Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 4 (2024) 542 - 554 e-ISSN: 2580-0760

Comparing Correlation-Based Feature Selection and Symmetrical Uncertainty for Student Dropout Prediction

Haryono Setiadi[1]*, Indah Paksi Larasati[2], Esti Suryani[3], Dewi Wisnu Wardani[4],
Hasan Dwi Cahyono[5], Ardhi Wijayanto[6], Afrizal Doewes[6]
[1-6] Research Group Data Information Knowledge and Engineering, Department of Informatics, Universitas Sebelas Maret, Surakarta, Indonesia
[6]Eindhoven University of Technology, The Netherlands
[1]hsd@staff.uns.ac.id, [2]ip_larasati@student.uns.ac.id, [3]estisuryani@staff.uns.ac.id, [4]dww_ok@uns.ac.id,
[5]hasandc@staff.uns.ac.id, [6]ardhi.wijayanto@staff.uns.ac.id, a.doewes @tue.nl

Abstract

Predicting student dropout is essential for universities dealing with high attrition rates. This study compares two feature selection (FS) methods—correlation-based feature selection (CFS) and symmetrical uncertainty (SU)—in educational data mining for dropout prediction. We evaluate these methods using three classification algorithms: decision tree (DT), support vector machine (SVM), and naive Bayes (NB). Results show that SU outperforms CFS overall, with SVM achieving the highest accuracy (98.16%) when combined with SU Moreover, this study identifies total credits in the fourth semester, cumulative GPA, gender, and student domicile as key predictors of student dropout. This study shows how using feature selection methods can improve the accuracy of predicting student dropout, helping educational institutions retain students better.

Keywords: educational data mining; performance evaluation; correlation-based feature selection; symmetrical uncertainty

How to Cite: Haryono Setiadi, I. P. Larasati, Esti Suryani, D. W. Wardani, H. D. C. Wardani, and Ardhi Wijayanto, "Comparing Correlation-Based Feature Selection and Symmetrical Uncertainty for Student Dropout Prediction", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 8, no. 4, pp. 542 - 554, Aug. 2024.
DOI: https://doi.org/10.29207/resti.v8i4.5911

1. Introduction

Students can be expelled from or have their study rights terminated by a higher education institution before completing their studies without being transferred to another educational institution [1], [2], referred to as a dropout. A high dropout rate can lead to a lack of high-quality university graduates, which is essential for the job market's growth[3]. Moreover, dropout has a significant impact on the costs that individuals, educational institutions, and society must bear [4], [5]. A high dropout rate can lower the quality of higher education and impact accreditation [6]. To reduce dropout rates, higher education institutions must analyse previous dropout cases. Universities can use these data to employ effective preventive measures [7], [8], [9].

Higher education institutions keep student track records, including academic transactional data and educational administration. Effective decision-making in data processing requires the use of appropriate methods to extract knowledge from large volumes of data. Data mining is a tool that helps find hidden patterns and connections in large datasets to aid decision-making [10], [11], [12]. Although it is considered a new paradigm, data mining has found applications in various fields, including education, owing to its significance in decision-making [13]. The application of data mining methods to educational data, namely educational data mining (EDM), aims to extract information about student academic performance, evaluate learning systems, provide feedback to faculty and instructors, and predict student academic results to prevent dropout [14], [15].

The issue of student dropout has been thoroughly examined through educational data mining (EDM) techniques. Hegde and Prageeth [1] predicted the dropout of higher education students using the Naive

Received: 08-07-2024 | Accepted: 22-08-2024 | Published Online: 26-08-2024

Bayes (NB) algorithm with 72% classification accuracy. In addition, the results showed that students who have failed a course (Fs) four times, three times failed due to attendance (FAs), suffer from health problems, and do not adapt to the institutional atmosphere can be expelled from their institutions.

Alban et al. [16] provided a systematic review of the literature regarding predicting university student dropout using data mining techniques and showed that the decision tree (DT) algorithm is the most commonly employed data mining technique, representing approximately 79% of the 28 studies examined, followed by neural networks and SVM as the second most prevalent techniques. Furthermore, they highlighted that DT classifiers, notably the C4.5, ID3, and CART classifiers, consistently achieved high accuracy, reaching 98%, 97.5%, and 97%, respectively.

Original academic data contains many irrelevant and redundant variables, which affect prediction results. Feature selection helps eliminate unnecessary variables from the data to make it easier to understand. Optimisation enhances prediction accuracy, streamlines computational demands, reduces data complexity, and expedites information extraction [17], [18]. By excluding unnecessary attributes, variables are eliminated, streamlining the mining process and making it more efficient [19].

Bhimavatapu [20] developed a deep learning model to forecast student performance in an online learning environment, using symmetrical uncertainty (SU) to identify relevant features associated with students and their contribution to academic performance assessment. The method measured the importance of each extracted feature compared to the target feature. The model detected at-risk students with an accuracy of 98.80%. Nuanmeesri et al. [21] used another feature selection method in EDM. Their objective was to enhance the performance of classification models by integrating feature selection into neural network methods. The feature selection methods used for comparison included gain ratio, chi-square, and correlation-based feature selection (CFS). They showed that CFS consistently outperformed the other feature selection methods. Furthermore, CFS yielded comparable or superior results to wrapper selection in specific cases [17].

Combining classification methods and feature selection techniques requires further investigation to enhance prediction quality. The study introduces a classification algorithm that integrates feature selection methods to forecast student dropout. We utilized two feature selection methods: CFS and SU. The dataset consisted of 2013 academic data from undergraduate students at Sebelas Maret University Indonesia, obtained from the academic information system. Of the 2,476 data entries, 2,267 were for students who did not drop out, whereas the remaining entries were for students who did drop out. This dataset had an unbalanced class distribution. To address the class imbalance, we employed the synthetic minority oversampling technique (SMOTE). This research offers insights into the efficacy of two feature selection methods in identifying the most influential attributes related to students' dropout tendencies. This information is crucial for evaluating and preventing student dropout tendencies. The approach adopted to achieve this goal followed two stages:

First, to construct the classification model, we applied SMOTE to address the issue of class imbalance within the dataset. Our objective was to enable the classification model to learn from the dropout class data with an uneven number of instances.

Second, using feature selection, we identified the most influential attributes in student dropout prediction. In addition, we evaluated the performance of both feature selection methods to support at-risk students and facilitate further decision-making regarding students displaying dropout potential. To achieve these goals, two feature selection methods, namely CFS and SU, were implemented in three classification algorithms: DT, support vector machine (SVM), and NB.

Based on a series of processes, we aimed to answer two specific research questions:

RQ1: How does the performance of the feature selection methods (CFS and SU) compare in predicting dropout within classification algorithms?
RQ2: Which features influence dropout based on the attributes selected by the two feature selection methods?

This article is organized as follows: the methodology is presented in Section 2, and the results are discussed in Section 3. Section 4 includes the conclusions and outlines future research directions.

This section discusses the previous applications of the selected algorithms, namely Decision Tree, Support Vector Machine, and Naive Bayes, as well as feature selection methods, namely Correlation-based Feature Selection and Symmetrical Uncertainty.

DT classifiers for dropout prediction: Many studies in EDM have focused on using DTs to predict students' dropout risk. Limsathitwong et al. [22] developed a dropout prediction system based on students' performance across various subjects, demonstrating that the DT could effectively identify students at risk of dropping out and assist them in improving their learning processes. Furthermore, Iqbal et al. [23] employed machine learning techniques to predict students' grades using 17 attributes categorized into four groups: gender, family-related information, educational and personal details, and academic performance. They showed that the DT algorithm yielded the highest accuracy, ranging from 95% to 100%.

