unpam.ac.id[1]*, dosen00216@unpam.ac.id[2]

Received: 21/07/2026 | **Revised:** 11/08/2026 | **Accepted:** 25/08/2026 | **Published:** 09/09/2026

Abstract

One issue that higher education institutions face is dropout rates. The issue of students quitting or dropping out of school can affect a college's reputation, accreditation, and long-term viability. The institution, the individual student, their family, and human resource development are all negatively impacted by student dropout. Universities must thus predict the probability of student dropout to take early preventive action. This study focuses on developing a prediction model for dividing students into three groups: successful dropouts, enrolled, and graduates, using data processing and machine learning approaches. The research data will include academic grades, semesters, attendance, economic background, occupation, marital or single status, and e-learning. High research results using these parameters and the SVM algorithm achieved an accuracy of 78.5%. A degree of accuracy was successfully achieved by the research results using these parameters and the Support Vector Machine (SVM) method, suggesting that the Support Vector Machine (SVM) algorithm is appropriate for predicting student dropout in higher education.

Keywords : Classification, Dropout, Suppot Vector Machine, SVM, Machine Learning

INTRODUCTION

Research conducted by Viloria et al. (2019) defines dropping out as the termination of studies before achieving a suitable degree. Dropping out of college or university students is a problem and challenge for higher education institutions due to its negative impacts (Delgado-García et al., 2025; Vaarma & Li, 2024) . According to Niyogisubizo et al. (2022) dropout is a global problem that affects not only students but also schools or universities, families, and the general public. Research by Deleña et al. (2025) states that the problem of student dropout remains a challenge in higher education, especially in developing countries due to limited institutional resources for intervention.

In their research González-Ortiz-de-Zárate et al. (2025) stated that student dropout is a phenomenon where students decide to stop or withdraw before completing their studies. This has a negative impact on both the students themselves and the educational institution, as it results in a loss of resources and overall academic achievement.
The increasing number of students dropping out of school due to social, economic, personal, and health problems is a worrying condition for the government, educators, and parents (Mustofa et al., 2025) . Information on students with the potential risk of dropping out can prevent dropouts, protect state resources, and decrease the quality and productivity of human resources in the long term (Delogu et al., 2024) . The ability of universities to retain students in completing their studies and obtaining degrees will impact the institution's reputation because it reflects the institution's ability to prepare its graduates (Cardona & Cudney, 2019) .

Many studies have applied machine learning methods to predict students at risk of dropping out of school (Pecuchova & Drlik, 2023) . One popular machine learning method for classification is the Support Vector Machine (SVM) (Maniyan et al., 2024; Mustofa et al., 2025; Vaarma & Li, 2024) . Accurate predictions in identifying students at risk of dropping out will minimize its impact and improve student academic success (Aini et al., 2025) . The SVM method has advantages in handling high-dimensional data and is able to provide a high level of accuracy in predictions with optimal separation margins. Based on research conducted by Hidayat et al. (2024) , the SVM method has advantages in handling high-dimensional data and is able to provide a high level of accuracy in predictions with optimal separation margins. The application of SVM in predicting student dropouts has been widely studied and proven effective, such as in research conducted at the Lhokseumawe State Polytechnic. This study produced a

OPEN ACCESS

prediction accuracy level of 86.92% with good precision and recall values for both dropout and non-dropout classes (successfully graduated), so this method is considered good in detecting students at risk of dropping out of college. Physics Research (2017) predicted the problems of Vocational High School (SMK) students, whether they could continue to the next semester or drop out based on student absence as the main causal factor. This student absence factor can be caused by several things, namely economic factors where students cannot afford school fees and due to illness. Several other factors considered were behavior, subject grades, teachers, and school fees. The method used in this study was the Support Vector Machine (SVM).

Artificial intelligence can provide fairly accurate predictions of the likelihood of student failure based on economic, health, and social factors. These factors significantly influence student academic performance (López-García et al., 2023) . Various artificial intelligence techniques have made significant contributions to the mission of predicting when a student is likely to drop out (Martínez & Castillo, 2024) . The development of artificial intelligence has given rise to machine learning, a subfield of artificial intelligence that studies how machines learn and improve statistical models to perform tasks without explicit instructions (Beckham et al., 2022) . These studies show a gap in the use of independent variables used, namely academic and economic data, therefore this study aims to analyze the use of independent variables (features) using academic, economic, and demographic data in the formation of training models and their variable correlations.

