FMDB *Transactions on Sustainable Techno Learning*

Predicting Student Academic Performance using Educational Data Mining and Feature Selection: Evidence from Higher Education

Roshan Renji[1,*], R. Mahalakshmi[2]

[1,2]Department of Advanced Computing and Analytics, School of Computing Science, Vels Institute of Science, Technology and Advanced Studies, Chennai, Tamil Nadu, India.
roshanrenji0@gmail.com[1], rmahalakshmi.scs@vistas.ac.in[2]

Abstract: Educational Data Mining (EDM) involves the extraction of valuable information and insights from educational data. EDM discerns patterns and trends in educational data, which may enhance academic curricula, pedagogical approaches, assessment techniques, and student performance. Recently, there has been renewed interest among EDM researchers in predicting students' academic performance based on factors such as entry-level attributes, family background, and educational accomplishments in the course. This study presents the findings of the EDM research conducted in the higher education institutions in Kerala, India. The data were collected from 1050 students studying their final year of a three-year course in various arts and science colleges in Kerala, India. The study focuses on developing data mining models to predict student achievement using their personal, pre-university, family, and other performance attributes. The comparison of performance accuracy when all the feature attributes are included and also with feature selection methods with reduced attributes revealed that the application of feature selection marginally enhances the performance of all the classifiers, Naive Bayes, SMO, IBK, JRip, and J48. The analysis demonstrated that the Relief ranking filter with the attribute ranking meta characterized the highest accuracy (79%) for the SMO classifier (25%). Correlation-based Feature Selection Greedy Stepwise search method produced the best prediction results for SMO and J48 classifiers (78.87%). Overall, results showed that the SMO classifier outperformed other standalone classifiers, such as Naïve Bayes, IBK, JRip, and J48, in terms of weighted average metrics such as Precision (0.760), Recall (0.789), Accuracy (0.788), F-measure (0.773), and Kappa statistic (0.604).

Keywords: Educational Data Mining; Prediction Techniques; Student Attributes; Feature Selection; Data Mining; Machine Learning; Artificial Neural Network; Generative Adversarial Network; Deep Learning.

Received on: 25/01/2025, **Revised on:** 26/03/2025, **Accepted on:** 03/06/2025, **Published on:** 05/06/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSTL

DOI: https://doi.org/10.69888/FTSTL.2026.000745

Cite as: R. Renji and R. Mahalakshmi, "Predicting Student Academic Performance using Educational Data Mining and Feature Selection: Evidence from Higher Education," *FMDB Transactions on Sustainable Techno Learning*, vol. 4, no. 2, pp. 57–67, 2026.

1. Introduction

The prediction of students' academic performance has long been considered an important indicator of an educational institution's quality [2]. Students' performance indicates the quality of education and its impact on an institution's reputation, its capacity to attract potential students, and its overall success. Higher education institutions aim to maintain high standards in

*Corresponding author.

teaching and learning so that graduates can meet industry requirements for skills and expertise while addressing societal needs [4]. There is increasing demand for data-driven methodologies in the education sector to predict student performance and improve educational outcomes [5]. Data mining (DM) and machine learning (ML) approaches have evolved as de facto standards for analyzing patterns [6]. EDM has evolved into an effective strategy that leverages data to uncover learning patterns, predict student performance, and enhance academic decision-making [7]. EDM is an expanding field that helps academic institutions enhance student performance. EDM facilitates individualized learning methodologies and enables educators to detect at-risk students and customize interventions productively [1]. The rapid development of modern educational technologies, greatly enhanced by data mining approaches, has revolutionized conventional education to address the changing needs of educators and students. EDM is pivotal in this change, enabling the derivation of important perceptions and insights from educational data to enhance educational performance [3]. Predicting student performance is a common use of EDM, helping educators create educational programs that are both successful and guided by data-driven insights. Recent advancements in data mining and machine learning (ML) have significantly enhanced student performance prediction systems, profoundly influencing subsequent elements such as personalized learning, targeted training and intervention programs, and improvements in an institution's educational outcomes [8]. Data mining is the process of extracting valuable patterns from existing datasets/extensive databases, often using advanced mining techniques.