Gil et al. [24] researched influential factors in dropout cases. Their study, which utilized a dataset containing

academic and demographic details, revealed that the C4.5 DT model achieved an accuracy of 98.9%. Meanwhile, Roslan et al. [25] compared the DT and logistic regression models. Their findings indicated that DT classification outperformed, achieving 89.49% accuracy with an 80/20 data split. Table 1 presents the results of studies that used Decision Trees for dropout prediction.

Table 1. Research relating to DT Classifiers

Author	Research Result
Limsathitwong et al. (2018) [22]	The DT can accurately predict student dropout and identify those who need special attention.
Iqbal et al. (2022) [23]	The DT was found to have the best performance, followed by DT regression and linear regression.
Gil et al. (2020) [24]	The DT produces the best performance, and the model can identify critical factors that cause dropout.
Roslan et al. (2021) [25]	Implementing the DT with the preprocessing method offers a high level of accuracy (89.49%).

SVM classifiers for dropout prediction: Cardona and Cudney [26] conducted SVM research to predict college graduation. SVM classifies input variables into expected classes, namely passing and non-passing, by maximizing the distance between points for different classes while limiting classification errors. In order to reduce the number of students dropping out of higher education, Lee et al. [27] investigated dropout prediction. The findings indicate that the SVM algorithm can forecast student dropout with high accuracy.

Lottering et al. [28] built a dropout prediction system using EDM. The evaluation results of five models trained to predict dropout indicated that the SVM algorithm performed the best based on the original dataset, achieving the highest score among other classification algorithms. Burman and Som [29] predicted student performance using linear and radial basis kernels for SVM classification. The radial basis kernel produced better performance, with an accuracy of 90.97%. Table 2 lists the research findings obtained using SVM for student dropout prediction.

Table 2. Research Related to SVM Classifiers

Author	Research Result
Cardona and Cudney (2019) [26]	This study illustrates the use of SVM in predicting graduation. The model's results indicate a strong performance.
Lee et al. (2020) [27]	SVM performed well, with 96.2% accuracy.
Lottering et al. (2020) [28]	The best-performing algorithm with the original data set is SVM.
Burman and Som (2019) [29]	SVM with a radial basis function (RBF) produces better results compared to linear kernels.

NB on dropout prediction: Nuankaew et al. [30] employed NB, ANN, and DT to develop a dropout prediction model. They assessed the model's performance using a confusion matrix, including recall, accuracy, and precision metrics. The results showed that the NB algorithm achieved an accuracy of 91.68% in the developed model. Tripathi et al. [31] used the NB algorithm to predict student performance and compared the accuracy and execution time of the proposed model with those of the existing SVM models. The results demonstrated that the proposed model outperformed the existing models in accuracy. Further, Triayudi and Widyarto [32] also looked at how well the NB and J48 algorithms predicted student performance and found that the NB algorithm did the best based on F-measure, recall, precision, and the number of correctly classified instances.

According to Saifudin et al. [33], the NB algorithm and feature selection can predict students who have the potential to drop out and identify influential attributes. They used NB to predict student performance and improved the model's performance by employing forward selection. Using the NB algorithm and feature selection, we identified the factors that cause students to experience difficulties in completing their education. Table 3 shows the research results obtained using NB for student dropout prediction.

Table 3. Research Related to NB

Author	Research Result
Nuankaew et al. (2020) [30]	The NB model has the highest accuracy for students in the 2012–2016 academic year
Tripathi et al. (2019) [31]	The proposed model has high accuracy and a low execution time compared to existing models.
Triayudi et al. (2021) [32]	Both the J48 and NB algorithms have high accuracy (>70%), but NB has the highest accuracy and the highest number of correctly classified cases.
Saifudin et al. (2020) [33]	NB's performance using forward selection increased from 85.56% to 94.43%.

Implementation of CFS in EDM: CFS can improve accuracy and model performance. Febro et al. [34] examined the factors influencing the success of first-year students, highlighting their impact on academic performance. The results were more accurate, reaching 92.09%. In addition, using feature selection techniques, they revealed that post-acceptance variables were the dominant predictors. Ghareeb et al. [35] used CFS to obtain the final feature set because it provides excellent results for student performance datasets.

Nuanmeesri et al. [21] reported that the CFS method provides better model performance results than the gain ratio and chi-square methods. CFS had the highest accuracy in attribute selection. Based on the research findings, the following factors influenced student dropout: grade point average (GPA), cumulative GPA (CGPA) for nonfaculty subjects (IPKnf), and participation in social media groups within subjects (social class). Alturki et al. [36] found that the correlated features for predicting academic performance using CFS included students' GPA during the first four semesters, course grades, and the number of courses that failed during the first four semesters. Table 4 shows the research results obtained using CFS in EDM.

Table 4. Research relating to CFS in EDM

Author	Research Result
Febro et al. (2019) [34]	After implementing feature selection, prediction accuracy improved.
Ghareeb et al. (2022) [35]	The student performance dataset was enhanced by CFS.
Nuanmeesri et al. (2022) [21]	Feature selection can improve the efficiency of neural network models in predicting student dropout during the COVID-19 pandemic.
Alturki et al. (2021) [36]	CFS can select significantly correlated attributes to predict students' academic performance.

Implementation of SU in EDM: Bhimavarapu [20] used SU feature selection to identify student-related features that help assess student performance, and applied hybrid deep learning. The proposed model achieved an accuracy of 98.80% in predicting student performance in online classes. Hammoodi et al. [37] used SU for feature selection and demonstrated that the selected features were adequate for predicting student graduation. Almalki [38] and Hussain et al. [19] employed SU and various feature selection methods to evaluate the performance of educational data using WEKA. The previous findings on the implementation of SU in EDM are listed in Table 5.

Table 5. Research relating to SU in EDM

Author	Research Result
Bhimavarapu (2023) [20]	The proposed deep learning model effectively predicts student performance in online classes.
Hammoodi et al. (2022) [37]	The model's accuracy increases after feature selection.
Almalki (2021) [38]	SU shows better results than the feature selection algorithm tested.
Hussain et al. (2018) [19]	The SU implementation enables selecting those with high influence and increased accuracy.

2. Research Method

This research implements a filter-based feature selection method using the CFS and SU algorithms for student dropout classification. The method comprises six stages: data collection, preprocessing, feature selection, data sampling, classification model, and evaluation of results. The overall research steps are shown in Figure 1.

Data collection: This study was conducted with strict adherence to ethical guidelines for research involving student data. All personal identifiers were removed from the dataset prior to analysis to ensure student privacy and confidentiality. The data was anonymized and aggregated, with no possibility of being traced back to individual students.

The dataset used in this study pertains to undergraduate students at Sebelas Maret University in 2013. The dataset encompasses a range of student information, including gender, place of birth, GPA during the first four semesters, CGPA, the number of credits undertaken, and other related details. Importantly, it includes information regarding students' dropout status, which is determined based on whether a student has enrolled for more than seven years without a recorded graduation date.