METHOD

This study uses an experimental model to analyze the correlation of independent variables in classification and accuracy for the problem of student dropouts using Support Vector Machine (SVM). This experiment uses 13 independent variables with several configurations of split dataset modeling, the first using a ratio of 70:30, the second 80:20, and the third 90:10 with 70%, 80%, 90% being training data and 30%, 20%, and 10% being test data. The value of C as a Regularization Parameter uses 10 and 100 to minimize classification errors in the training data. The stages of this study are illustrated in Figure 1.

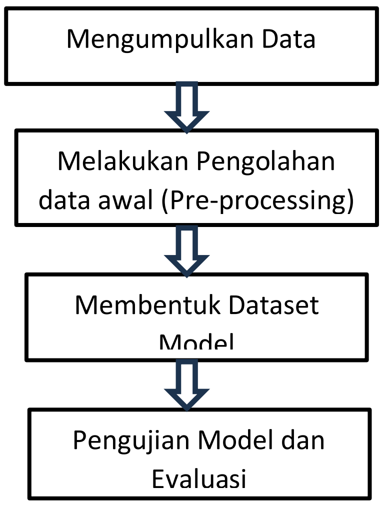

Figure 1. Research Stages

Based on Figure 1, the initial stage of the research was to collect a dataset of dropout students, then preprocessing was performed to handle incomplete data (missing values) and duplicate data. Next, the model was built using the Support Vector Machine (SVM) algorithm by dividing the dataset into training data and test data. Then, the model was tested to determine its accuracy, and finally, validation and evaluation were performed using a Confusion Matrix.

a. Data collection

This study used a dataset of dropout students sourced from www.kaggle.com and the UCI (University of California, Irvine) Machine Learning Repository. The dataset consists of 37 features or variables divided into three categories: academic, economic, and demographic data. The 37 variables are divided into 36 independent variables and one dependent variable (Target). The research experiment used 14 variables: 13 independent variables and one dependent variable.

Table 1. Dropout Student Dataset

Marital status	Applicatio	Applicatio	Course	Daytime/evening atte		Previous q	Previous q	Nacionalit	Mother's q	Father's qu	Mother's o	Father's oc	Admission	Displaced	Educationa	Debtor	Tuition fee	Gender	Scholarshi	Age at enr	Internation
1	17	5	171	1	1	122.0	1	19	12	5	9	127.3	1	0	0	1	1	0	20	0	0
1	15	1	9254	1	1	160.0	1	1	3	3	3	142.5	1	0	0	0	1	0	19	0	0
1	1	5	9070	1	1	122.0	1	37	37	9	9	124.8	1	0	0	0	1	0	19	0	0
1	17	2	9773	1	1	122.0	1	38	37	5	3	119.6	1	0	0	1	0	0	20	0	0
2	39	1	8014	0	1	100.0	1	37	38	9	9	141.5	0	0	0	1	0	0	45	0	0
2	39	1	9991	0	19	133.1	1	37	37	9	7	114.8	0	0	1	1	1	0	50	0	0
1	1	1	9500	1	1	142.0	1	19	38	7	10	128.4	1	0	0	1	0	1	18	0	0
1	18	4	9254	1	1	119.0	1	37	37	9	9	113.1	1	0	0	0	1	0	22	0	0
1	1	3	9238	1	1	137.0	62	1	1	9	9	129.3	0	0	0	1	0	1	21	1	0
1	1	1	9238	1	1	138.0	1	1	19	4	7	123.0	1	0	1	0	0	0	18	0	0
1	1	1	9670	1	1	139.0	1	38	19	5	7	130.6	1	0	0	1	0	0	18	0	0
1	1	1	9500	1	1	136.0	1	19	38	9	9	119.3	1	0	0	1	0	1	18	0	0
1	1	2	9853	1	1	133.0	1	19	37	4	9	130.2	1	0	0	1	0	0	19	0	0
1	53	1	9254	1	42	110.0	1	1	1	4	7	111.8	1	0	0	1	0	1	21	0	0
1	1	1	9085	1	1	149.0	1	38	37	5	5	137.1	1	0	0	1	0	1	18	0	0
1	1	1	9773	1	1	127.0	1	19	37	9	3	120.7	1	0	0	1	0	0	20	0	0
1	18	1	9238	1	1	137.0	1	19	38	5	8	137.4	1	0	0	1	0	0	18	0	0
1	17	2	9500	1	1	135.0	1	19	1	5	4	127.3	1	0	0	1	0	0	18	0	0
1	1	1	9130	1	1	137.0	1	3	19	3	5	136.3	1	0	0	1	0	0	20	0	0
1	1	1	9853	1	1	140.0	1	19	19	7	7	124.6	1	0	0	1	0	0	18	0	0