2. Background and Related Work

Research on predicting student performance using ML is a prominent field that consistently attracts scholars. Arsad and Buniyamin [10] developed an Artificial Neural Network (ANN) model to predict students' academic performance in engineering courses. They used the Cumulative Grade Point Average (CGPA) system to assess academic performance in the eighth semester. The students' achievements in essential topics from the first semester served as the predictors, while the CGPA in the eighth semester functioned as the output. The models' performances were evaluated using the statistical parameters, namely "correlation coefficient" and "mean squared error". The results indicated that core subjects in the first and third semesters significantly impacted the ultimate CGPA upon graduation. Amrieh et al. [11] developed a novel model for predicting students' performance using data mining techniques and included additional data elements, such as behavioral aspects of students [3]. The suggested model utilizing behavioral features demonstrated an accuracy improvement of up to 22.1% compared to results obtained without these features, and an additional improvement of up to 25.8% with the application of ensemble methods [2]. The model had an accuracy exceeding 80% when evaluated with novice pupils. Zaffar et al. [12] have analyzed the efficacy of various feature selection and classification algorithms on two distinct student datasets for predicting academic performance [4]. The outcomes from various feature selection algorithms and classifiers applied to two student datasets with differing feature counts have helped the researchers identify optimal combinations of filter feature selection methods and classifiers [5]. The study emphasizes the importance of feature selection for predicting student performance, as effective instructional techniques can be developed from the relevant set of features. The investigation found a 10% difference in prediction accuracy across datasets with varying feature counts [1].

Hirokawa [13] assessed predictive performance using all conceivable combinations of four attribute types: "behavioral features", "demographic features", "academic background", and "parental involvement". The behavioral characteristics were presented as numerical data. However, it was presented as a pair consisting of an attribute name and its corresponding value [6]. This vectorization produced data with 417 dimensions, whereas the naively characterized data had 68 dimensions. By employing support vector machines with a feature selection algorithm, an optimal predictive performance was achieved, with an accuracy of 0.8096 and an F-measure of 0.7726 [20]. The study found that the behavioral feature is essential, achieving an accuracy of 0.7905 in isolation from other characteristics [26]. The integration of behavioral and demographic features achieved an F-measure of 0.7662. Gunawan et al. [14] have predicted students' on-time graduation using decision tree-based data mining with the C4.5 method and identified the characteristics that affect this prediction. The findings indicated that the GPA was the most significant attribute. The C4.5 algorithm achieved classification accuracy of 78.612%, precision of 0.762, and recall of 0.786. Li et al. [15] have analyzed student retention statistics using data science methodologies [8]. A comprehensive study of student characteristics reveals a significant correlation between student retention outcomes and academic success, specifically college GPA, and its implications for students' financial circumstances [21]. A predictive feature ranking and selection analysis has revealed a set of traits that serve as strong markers of student retention outcomes across the four student categories. The categorization analysis indicated that the models effectively identified the traits of "Stayed" students, namely those who persisted in completing their studies at MTSU [7].

The categorization accuracy for this student class exceeded 90%. However, the models were inadequate in accurately representing the attributes of the students who "Dropped" from their studies at MTSU [22]. The highest classification accuracy achieved for this category of students was approximately 66%, specifically for first-generation students [23]. This study concluded that students who are academically capable of remaining in school are withdrawing when they fall below the grade threshold required to maintain their HOPE grant [15]. Bujang et al. [16] highlighted the pressing need for predictive analytics in the higher education sector and the difficulties in applying ML to forecast student performance in terms of grades. The authors conducted a thorough investigation of six ML approaches, evaluating their effectiveness in predicting grades using actual data from various courses [24]. Furthermore, they suggested a multi-class predictive model to tackle unbalanced datasets. The model uses the Random Forest algorithm to achieve substantial improvements in f-measure. This approach demonstrated encouraging outcomes in improving predictive efficacy for unbalanced multi-class categorization in student grade forecasting. Gunawan et al. [17] proposed an innovative data-mining and feature-selection-based approach to analyze factors that may affect the time it takes students to graduate [27]. The use of Learner-Based Feature Selection, which reduces the number of features, enhanced accuracy with the Naïve Bayes algorithm, achieving 70.06% and 66.53%. Hemdanou et al. [18] have applied various feature selection and extraction methods, some of which are not typical for the education domain but are known to be effective in other domains. They predicted student success by integrating ML and DL with the identified criteria. They tailored Feature Selection using Mutual Information Maximization (FSMRMR), an Autoencoder, and a Generative Adversarial Network (GAN) to highlight the vital components that influence student performance [9].