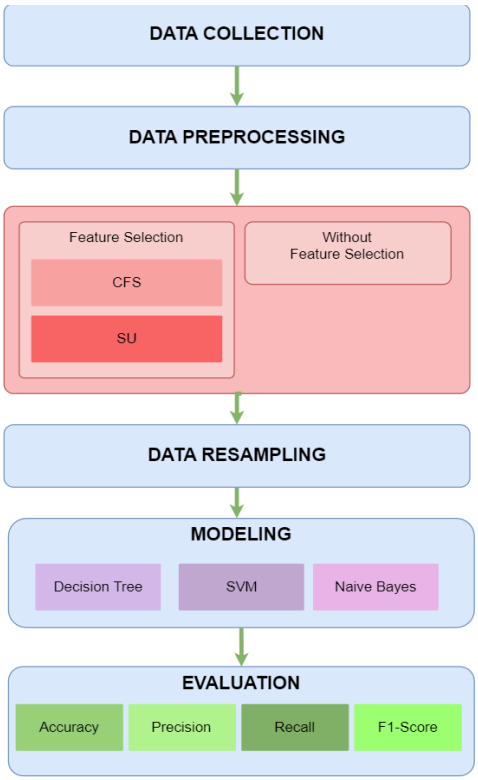

Fig. 1. Research flow diagram

Data preprocessing: The preprocessing stage transforms raw data into a suitable format for effective data mining analysis. During this stage, data cleaning is conducted to rectify noisy data, address missing values, eliminate duplicate data, and correct data with incorrect formats. A dataset with a specific set of features can be transformed into one with a different set of attributes of the same or different types.

This process contributes to acquiring high-quality training data, enhancing detection accuracy, and creating a more efficient training process [39]. We converted several features from numerical to categorical values during the data transformation stage.

Feature Selection: Feature selection is performed to identify the required attributes for data mining. It makes the dataset more efficient and captures a subset of useful features [40]. Each feature selection uses multiple features, with the highest value from each method based on percentages, namely 25%, 50%, and 75% of features, to avoid tying them to a specific number. The percentage-based feature selection method enables a more comprehensive evaluation of performance compared to other threshold techniques, such as selecting the top n features [41].

CFS: CFS is a filter-based feature selection that ranks features based on correlation. This method evaluates the value of a subset of attributes by considering each feature's predictive ability and its level of redundancy [42]. It calculates the correlation between each attribute

and the class variable, selects attributes with moderate-to-high correlation values (close to 1), and deletes attributes with low correlation values (close to 0) [43]. CFS is employed based on Equation 1:

$$Merit_s = \frac{kr_{zf}}{\sqrt{k + k(k-1)\,r_{ff}}} \quad (1)$$

$Merit_s$ is a feature's contribution value in predicting prediction results, K is the number of subset features in the dataset, r_{zf} is the correlation between features and class variables, and r_{ff} is the intercorrelation of subset features.

SU: The SU method is an entropy-based feature selection method. SU is a derivative form of information gain, which is a method used to compensate for bias in information gain. Information gain bias occurs because features with many possible values are preferred. Thus, features with many possible values tend to have higher information gain compared to those with few possible values. SU is applied to compensate for information gain bias and normalize it to a range between 0 and 1 [44]. SU is used to measure the relevance of independent feature classes.

$$SU(A) = 2 \times \frac{infoGain(D,A)}{I(D) + I(A)} \quad (2)$$

In Equation 2, the variable info Gain (D, A) represents the feature information gain value, I(D) is the feature entropy, and I(A) represents the class entropy.

SMOTE: SMOTE is a resampling method used to handle imbalanced data by increasing minority class representation by creating artificial (synthetic) data from k-nearest neighbors [45]. The SMOTE can generate new artificial data between the existing minority and nearest-neighbor samples. These artificial data help increase the representation of the minority class and bring its distribution closer to that of the majority class.

Algorithms and FS in previous studies: To contextualize our methodological choices, we conducted a comprehensive review of recent dropout prediction studies in higher education. Table 6 summarizes key aspects of these studies, including the machine learning algorithms employed, feature selection methods, number of variables, and dataset sizes.

In the context of educational data mining, feature selection plays a crucial role in enhancing the performance of machine learning algorithms for dropout prediction. A study by Febro [34] demonstrated that using CFS improved the accuracy of various classification models by identifying the most relevant features. This approach reduced the dimensionality of the dataset from 29 variables to the most predictive subset, resulting in a more efficient and interpretable model. Similarly, Bhimavarapu [20] employed SU to enhance a deep learning model, achieving an impressive accuracy of 98.80%. These findings underscore the importance of feature selection in EDM, as it helps in eliminating irrelevant and redundant features, thereby improving the predictive performance of machine learning algorithms.

Table 6: Comparison of Previous Dropout Prediction Studies

Study	ML Algorithms	Feature Selection	number of variables	Dataset Size
Febro (2019) [34]	Various	CFS, Gain Ratio, Chi Square	29	7,936
Ghareeb et al. (2022) [35]	RFC, ANN	CFS	25	1550
Nuanmeesri et al. (2022) [21]	LR, DT, RF, NB, SVM, Multilayer Perceptron Neural Network	CFS, Gain Ratio, Chi-square	16	1650
Alturki et al. (2021) [36]	J48, Simple Cart, LADTree, NB, RF	Search-Based, Correlation Based, Information Gain Based	18	300
Bhimavarapu (2023) [20]	Deep Learning	SU	22	32,593
Lottering et al. (2020) [28]	SVM, NB, DT, KNN, RF	none	19	4417
Our Study	DT, SVM, NB	CFS, SU	34	2,463

The implementation of advanced machine learning algorithms has significantly improved dropout prediction models. Lottering et al. [28] demonstrated that SVM outperformed other classifiers such as decision trees and naive bayes in handling high-dimensional educational datasets. Their study showed that SVM, combined with feature selection techniques, achieved the highest accuracy of 99.31%, proving its effectiveness in capturing complex patterns within the data. Additionally, Nuanmeesri et al. [21] highlighted the superior performance of neural network models when integrated with multiple feature selection methods, including CFS, gain ratio, and chi-square. This integration resulted in more robust and accurate predictions of student dropout, emphasizing the need for advanced algorithms and comprehensive feature selection in EDM.

Modeling: The preprocessing data includes a set of attributes that serve as the input. Classification model testing uses DT, SVM, and NB based on training data and test data formed by the K-fold cross-validation (tenfold) process. In K-fold cross-validation, we *randomly* divide the dataset into multiple partitions and perform data processing k times. K-fold cross-validation with k = 10 provides a more stable estimate because it allows the data to be used as training data and test data as much as 90% and 10% of the total data, respectively [46].

DT: DT is a classification algorithm that models the classification process via hierarchical decisions organized in a tree structure. This algorithm offers easily interpretable, flexible, and visualizable DTs [47]. Information gain metrics are critical when selecting the testing attributes for each node in the DT algorithm [42]. The modelling begins with preparing training data and determining the root attribute based on gain value calculations. The entropy value is needed to calculate the gain value using Equation 3.

$$I(S_A) = \sum_{i=1}^{k} -P_i log_2(P_i) \quad (3)$$

S is the set of cases in attribute A, K is the number of partitions in S, and Pi is the probability of Si regarding S. The entropy value above is then used to calculate the information gain value using Equation 4.

$$Gain\ (S, A) = I(S_A) - \sum_{i=1}^{k} \frac{|Si|}{|S|} \times I(S_i) \quad (4)$$

Using |Si| number of cases in the ith partition and |S| number of cases in S, root determination is reiterated until all data are partitioned.

SVM: SVM uses a high-dimensional feature space for classification. SVM uses a hypothesis space as linear functions in a feature space, trained with a learning algorithm based on optimization theory by implementing learning bias derived from statistical learning theory [48]. Its concept involves attempting to find the best multiple paths or decision boundaries that separate two classes. This optimal boundary or region is known as the hyperplane.