Curricular	Curricular	Curricular	Curricular	Curricular	Curricular	Curricular	Curricular	Curricular	Curricular	Curricular	Curricular	Unemploy	Inflation rate	Target
0	0	0	0.0	0	0	0	0	0	0.0	0	10.8	1.4	1.74	Dropout
6	6	6	14.0	0	0	6	6	6	1,37E+16	0	13.9	-0.3	0.79	Graduate
6	0	0	0.0	0	0	6	0	0	0.0	0	10.8	1.4	1.74	Dropout
6	8	6	1,34E+16	0	0	6	10	5	12.4	0	9.4	-0.8	-3.12	Graduate
6	9	5	1,23E+16	0	0	6	6	6	13.0	0	13.9	-0.3	0.79	Graduate
5	10	5	1,19E+16	0	0	5	17	5	11.5	5	16.2	0.3	-0.92	Graduate
7	9	7	13.3	0	0	8	8	8	14345	0	15.5	2.8	-4.06	Graduate
5	5	0	0.0	0	0	5	5	0	0.0	0	15.5	2.8	-4.06	Dropout
6	8	6	13875	0	0	6	7	6	1,41E+16	0	16.2	0.3	-0.92	Graduate
6	9	5	11.4	0	0	6	14	2	13.5	0	8.9	1.4	3.51	Dropout
6	6	6	1,23E+16	0	0	6	7	5	14.2	0	13.9	-0.3	0.79	Graduate
8	8	7	1,32E+16	0	0	8	8	7	1,32E+16	0	12.7	3.7	-1.7	Graduate
6	6	0	0.0	0	0	6	0	0	0.0	0	12.7	3.7	-1.7	Dropout
6	7	6	1,06E+16	0	0	6	8	5	11.0	0	8.9	1.4	3.51	Graduate
5	7	4	13.25	0	0	5	5	5	12.0	0	10.8	1.4	1.74	Graduate
6	6	5	13.2	0	0	6	7	0	0.0	0	15.5	2.8	-4.06	Dropout
6	10	1	12.0	0	0	6	14	2	11.0	0	10.8	1.4	1.74	Enrolled
7	8	7	1330625	0	0	8	8	8	14545	0	15.5	2.8	-4.06	Graduate
5	8	4	12.5	1	0	5	8	4	12.25	2	10.8	1.4	1.74	Graduate
7	7	6	1,17E+16	0	0	7	8	6	13.5	0	16.2	0.3	-0.92	Enrolled

The dataset consists of 3 classes, the Dropout class with 1421 data, the Enrolled class with 794 data, and the Graduate class with 2209 data. Then feature selection was carried out, taking 14 variables.