They compared these models with commonly used ones. They examined the role of each component and validated its impact using various ML and DL methods. To determine the best model, a comparison analysis was conducted that accounted for the importance of multiple parameters. Timely graduation rates are crucial for universities, as they are directly related to institutional effectiveness and student success. In the field of educational data mining, Setiadi et al. [19] compared the performance of NB, KNN, and DT for classification. To improve feature selection while maintaining low complexity and high predictive power, two methods were used: Forward selection (FwS) and Backward elimination (BE). This research compared and contrasted the NB, KNN, and DT algorithms, using FwS and BE as the feature optimizers. The results show that model performance improved after feature selection across all algorithms. The KNN and DT algorithms yielded better results for FwS, while BE performed better for NB [1]. The KNN-based FwS algorithm achieved an accuracy of 83.8%, a significant improvement over the baseline of 76.64%. Students entering the degree course often have similar test scores when they apply. Nevertheless, there may be various reasons why differences in pupil attainment are increasing over time. Therefore, it is important to address these gaps to help all pupils learn and develop equally. The study's importance lies in the growing use of data-driven approaches in education to predict student performance [3]. This may allow the identification of pupils at risk and the making of recommendations for intervention at an early stage. Data mining and ML methods are deemed suitable for solving this problem by analyzing the complicated patterns in educational data. A review of past studies has clearly shown that student attributes, such as "academic marks," "socio-economic status," and "demographic factors," may significantly impact academic performance [2].

3. Materials and Methods

The conceptual framework of the proposed EDM model for predicting students' academic performance is shown in Figure 1.

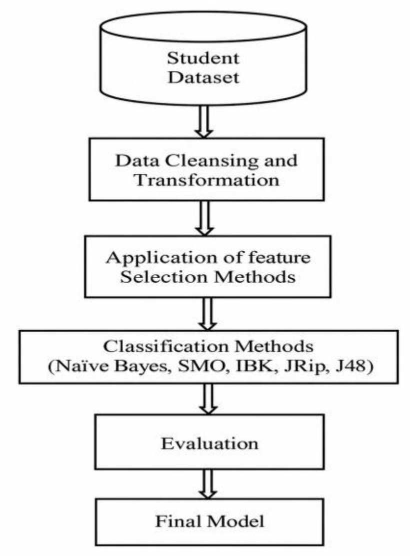

Figure 1: Conceptual framework of the study

3.1. Dataset

The dataset used in this study was collected from students enrolled in various higher educational institutions in the state of Kerala, India. 1050 student records were collected from various Arts and Science Colleges. Each record contains seventeen input attributes and one outcome attribute. The input attributes were based on students' demographic characteristics, pre-collegiate information, family-related information, and academic performance throughout the year. The output attribute measures the student's final academic performance at the end of the semester. All the variables/attributes, including the predicted class variable, were nominal with a finite number of possible values. Table 1 presents the attributes and the possible values.

Table 1: Student academic dataset

Variable Name	Description	Description of Categories
Gender	Gender of Students	Male, Female
Community	Community of Students	Other Classes (OC), Backward Classes (BC), Scheduled Castes (SC)
Higher Secondary Performance	Students grade in High School	First: Above 60% Second: 50% to 60% Third: Below 50%
Senior Secondary Performance	Students grade in Senior Secondary School	First: Above 60% Second: 50% to 60% Third: Below 50%
Residential Locality	Living Location of Student	Rural, Urban
Stay Type	Nature of Stay (hostel or Day scholar)	No, Yes
F_Size	Family size	1, 2, 3, >3
FType	Student’s family type	Joint, Individual
INC	Annual Income	Poor, Medium, High
F_Qual	Father’s qualification	HSE & Below, UG, PG & Above, Professional
M_Qual	Mother’s Qualification	HSE & Below, UG, PG & Above, Professional
PSM	Previous Semester Mark	Fail < 35%, Third >36 &<45%, Second >45 &<60%, First > 60%
CTG	Class Test Grade	Poor, Average, Good
SEM_P	Seminar Performance	Poor, Average, Good
ASS	Assignment	No (Not Submitted), Yes (Submitted)
GP	General Proficiency	Poor, Average, Good
ATT	Attendance	Poor, Average, Good
ESM	End-of-Semester Marks	Fail < 35%, Third >36 & <45%, Second >45 & <60%, First > 60%