By using the kernel concept, SVM can be applied to both linear and nonlinear data. Processing data with numerous features can become complex when handled linearly, which can be resolved using kernel functions to transform the data into a higher-dimensional space [49]. Choosing the proper kernel function is crucial because it determines the appropriate feature space to produce classification with the best accuracy. The recommended kernel to test first is RBF, which offers the same efficiency as a linear kernel and exhibits behaviour similar to a sigmoid kernel function with other parameters[50].

NB: NB is a probabilistic classification algorithm developed based on Bayes decision theory. This algorithm uses the theory to calculate the conditional probability of a class variable, taking into account the observed values of other feature variables [51]. The NB classifier describes training assuming that features do not depend on specific classes. The general form of Bayes' theorem is described using Equation 5.

$$P(C_k \mid X_i) = \frac{P(C_k) \times P(X_i \mid C_k)}{P(X_i)} \quad (5)$$

In Equation 5, C represents the class to be predicted, and Xi represents the attributes used in classification with i = 1, 2,..., n. Equation (5) explains that the probability that a sample is included in class C (posterior) is calculated by multiplying the probability of the appearance of class C before observing the sample (prior) by the probability of the appearance of the sample attributes in class C (likelihood), which is then divided by the probability of the occurrence of all sample attributes (evidence).

In NB, a Gaussian distribution approach calculates conditional probability or likelihood values. Thus, Equation 6 calculates the likelihood with a Gaussian distribution.

$$P(X_i \mid C_k) = \frac{1}{\sqrt{2\pi\ \sigma_{ik}{}^2}} exp\left(-\frac{1}{2}\left(\frac{(X_i - \mu_{ik})^2}{\sigma_{ik}{}^2}\right)\right) \quad (6)$$

The variable μ represents the average parameter, whereas σ^2 represents the variance.

Evaluation: The model was analyzed using the confusion matrix method. The confusion matrix is a table used to assess the performance of classification models in data mining. The binary classification confusion matrix table is a 2-by-2 table formed by calculating four results from the binary classifier [52]. The evaluation of the confusion matrix model is shown in Table 7.

Table 7. Evaluation using the confusion matrix

	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

The evaluation matrix is used to calculate the classification model's accuracy, precision, recall, and F1 score values, as respectively shown in Equations 7–10.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\% \quad (7)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \quad (8)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \quad (9)$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

3. Results and Discussion

3.1. Result

This research involves three experimental scenarios, starting with the data collection and preprocessing stages. The first scenario used a dataset without feature selection; the second scenario used a dataset after preprocessing with attribute selection based on the CFS method; and the third scenario used a dataset with attributes selected using SU. Each scenario was tested using the DT, SVM, and NB classification algorithms.

The dataset used belonged to the 2013 undergraduate student academic data at Sebelas Maret University, obtained from the academic information system. The

data were obtained from the Information and Communication Technology Management Unit of Sebelas Maret University and have not been processed. We had 2.476 rows of data comprising 65 features and one target, with two classes: not dropping out and dropping out. Table 8 shows a list of attributes in the data collection.

Table 8. Data collection

GPA 1	Gender	…	Home Status	Target
3.1	1	…	1	Not Dropout
3.08	0	…	5	Not Dropout
2.95	0	…	1	Not Dropout
…	…	…	…	…
2.7	1	…	1	Dropout

Data cleaning was conducted in four stages, including cleaning features that do not provide information, cleaning features that duplicate information, mapping the features used, and cleaning data rows with missing values. Cleaning missing values was conducted by deleting rows containing missing values in the dataset. The final data cleaning process resulted in 2.463 data rows and 34 features.

Data transformation is the process of converting data from a nominal to a categorical format by establishing a range. This process involves grouping data values into defined categories based on specific ranges of values. The results of changing numerical data to categorical data are shown in Table 9.

Table 9. Data Transformation Results

Attribute	Value	Description
Total credits per 4th semester	{1,2,3,4,5}	1 = 0–20 credits 2 = 21–40 credits 3 = 41–60 credits 4 = 61–80 credits 5 = 81–96 credits
Average high school final score	{1,2,3,4,5}	1 = 0–4.50 2 = 4.50–5.50 3 = 5.50–6.50 4 = 6.50–7.50 5 = >7.50
GPA 1–4	{1,2,3,4,5}	1 = 0–1.50 2 = 1.51–2.00 3 = 2.01–2.50 4 = 2.51–3.50 5 = 3.51–4.00

In the initial stage of CFS, the average correlation between each feature and the target variable was calculated using the Pearson correlation coefficient as a measurement tool. Furthermore, in merit calculations, the average correlation between features in a subset was calculated using the Pearson correlation coefficient. Then, these features were sorted from the feature with the highest merit to the feature with the lowest merit. The number of features to be taken in this stage was determined based on the top-ranking order of all features. The percentage levels used as a reference in feature retrieval were 25%, 50%, and 75%. Table 10 displays the features grouped based on the percentage of feature retrieval.

Table 10. Selected Features from CFS and Their Retrieval Percentages

Percentage	Number of Features	Selected Features
25%	8	Total credits per 4th semester, 4th semester GPA, 2nd semester GPA, 3rd semester GPA, gender, 1st semester GPA, Final high school score, student's district
50%	17	Total credits per 4th semester, 4th semester GPA, 2nd semester GPA, 3rd semester GPA, gender, 1st semester GPA, Final high school score, Student's District, Hobbies, Student's Province, Admission Path, Parent's District, High School Major, Father's Income, Father's Educational Background, Student Activities, Achievements
75%	25	Total credits per 4th semester, 4th semester GPA, 2nd semester GPA, 3rd semester GPA, gender, 1st semester GPA, Final high school score, Student District, Hobbies, Student Province, Admission Path, Parent's District, High School Major, Father's Income, Father's Educational Background, Student Activities, Achievements, Parent's Province, Religion, Mother's Educational Background, Father's Occupation, Faculty, Citizenship, Department, Scholarship

For feature selection using SU, first, an entropy calculation was performed for each feature within the dataset. Subsequently, information gain was calculated to assess each feature's relevance to the classification. Then, the results obtained from the entropy and gain calculations were utilized to compute the SU value using Equation 2. The selection of the number of features that serve as samples in the training model was based on three distinct percentage levels: 25%, 50%, and 75%. Table 10 presents the categorization of features based on these different percentages.

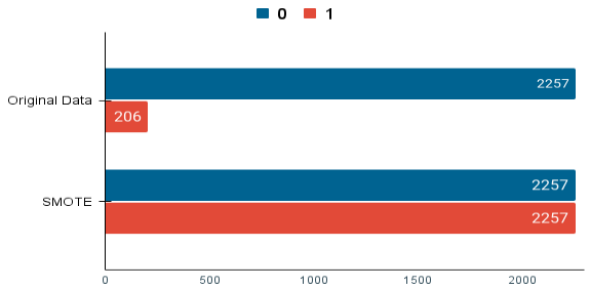

Fig. 2. Comparison between the original dataset and the dataset after applying the SMOTE

Preprocessed datasets exhibit a notable imbalance between the majority and minority classes. The SMOTE was employed to address this issue as the dataset required a more equitable data distribution. Then, the resulting synthetic data was added to the dataset as new data. Creating synthetic data was conducted on student data from class 1 (dropout) until the minority class data were balanced with class 0 data

(no dropout). A comparison of the original dataset compared to the data after undergoing the SMOTE is shown in Figure 2.