Table 2. Feature Selection Result Dataset

	Mothers qualification	Previous qualification (grade)	Tuition fees up to date	Age at enrollment	Curricular units 1st sem (approved)	Curricular units 1st sem (grade)	Curricular units 1st sem (enrolled)	Curricular units 2nd sem (approved)	Curricular units 2nd sem (grade)	Curricular units 2nd sem (enrolled)	Curricular units 2nd sem (evaluations)	Admission grade	Daytime-evening attendance	Target
0	19	122.0	1	20	0	0.000000e+00	0	0	0.000000e+00	0	0	127.3	1	0
1	1	160.0	0	19	6	1.400000e+01	6	6	1.366667e+16	6	6	142.5	1	2
2	37	122.0	0	19	0	0.000000e+00	6	0	0.000000e+00	6	0	124.8	1	0
3	38	122.0	1	20	6	1.342857e+16	6	5	1.240000e+01	6	10	119.6	1	2
4	37	100.0	1	45	5	1.233333e+16	6	6	1.300000e+01	6	6	141.5	0	2
5	37	133.1	1	50	5	1.185714e+16	5	5	1.150000e+01	5	17	114.8	0	2
6	19	142.0	1	18	7	1.330000e+01	7	8	1.434500e+04	8	8	128.4	1	2
7	37	119.0	0	22	0	0.000000e+00	5	0	0.000000e+00	5	5	113.1	1	0
8	1	137.0	1	21	6	1.387500e+04	6	6	1.414286e+16	6	7	129.3	1	2
9	1	138.0	0	18	5	1.140000e+01	6	2	1.350000e+01	6	14	123.0	1	0
10	38	139.0	1	18	6	1.233333e+16	6	5	1.420000e+01	6	7	130.6	1	2
11	19	136.0	1	18	7	1.321429e+16	8	7	1.321429e+16	8	8	119.3	1	2
12	19	133.0	1	19	0	0.000000e+00	6	0	0.000000e+00	6	0	130.2	1	0
13	1	110.0	1	21	6	1.057143e+16	6	5	1.100000e+01	6	8	111.8	1	2
14	38	149.0	1	18	4	1.325000e+01	5	5	1.200000e+01	5	5	137.1	1	2
15	19	127.0	1	20	5	1.320000e+01	6	0	0.000000e+00	6	7	120.7	1	0
16	19	137.0	1	18	1	1.200000e+01	6	2	1.100000e+01	6	14	137.4	1	1
17	19	135.0	1	18	7	1.330625e+06	7	8	1.454500e+04	8	8	127.3	1	2
18	3	137.0	1	20	4	1.250000e+01	5	4	1.225000e+01	5	8	136.3	1	2
19	19	140.0	1	18	6	1.166667e+16	7	6	1.350000e+01	7	8	124.6	1	1

Figure 3. Feature Selection Results

The correlation between the 14 independent and dependent variables will be examined. The correlation between variables influences student dropout rates. The following correlation between the independent and dependent variables is illustrated in Figure 2.

OPEN ACCESS

```
Target                                    1.000000
Curricular units 2nd sem (approved)       0.624157
Curricular units 1st sem (approved)       0.529123
Tuition fees up to date                   0.409827
Curricular units 1st sem (grade)          0.311004
Curricular units 2nd sem (grade)          0.307311
Curricular units 2nd sem (enrolled)       0.175847
Curricular units 1st sem (enrolled)       0.155974
Admission grade                           0.120889
Previous qualification (grade)            0.103764
Curricular units 2nd sem (evaluations)    0.092721
Daytime-evening attendance                0.075107
Mothers qualification                    -0.043178
Age at enrollment                        -0.243438
```

Figure 4. Variable Correlation Values

Academic data variables influence the Target variable, and economic data, namely accuracy in paying tuition fees (up-to-date tuition fees), plays a significant role in student dropout outcomes. Meanwhile, demographic data, namely Daytime-evening attendance, Mother's qualification, and Age at enrollment, have low values, meaning they are not very significant for dropout but still have an influence.

a. Pre-processing

This initial process is for data cleaning, eliminating missing data and duplicate data to obtain the correct data. This stage is crucial because it impacts the model. The results for missing data and duplicate data are 0, so the dataset remains at 4,424.

```
Total Missing Value:
0
Mothers qualification                     0.0
Previous qualification (grade)            0.0
Tuition fees up to date                   0.0
Age at enrollment                         0.0
Curricular units 1st sem (approved)       0.0
Curricular units 1st sem (grade)          0.0
Curricular units 1st sem (enrolled)       0.0
Curricular units 2nd sem (approved)       0.0
Curricular units 2nd sem (grade)          0.0
Curricular units 2nd sem (enrolled)       0.0
Curricular units 2nd sem (evaluations)    0.0
Admission grade                           0.0
Daytime-evening attendance                0.0
Target                                    0.0
```