The results of implementing the DT algorithm, which was processed using tenfold cross-validation, are shown in Table 11.

Table 11. Results of DT Implementation

Feature selection	Percentage	Number of features	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)
Without Feature Selection	-	34	90.94	88.65	93.87	91.17
CFS	25%	8	87.11	84.71	90.71	87.53
	50%	17	90.61	88.06	93.94	90.88
	75%	25	90.83	88.23	94.22	91.11
SU	25%	8	83.50	84.18	83.50	84.18
	50%	17	91.29	91.55	91.29	91.55
	75%	25	90.83	91.08	90.83	91.08

Classification using a DT with CFS feature selection cannot improve performance. Meanwhile, the DT with SU feature selection enhanced feature retrieval performance by 50%, with an accuracy of 91.29% and an F1 score of 91.55%.

The following classification algorithm uses an SVM with RBF. Data division in SVM testing uses tenfold cross-validation to avoid overfitting. The accuracy, precision, recall, and F1 score results are shown in Table 12.

Table 12. SVM Implementation Results

Feature selection	Percentage	Number of features	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)
without feature selection	-	34	97.94	99.68	96.19	97.90
Figure 4	25%	8	87.70	86.52	89.37	87.87
	50%	17	94.55	94.04	95.16	94.59
	75%	25	98.01	99.24	96.76	97.98
SU	25%	8	82.94	81.46	85.57	83.35
	50%	17	95.59	95.21	96.03	95.62
	75%	25	98.16	99.10	97.23	98.14

Table 12 shows the results of SVM testing using CFS at a percentage of 75%; the accuracy and F1 score increased to 98.01% and 97.98%, respectively. In addition, SVM testing using SU showed a 75% performance enhancement with an accuracy of 98.16% and an F1 score of 98.14%. The dropout classification results obtained using the NB algorithm, processed using tenfold cross-validation, are shown in Table 13.

Table 13. Results of NB Implementation

Feature selection	Percentage	Number of features	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)
Without feature selection	-	34	74.70	70.08	86.20	77.31
CFS	25%	8	69.12	82.50	48.59	61.14
	50%	17	69.96	82.42	50.86	62.85
	75%	25	76.47	80.53	69.89	74.76
SU	25%	8	68.19	81.56	47.10	59.67
	50%	17	68.39	80.01	49.18	60.87
	75%	25	75.99	77.91	72.57	75.11

CFS in the NB model produced higher accuracy (76.47%) than that obtained using all the features. Meanwhile, the implementation of SU increased performance at the 75% percentile, with accuracy reaching 75.99%. Although accuracy increased with CFS and SU, the F1 score did not significantly improve.

3.2. Discussion

This section presents the results obtained based on (1) predicting dropout using classifiers and (2) identifying features that influence dropout prediction.

The evaluation results of each model without a feature selection process are illustrated in Figure 3. In this research, we compared the performance of the DT, SVM, and NB algorithms using 34 features (without feature selection), with the dropout status attribute set as the target class. The SVM with RBF achieved the best performance, with 97.94% accuracy, 99.68% precision, 96.19% recall, and a 97.90% F1 score. Meanwhile, the NB model performed the least favourably, with an accuracy rate of 74.70%.

Fig. 3. Comparison of the model evaluation of all features

After applying the three models to predict student dropout using the features selected via CFS and SU, the performance results were obtained, as shown in Table 14.

Table 14. Performance of the Algorithms Combined with Feature Selection Methods

Model	Accuracy	Precision	Recall	F1 Score
DT + CFS (25%)	87.11%	84.71%	90.71%	87.53%
DT + SU (25%)	83.50%	80.88%	88.02%	84.18%
DT + CFS (50%)	90.61%	88.06%	93.94%	90.88%
DT + SU (50%)	91.29%	88.85%	94.44%	91.55%
DT + CFS (75%)	90.83%	88.23%	94.22%	91.11%
DT + SU (75%)	90.83%	88.56%	93.76%	91.08%
SVM + CFS (25%)	87.70%	86.52%	89.37%	87.87%
SVM + SU (25%)	82.94%	81.46%	85.57%	83.35%
SVM + CFS (50%)	94.55%	94.04%	95.16%	94.59%
SVM + SU (50%)	95.59%	95.21%	96.03%	95.62%
SVM + CFS (75%)	98.01%	99.24%	96.76%	97.98%
SVM + SU (75%)	98.16%	99.09%	97.23%	98.14%
NB + CFS (25%)	69.12%	82.50%	48.59%	61.14%
NB + SU (25%)	68.19%	81.56%	47.10%	59.67%
NB + CFS (50%)	69.96%	82.42%	50.860%	62.850%
NB + SU (50%)	68.39%	80.01%	49.18%	60.87%
NB + CFS (75%)	76.47%	80.53%	69.89%	74.76%
NB + SU (75%)	75.99%	77.91%	72.57%	75.11%

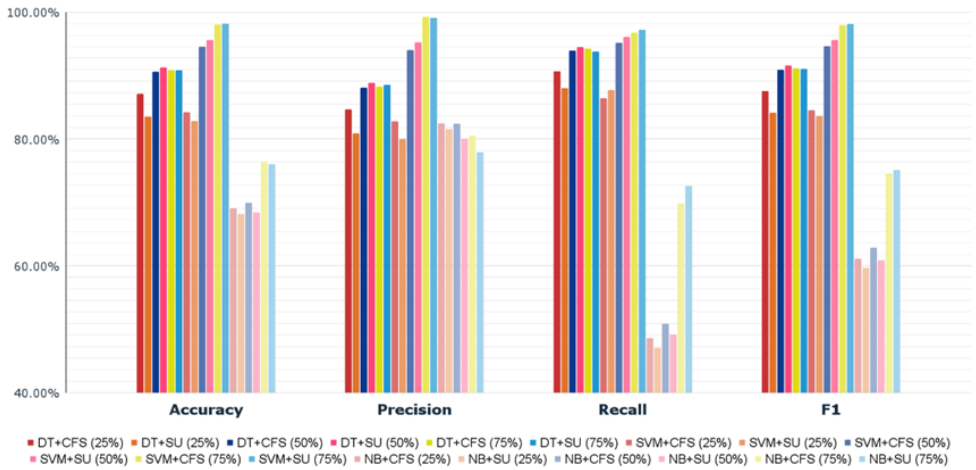

Fig. 4. Performance of the algorithms combined with feature selection methods

Based on the comparison results shown in Table 14 and Figure 4, SU provided increased accuracy in all classification algorithms applied in this research. The DT algorithm's use of CFS decreased matrix accuracy, precision, recall, and F1 score performance. Meanwhile, the DT algorithm, which used 17 relevant features based on SU, further improved the accuracy, precision, and F1 scores compared with the models built without feature selection. The evaluation results obtained using NB showed that the model significantly increased matrix precision and accuracy. However, after using CFS and SU, the recall and F1 scores decreased. The increased accuracy and precision indicate that the model is more likely to provide correct predictions (both positive and negative) and be better at correctly identifying positives, but it does contain some cases that should be predicted to be positive.

Incorporating feature selection techniques into the DT and NB models improves accuracy; however, our evaluations reveal that SVM outperforms the other algorithms. The SVM + SU model with 25 features does the best job of predicting dropout when comparing the CFS and SU feature selection models across the three classification algorithms. This model has the highest accuracy, recall, and F1 scores. Alternatively, the SVM + CFS model with 25 features has the highest precision value, demonstrating the accurate identification of positive classes.

In addition to improving prediction accuracy, comparing two feature selection techniques can provide insights into which attributes significantly impact a student's potential to drop out. Implementing CFS and

SU's dropout student data results in similar rankings of the top features. However, there is a difference between "Gender" and "GPA 1st Semester," which appear in fifth and sixth place in both feature selections. However, CFS and SU have the same top features: total credits per 4th semester, gender, and GPA from 1st to 4th semester.

In addition, the domicile of origin is vital in student dropout prediction, as can be seen from the same ranking for the features "student's district" and "student's province" in both methods. The feature selection computing results are sorted from the highest to the lowest value, as listed in Table 15.

Table 15. List of Feature Importance Based on the Feature Selection Method

Rank	CFS	SU
1	Total credits per 4th semester	Total credits per 4th semeste
2	4th semester GPA	4th semester GPA
3	2nd semester GPA	2nd semester GPA
4	3rd semester GPA	3rd semester GPA
5	Gender	1st semester GPA
6	1st Semester GPA	Gender
7	Final high school score	Parent's District
8	Student District	Student's District
9	Hobbies	Department
10	Student Province	Student's Province
11	Admission Path	Parent's Province
12	Parent's District	Final high school score
13	High School Major	Scholarship
14	Father's Income	Admission path
15	Father's Educational Background	Achievements
16	Student Activities	Father's Income
17	Achievements	High School Major
18	Parent's Province	Religion
19	Religion	Citizenship
20	Mother's Educational Background	Faculty
21	Father's Occupation	Hobbies
22	Faculty	Mother's Occupation
23	Citizenship	Student Activities
24	Department	Mother's Educational Background
25	Scholarship	Home Status
26	Source of costs	Father's occupation
27	Home Status	Father's Educational Background
28	Scholarship provider	Source of costs
29	Mother's income	Mother's income
30	Mastery of foreign texts	Marital status
31	Marital status	Scholarship provider
32	Mother's occupation	Mastery of foreign texts
33	Suffering from illness	Suffering from illness
34	Gap year	Gap year

The CFS and SU methods produce the nine lowest-ranking features, with similar results. Some of the features eliminated by both methods include the source of costs, mother's income, marital status, scholarship provider, mastery of foreign texts, suffering from illness, and gap year. Besides these similarities, they also exhibited some differences.

For example, CFS excludes home status and the mother's occupation, whereas SU excludes the father's occupation and educational background. The comparison of feature selection results (Table 15) shows that the two feature selection methods have slightly different perspectives on evaluating the relevance of features in the context of student dropout prediction. These differences can affect the performance of the models built based on the selected features.

The results of the CFS and SU methods employed in the DT, SVM, and NB algorithms provide significant comparisons with previous research. Preprocessing revealed data imbalances, particularly in students at risk of dropout compared to non-drops. To address this, we applied the SMOTE, as reported by Flores et al. [53], to enhance dropout prediction. In both studies, the NB classifier and tree-based algorithms were used. However, in our study and the work conducted by Flores et al. [53], these algorithms did not consistently outperform the proposed algorithm in all evaluations.

This research used a dataset containing personal information, academic track records, and family-related information. To determine whether total credits and GPA had a significant effect on dropout prediction, we used CFS and SU. Notably, the importance of total credits as a predictive feature aligns with prior research conducted by Bedregal-Alpaca et al. [54] using academic data and the DT algorithm. The difference is that there is a feature selection process to select relevant features and optimize prediction performance. Implementing SU in the DT produces an accuracy of 91.29%, higher than the previously reported accuracy of 83.19%.

The findings about how important GPA is in predicting student dropout are similar to those of Febro et al. [34]. They discovered that GPA is very important in predictions made using three feature selection methods: CFS, information gain, and chi-square. These findings confirm our research that GPA, or aspects of academic grades, are very relevant in predicting dropout, as identified by Febro et al. [34].

The CFS and SU feature selection results show that the SVM model works better than the others, which backs up what Lottering et al. [28]. However, our research demonstrates the performance enhancement of SVM via feature selection, unlike the study conducted by Lottering et al. [28], which showed no accuracy improvement following feature selection. In this study, the SVM + SU model achieved the highest accuracy of 98.16%, whereas in the previous research, the accuracy reached 99.31%.

In this study, implementing CFS did not increase the DT's accuracy. This decrease is due to the DT's focus on finding key features when dividing nodes in a tree. However, when applying feature selection to the DT, the selection method may eliminate essential features for accurate classification [55].

*T*o provide a clear overview of the most influential features identified by each model combination, we present Table 16. This table showcases the top five features selected by each algorithm and feature selection method pairing.

Table 16: Top 5 Features Identified by Different Model Combinations

Model	Top 5 Feature
DT + CFS	1. Total credits per 4th semester
	2. 4th semester GPA
	3. 2nd semester GPA
	4. 3rd semester GPA
	5. Gender
DT + SU	1. Total credits per 4th semester
	2. 4th semester GPA
	3. 2nd semester GPA
	4. 3rd semester GPA
	5. 1st semester GPA
SVM + CFS	1. Total credits per 4th semester
	2. 4th semester GPA
	3. Gender
	4. Student's District
	5. Final high school score
SVM + SU	1. Total credits per 4th semester
	2. 4th semester GPA
	3. 2nd semester GPA
	4. 3rd semester GPA
	5. Parent's District
NB + CFS	1. Total credits per 4th semester
	2. 4th semester GPA
	3. 2nd semester GPA
	4. 3rd semester GPA
	5. Student's Province
NB + SU	1. Total credits per 4th semester
	2. 4th semester GPA
	3. 2nd semester GPA
	4. Department
	5. Student's District

Analyzing Table 14 reveals consistent patterns across different model combinations. We unanimously identify the total credits per 4th semester and the 4th semester GPA as the two most influential features, aligning with findings from previous studies [56] [57]. This underscores the critical role of academic performance, particularly in later semesters, in predicting dropout risk. Interestingly, geographical factors such as the student's district or province appear in several models, suggesting that socio-geographic background may play a significant role in dropout prediction, a finding echoed in recent literature [58]. Some models also include gender, which warrants further investigation into potential gender-based disparities in dropout risk [59].

Moreover, the appearance of 'Department' in the NB + SU model aligns with findings from Hu et al. [60], who

identified course-specific factors as significant predictors of student performance and retention. The consistency of GPA across semesters as a top feature supports the findings of Fernandes-Garcia et al. [61], who emphasized the importance of continuous academic monitoring in dropout prevention strategies.

It's noteworthy that our models identified 'Final high school score' as an influential feature, corroborating the findings of Gafarov et al. [62], who found pre-university academic performance to be a significant predictor of university success and retention. This suggests that early interventions, even before university enrollment, could be crucial in mitigating dropout risk.