Figure 5. Missing Data

```
Jumlah data setelah menghapus duplikat:
(4423, 14)
```

Figure 6. Results of the Data Duplication Process

After feature selection and duplication process, there was one data that was detected as duplicate data and was removed with the result that the dataset was reduced by one to 4423.

b. Modeling

The modeling process was carried out by splitting the dataset with several models, using a ratio of 70% training data : 30% test data, the second experiment with a ratio of 80% training data : 20%, and the third experiment using a ratio of 90% training data : 10%. The parameters used were C and the RBF kernel.

c. Model Testing and Evaluation

After splitting, model testing is carried out, namely testing the training data with the test data. The experiment was carried out with several tests, namely 1) the first experiment using a ratio of 70% training data (training data): 30% test data (test data), the second experiment with a ratio of 80% training data (training data): 20%, and the third experiment using a ratio of 90% training data (training data): 10%, 2) using academic data variables; academic and economic data variables; academic data variables, economic data, and demographic data, 3) using three groups or three classes, namely dropout, enrolled, and graduate classes. and economic data.

The images below are the results of experiments using academic, economic, and demographic data.

OPEN ACCESS

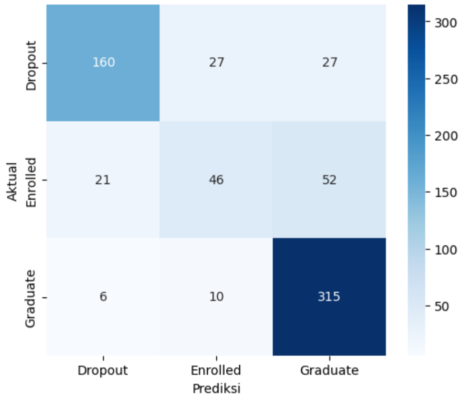

Figure 7. Academic Data Testing

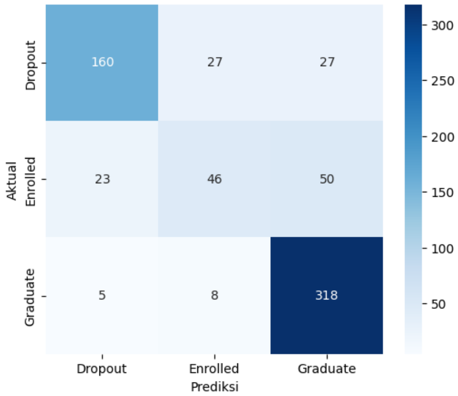

Figure 8. Academic & Economic Data Testing

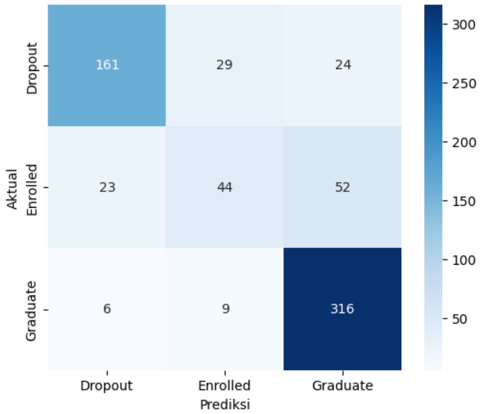

Figure 9. Testing of Academic, Economic and Demographic Data

After conducting experiments with three groups of variables, we continued with experiments based on several model dataset comparisons. The figures below show the results of comparing the data splits 70:30, 80:20, and 90:10.