4. Conclusions

We examined two research questions to help higher education institutions make decisions about student dropout. The first question concerns the efficacy of filter-based feature selection methods (CFS and SU) in classifying student attrition risks using academic data. The findings indicate that both CFS and SU significantly impact dropout classification performance. The application of SVM and DT algorithms reveals that SU offers a more substantial contribution than CFS. In the DT, SU improves the prediction accuracy to 91.29%, showing a significant enhancement. Meanwhile, SVM yields the best results with SU, boasting a remarkable accuracy of 98.16% and an F1 score of 98.14%. Using CFS produces better performance on NB than SU, with an accuracy of 76.47%. However, this increment in accuracy does not coincide with a proportional improvement in the F1 score, underscoring the need to optimize the trade-off between precision and recall. The second question identifies what attributes influence undergraduate students' dropout propensity. Applying CFS and SU can be essential in selecting high-relevance features for predicting potential student attrition. We revealed many factors that wield substantial influence in elucidating students' predisposition to discontinue their studies, including total semester IV credits, GPA from the 1st to the 4th semester, gender, and student domicile. This study highlights the significance of the number of features selected within the CFS and SU methods, as it can substantially influence model classification performance. Taking too few or too many features can reduce these selection methods' classification ability. Future research may investigate hybrid approaches to optimize feature selection. Particle swarm optimization or genetic algorithms can help identify the best attribute combination. Moreover, incorporating optimization parameters into the classification model becomes imperative to improve accuracy. Thus, our results can help improve the efficiency and accuracy of the classification models built within the framework of this research.

Acknowledgements

We would like to thank Sebelas Maret University for providing the Research Group Research Grant (Grant-MRG), with Contract Number: 194.2/UN27.22/PT.01.03/2024, so that this research can be carried out.

References

[1] V. Hegde and P. P. Prageeth, "Higher education student dropout prediction and analysis through educational data mining," in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, Coimbatore: IEEE, Jan. 2018, pp. 694–699. doi: 10.1109/ICISC.2018.8398887.

[2] R. W. Rumberger, *Dropping Out: Why Students Drop Out of High School and What Can Be Done About It*. Harvard University Press, 2011. doi: 10.4159/harvard.9780674063167.

[3] A. Akkari, "Education in the Middle East and North Africa," in *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier, 2015, pp. 210–214. doi: 10.1016/B978-0-08-097086-8.92149-4.

[4] A. Behr, M. Giese, H. D. Teguim Kamdjou, and K. Theune, "Dropping out of university: a literature review," *Rev Educ*, vol. 8, no. 2, pp. 614–652, Jun. 2020, doi: 10.1002/rev3.3202.

[5] M. P. Marchbanks *et al.*, "More than a Drop in the Bucket: The Social and Economic Costs of Dropouts and Grade Retentions Associated With Exclusionary Discipline," *Journal of Applied Research on Children: Informing Policy for Children at Risk*, vol. 5, no. 2, Feb. 2015, doi: 10.58464/2155-5834.1226.

[6] E. Arias Ortiz and C. Dehon, "Roads to Success in the Belgian French Community's Higher Education System: Predictors of Dropout and Degree Completion at the Université Libre de Bruxelles," *Res High Educ*, vol. 54, no. 6, pp. 693–723, Sep. 2013, doi: 10.1007/s11162-013-9290-y.

[7] B. Daniel, "Big Data and analytics in higher education: Opportunities and challenges," *Brit J Educational Tech*, vol. 46, no. 5, pp. 904–920, Sep. 2015, doi: 10.1111/bjet.12230.

[8] A. Namoun and A. Alshanqiti, "Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review," *Applied Sciences*, vol. 11, no. 1, p. 237, Dec. 2020, doi: 10.3390/app11010237.

[9] C. S. Lyche, "Taking on the Completion Challenge: A Literature Review on Policies to Prevent Dropout and Early School Leaving," OECD Education Working Papers 53, Nov. 2010. doi: 10.1787/5km4m2t59cmr-en.

[10] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques Third Edition*. Elsevier, 2012.

[11] M. S. P. Babu and S. H. Sastry, "Big data and predictive analytics in ERP systems for automating decision making process," in *2014 IEEE 5th International Conference on Software Engineering and Service Science*, Beijing: IEEE, Jun. 2014, pp. 259–262. doi: 10.1109/ICSESS.2014.6933558.

[12] W.-W. Wu, Y.-T. Lee, M.-L. Tseng, and Y.-H. Chiang, "Data mining for exploring hidden patterns between KM and its performance," *Knowledge-Based Systems*, vol. 23, no. 5, pp. 397–401, Jul. 2010, doi: 10.1016/j.knosys.2010.01.014.

[13] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques," *IJMECS*, vol. 8, no. 11, pp. 36–42, Nov. 2016, doi: 10.5815/ijmecs.2016.11.05.

[14] C. Romero and S. Ventura, "Data mining in education: Data mining in education," *WIREs Data Mining Knowl Discov*, vol. 3, no. 1, pp. 12–27, Jan. 2013, doi: 10.1002/widm.1075.

[15] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," *IEEE Trans. Syst., Man, Cybern. C*, vol. 40, no. 6, pp. 601–618, Nov. 2010, doi: 10.1109/TSMCC.2010.2053532.

[16] M. Alban and D. Mauricio, "Predicting University Dropout trough Data Mining: A systematic Literature," *Indian Journal*

of Science and Technology, vol. 12, no. 4, pp. 1–12, Jan. 2019, doi: 10.17485/ijst/2019/v12i4/139729.

[17] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," Doctoral dissertation, University of Waikato, 1999.

[18] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, Jan. 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[19] S. Hussain, N. Abdulaziz Dahan, F. M. Ba-Alwi, and N. Ribata, "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA," *IJEECS*, vol. 9, no. 2, p. 447, Feb. 2018, doi: 10.11591/ijeecs.v9.i2.pp447-459.

[20] U. Bhimavarapu, "Analysing student performance for online education using the computational models," *Univ Access Inf Soc*, Aug. 2023, doi: 10.1007/s10209-023-01033-7.

[21] S. Nuanmeesri, L. Poomhiran, S. Chopvitayakun, and P. Kadmateekarun, "Improving Dropout Forecasting during the COVID-19 Pandemic through Feature Selection and Multilayer Perceptron Neural Network," *IJIET*, vol. 12, no. 9, pp. 851–857, 2022, doi: 10.18178/ijiet.2022.12.9.1693.

[22] K. Limsathitwong, K. Tiwatthanont, and T. Yatsungnoen, "Dropout prediction system to reduce discontinue study rate of information technology students," in *2018 5th International Conference on Business and Industrial Research (ICBIR)*, Bangkok: IEEE, May 2018, pp. 110–114. doi: 10.1109/ICBIR.2018.8391176.

[23] Sumaiya Iqbal, Mahjabin Muntaha, Jerin Ishrat Natasha, and Dewan Sakib, "Early Grade Prediction Using Profile Data," *IJMLC*, vol. 12, no. 5, Sep. 2022, doi: 10.18178/ijmlc.2022.12.5.1100.

[24] J. S. Gil, "Predicting Students' Dropout Indicators in Public School using Data Mining Approaches," *IJATCSE*, vol. 9, no. 1, pp. 774–778, Feb. 2020, doi: 10.30534/ijatcse/2020/110912020.

[25] Nurhana Roslan, Jastini Mohd Jamil, and I. N. Mohd. Shaharanee, "Prediction of Student Dropout in Malaysian's Private Higher Education Institute using Data Mining Application," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 3, pp. 2326–2334, 2021.

[26] T. A. Cardona and E. A. Cudney, "Predicting Student Retention Using Support Vector Machines," *Procedia Manufacturing*, vol. 39, pp. 1827–1833, 2019, doi: 10.1016/j.promfg.2020.01.256.

[27] L. E. Lee *et al.*, "Evaluation of Prediction Algorithms in the Student Dropout Problem," *JCC*, vol. 08, no. 03, pp. 20–27, 2020, doi: 10.4236/jcc.2020.83002.

[28] R. Lottering, R. Hans, and M. Lall, "A model for the identification of students at risk of dropout at a university of technology," in *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, Durban, South Africa: IEEE, Aug. 2020, pp. 1–8. doi: 10.1109/icABCD49160.2020.9183874.

[29] I. Burman and S. Som, "Predicting Students Academic Performance Using Support Vector Machine," in *2019 Amity International Conference on Artificial Intelligence (AICAI)*, Dubai, United Arab Emirates: IEEE, Feb. 2019, pp. 756–759. doi: 10.1109/AICAI.2019.8701260.

[30] P. Nuankaew, W. Nuankaew, D. Teeraputon, K. Phanniphong, and S. Bussaman, "Prediction Model of Student Achievement in Business Computer Disciplines," *Int. J. Emerg. Technol. Learn.*, vol. 15, no. 20, p. 160, Oct. 2020, doi: 10.3991/ijet.v15i20.15273.

[31] A. Tripathi, S. Yadav, and R. Rajan, "Naïve Bayes Classification Model for the Student Performance Prediction," *2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, 2019, doi: 10.1109/ICICICT46008.2019.8993237.

[32] A. Triayudi and W. O. Widyarto, "Comparison J48 And Naïve Bayes Methods in Educational Analysis," *J. Phys.: Conf. Ser.*, vol. 1933, no. 1, p. 012062, Jun. 2021, doi: 10.1088/1742-6596/1933/1/012062.

[33] A. Saifudin, Ekawati, Yulianti, and T. Desyani, "Forward Selection Technique to Choose the Best Features in Prediction of Student Academic Performance Based on Naïve Bayes," *J. Phys.: Conf. Ser.*, vol. 1477, no. 3, p. 032007, Mar. 2020, doi: 10.1088/1742-6596/1477/3/032007.

[34] J. D. Febro, "Utilizing Feature Selection in Identifying Predicting Factors of Student Retention," *IJACSA*, vol. 10, no. 9, 2019, doi: 10.14569/IJACSA.2019.0100934.

[35] S. Ghareeb *et al.*, "Evaluating student levelling based on machine learning model's performance," *Discov Internet Things*, vol. 2, no. 1, p. 3, Dec. 2022, doi: 10.1007/s43926-022-00023-0.

[36] S. Alturki and N. Alturki, "Using Educational Data Mining to Predict Students' Academic Performance for Applying Early Interventions," *JITE:IIP*, vol. 20, pp. 121–137, 2021, doi: 10.28945/4835.

[37] Mahmood Shakir Hammoodi and Ahmed Al-Azawei, "Using Socio-Demographic Information in Predicting Students' Degree Completion based on a Dynamic Model," *IJIES*, vol. 15, no. 2, pp. 107–115, Apr. 2022, doi: 10.22266/ijies2022.0430.11.

[38] A. J. Almalki, "Accuracy analysis of Educational Data Mining using Feature Selection Algorithm," 2021, doi: 10.48550/ARXIV.2107.10669.

[39] J. Gu, L. Wang, H. Wang, and S. Wang, "A novel approach to intrusion detection using SVM ensemble with feature augmentation," *Computers & Security*, vol. 86, pp. 53–62, Sep. 2019, doi: 10.1016/j.cose.2019.05.022.

[40] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 1–4, pp. 131–156, 1997, doi: 10.1016/S1088-467X(97)00008-5.

[41] R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, and J. H. Moore, "Benchmarking relief-based feature selection methods for bioinformatics data mining," *Journal of Biomedical Informatics*, vol. 85, pp. 168–188, Sep. 2018, doi: 10.1016/j.jbi.2018.07.015.

[42] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative Study of Attribute Selection Using Gain Ratio and Correlation-Based Feature Selection," 2013.

[43] E. C. Blessie and E. Karthikeyan, "Sigmis: A Feature Selection Algorithm Using Correlation Based Method," *Journal of Algorithms & Computational Technology*, vol. 6, no. 3, pp. 385–394, Sep. 2012, doi: 10.1260/1748-3018.6.3.385.

[44] G. Sosa-Cabrera, M. García-Torres, S. Gómez-Guerrero, C. E. Schaerer, and F. Divina, "A multivariate approach to the symmetrical uncertainty measure: Application to feature selection problem," *Information Sciences*, vol. 494, pp. 1–20, Aug. 2019, doi: 10.1016/j.ins.2019.04.046.

[45] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *jair*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[46] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," 1995.

[47] D. T. Larose and C. D. Larose, "DISCOVERING KNOWLEDGE IN DATA An Introduction to Data Mining." John Wiley & Sons, Inc, 2014.

[48] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods." 2000.

[49] M. Zareapoor, P. Shamsolmoali, D. Kumar Jain, H. Wang, and J. Yang, "Kernelized support vector machine with deep learning: An efficient approach for extreme multiclass dataset," *Pattern Recognition Letters*, vol. 115, pp. 4–13, Nov. 2018, doi: 10.1016/j.patrec.2017.09.018.

[50] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification," 2003.

[51] C. C. Aggarwal, *Data Mining: The Textbook*. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-14142-8.

[52] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[53] V. Flores, S. Heras, and V. Julian, "Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education," *Electronics*, vol. 11, no. 3, p. 457, Feb. 2022, doi: 10.3390/electronics11030457.

[54] N. Bedregal-Alpaca, V. Cornejo-Aparicio, J. Zárate-Valderrama, and P. Yanque-Churo, "Classification Models for Determining Types of Academic Risk and Predicting Dropout in University Students," *IJACSA*, vol. 11, no. 1, 2020, doi: 10.14569/IJACSA.2020.0110133.

[55] J. Sadhasivam, V. Muthukumaran, J. Thimmia Raja, R. B. Joseph, M. Munirathanam, and J. M. Balajee, "Diabetes disease prediction using decision tree for feature selection," *J. Phys.: Conf. Ser.*, vol. 1964, no. 6, p. 062116, Jul. 2021, doi: 10.1088/1742-6596/1964/6/062116.

[56] A. Slim, G. L. Heileman, J. Kozlick, and C. T. Abdallah, "Predicting student success based on prior performance," in *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2014, pp. 410–415. doi: 10.1109/CIDM.2014.7008697.

[57] E. Yukselturk, S. Ozekes, and Y. K. Türel, "Predicting dropout student: an application of data mining methods in an online education program," *European Journal of Open, Distance and e-learning*, vol. 17, no. 1, pp. 118–133, 2014.

[58] A. Sarra, L. Fontanella, and S. Di Zio, "Identifying students at risk of academic failure within the educational data mining framework," *Social Indicators Research*, vol. 146, pp. 41–60, 2019.

[59] J. R. Casanova, A. Cervero, J. C. Núñez, L. S. Almeida, and A. Bernardo, "Factors that determine the persistence and dropout of university students," 2018.

[60] Q. Hu, A. Polyzou, G. Karypis, and H. Rangwala, "Enriching Course-Specific Regression Models with Content Features for Grade Prediction." 2017. doi: 10.1109/DSAA.2017.74.

[61] A. J. Fernández-García, J. C. Preciado, F. Melchor, R. Rodríguez-Echeverría, J. M. Conejero, and F. Sánchez-Figueroa, "A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data," *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3115851.

[62] F. Gafarov, Y. Rudneva, and U. Y. Sharifov, "Predictive Modeling in Higher Education: Determining Factors of Academic Performance," *Vysšee obrazovanie v Rossii*, 2023, doi: 10.31992/0869-3617-2023-32-1-51-70.