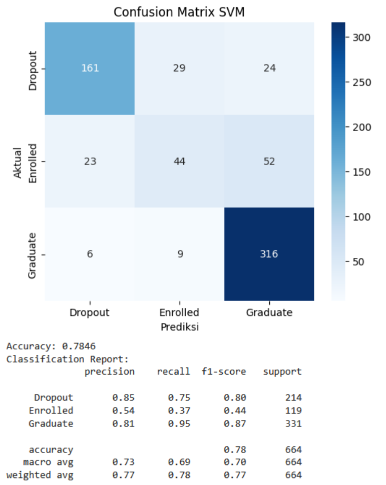

Figure 10. Results of the 70:30 Comparison

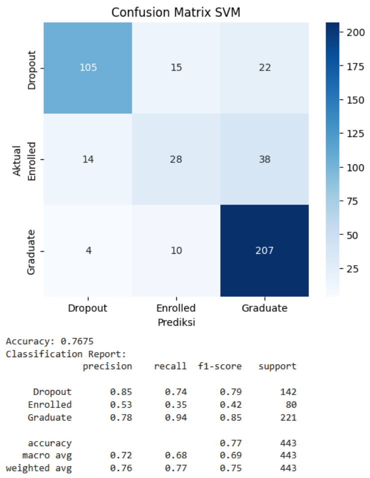

Figure 11. Results of the 80:20 Comparison

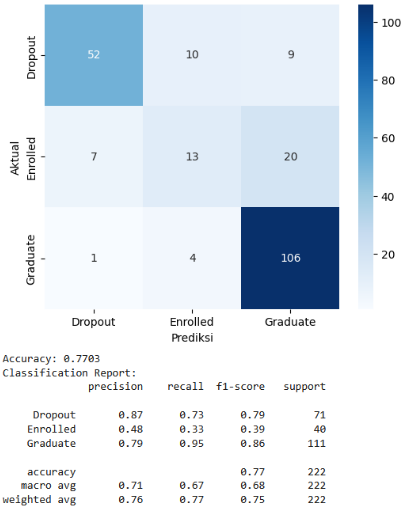

Figure 12. Result of 90:10 Comparison

RESULTS AND DISCUSSION

Based on the experiments conducted, this research found:

a. Results

Experiments based on the categories of academic data variables, economic data and demographic data are as follows:

Table 3. Variable Comparison Results

Variables	Class	Precision	Recall	F1-score
Academic	Dropout	0.85	0.75	0.80
	Enrolled	0.54	0.37	0.44
	Graduate	0.81	0.95	0.87
Academic & Economics	Dropout	0.83	0.74	0.79
	Enrolled	0.55	0.38	0.39
	Graduate	0.78	0.96	0.86
Academic, Economic, and Demographic	Dropout	0.84	0.76	0.80
	Enrolled	0.58	0.39	0.47
	Graduate	0.81	0.95	0.87

Table 4. Comparison of Accuracy Levels

Split Data	Accuracy
70:30	78.46%
80:20	76.75%
90:10	77.03%

b. Discussion

The experiment used unbalanced class data, with 1421 dropouts, 794 enrolled, and 2209 graduates significantly impacting the accuracy rate. The enrolled class had the least amount of data, and when class predictions were made, the graduate class, the class with the most data, was heavily targeted. The highest dropout precision was achieved when using only academic data at 85%, but the highest dropout recall was achieved when using academic, economic, and demographic variables. The highest accuracy was achieved using academic, economic, and demographic data variables, using a training data ratio of 70% to test data of 30%.

CONCLUSION

Prediction accuracy is influenced by the balance of data in each class; data from the class with the lowest number of data points is then directed to the class with the highest number. This error significantly impacts accuracy.

Furthermore, accuracy is also influenced by the ratio of training data to test data. Based on an accuracy rate of 78.46%, the Support Vector Machine (SVM) algorithm remains effective in predicting student dropout.

FUTURE WORK

The following research development should utilize classes with balanced datasets and study program datasets to obtain more accurate results. The method used could be a hybrid algorithm, combining two algorithms to increase accuracy, such as SVM and Xgboost.

REFERENCES

Aini, Q., Rahajeng, E., Tiohandra, M., Pratama, H. A., & Hammad, J. (2025). Predictive Modeling of Student Dropout Using Academic Data and Machine Learning Techniques. *Applied Information System and Management (AISM)*, *8*(2), 203–212. https://doi.org/10.15408/aism.v8i2.46659

Beckham, N. R., Akeh, L. J., Mitaart, G. N. P., & Moniaga, J. V. (2022). Determining factors that affect student performance using various machine learning methods. *Procedia Computer Science*, *216*(2022), 597–603. https://doi.org/10.1016/j.procs.2022.12.174

Cardona, T. A., & Cudney, E. A. (2019). Predicting student retention using support vector machines. *Procedia Manufacturing*, *39*(2019), 1827–1833. https://doi.org/10.1016/j.promfg.2020.01.256

Deleña, R. D., Dia, N. J., Sacayan, R. R., Sieras, J. C., Khalid, S. A., Macatotong, A. H. T., & Gulam, S. B. (2025). Predicting student retention: A comparative study of machine learning approach utilizing sociodemographic and academic factors. *Systems and Soft Computing*, *7*(June). https://doi.org/10.1016/j.sasc.2025.200352

Delgado-García, M., García-Prieto, F. J., López-Aguilar, D., & Ceinos-Sanz, C. (2025). Association between procrastination and academic dropout intention in university students. *Social Sciences and Humanities Open*, *11*(February). https://doi.org/10.1016/j.ssaho.2025.101506

Delogu, M., Lagravinese, R., Paolini, D., & Resce, G. (2024). Predicting dropout from higher education: Evidence from Italy. *Economic Modelling*, *130*(October 2023), 106583. https://doi.org/10.1016/j.econmod.2023.106583

Fiska, R. R. (2017). Penerapan Teknik Data Mining dengan Metode Support Vector Machine. *Sains Dan Teknologi Informasi (SATIN)*, *3*(1), 16–23.

González-Ortiz-de-Zárate, A., Alonso-García, M. A., Gómez-Flechoso, M. de los Á., & Aliagas, I. (2025). Peer mentoring, university dropout and academic performance before, during, and after the pandemic in Spain. *Evaluation and Program Planning*, *113*(February). https://doi.org/10.1016/j.evalprogplan.2025.102676

Hidayat, R., Haris, M., & Simbolon, Z. K. (2024). *Implementasi Metode Support Vector Machine (SVM) Pada Klasifikasi Drop Out (DO) Mahasiswa*. *9*, 51–59.

López-García, A., Blasco-Blasco, O., Liern-García, M., & Parada-Rico, S. E. (2023). Early detection of students' failure using Machine Learning techniques. *Operations Research Perspectives*, *11*(April), 100292. https://doi.org/10.1016/j.orp.2023.100292

Maniyan, S., Ghousi, R., & Haeri, A. (2024). Data mining-based decision support system for educational decision makers: Extracting rules to enhance academic efficiency. *Computers and Education: Artificial Intelligence*, *6*(May), 100242. https://doi.org/10.1016/j.caeai.2024.100242

Martínez, J., & Castillo, D. (2024). Prediction of student dropout using Artificial Intelligence algorithms. *Procedia Computer Science*, *251*, 764–770. https://doi.org/10.1016/j.procs.2024.11.182

Mustofa, S., Emon, Y. R., Mamun, S. Bin, Akhy, S. A., & Ahad, M. T. (2025). A novel AI-driven model for student dropout risk analysis with explainable AI insights. *Computers and Education: Artificial Intelligence*, *8*(November 2024), 100352. https://doi.org/10.1016/j.caeai.2024.100352

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, *3*(November 2021), 100066. https://doi.org/10.1016/j.caeai.2022.100066

Pecuchova, J., & Drlik, M. (2023). Predicting Students at Risk of Early Dropping Out from Course Using Ensemble Classification Methods. *Procedia Computer Science*, *225*, 3223–3232. https://doi.org/10.1016/j.procs.2023.10.316

Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technology in Society*, *76*(September 2023), 102474. https://doi.org/10.1016/j.techsoc.2024.102474

OPEN ACCESS

Viloria, A., Padilla, J. G., Vargas-Mercado, C., Hernández-Palma, H., Llinas, N. O., & David, M. A. (2019). Integration of data technology for analyzing university dropout. *Procedia Computer Science*, *155*(2018), 569–574. https://doi.org/10.1016/j.procs.2019.08.079

OPEN ACCESS