3.2. Data Cleansing and Transformation

Data cleansing and data transformation are crucial phases of data mining. One key goal of data cleaning is to correct inaccuracies, remove redundancies, and fill in missing values to improve data quality. The data records were cleaned (inappropriate entries deleted), and missing data were filled in with standard procedures. This helped ensure the integrity of the data set, leading to more accurate analysis and results.

Thus, the results can be interpreted with higher certainty. Data transformation involves converting data into a suitable format for analysis, such as aggregating, normalizing, or discretizing it. Both techniques intend to prepare the data for efficient analysis and modeling. The categorical and nominal data entries were assigned numbers. This made it easier to analyze and compare results with data from various sources. The researchers were able to use statistical methods more effectively by converting qualitative data into numerical values.

3.3. Feature Selection Methods

Feature selection is an extensively used pre-processing method in data mining. The main task of feature selection is to reduce the number of features in a feature set without affecting performance. Feature selection enhances classification by discarding irrelevant and redundant data from the source data sets. It is commonly used to determine and select the most important attributes in a data set. The feature selection process is shown in Figure 2. The feature selection process comprises five essential components: the original dataset, the selection of a feature subset, evaluation of the selected subset, selection criteria, and validation techniques.

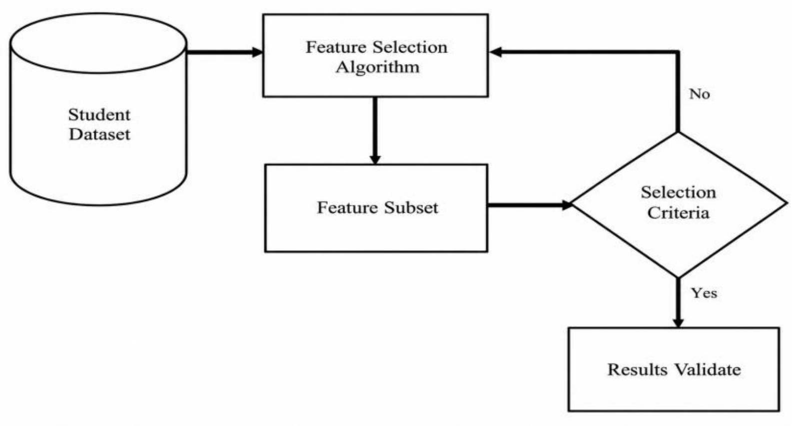

Figure 2: Feature selection process

A variety of feature selection approaches have been developed to generate the optimal subset of attributes. This study considered four feature selection algorithms based on historical performance, including Correlation Ranking Filter with Attribute Ranking (CRF_AR), CFS Subset Evaluator with Greedy Stepwise Method (CFSSE_GSM), Information Gain Ranking Filter with Attribute Ranking (IGRF_AR), and Relief Ranking Filter with Attribute Ranking (RRF_AR).

3.3.1. Correlation Ranking Filter with Attribute Ranking (CRF_AR)

The Correlation Ranking Filter is a statistical feature selection technique that identifies and ranks features based on their correlation with the target variable and with other features. It seeks to identify relevant elements related to the result while minimizing repetition. The procedure comprises many stages: Initially, it computes the correlation for each characteristic with the target variable, using coefficients such as Pearson's (for linear associations) and Spearman's rank (for ordinal data). Subsequently, characteristics are prioritized based on their correlation values, with higher correlations indicating greater significance. A filtering procedure is then implemented: characteristics that exceed a predetermined correlation threshold with the target are considered relevant, while redundancy is evaluated by analyzing feature correlations. Redundant characteristics, which exhibit significant correlation with each other, result in the elimination of the feature that has a lesser association with the target variable. This yields a final feature set that is both predictive and only weakly intercorrelated, thereby successfully mitigating multicollinearity. Attribute ranking is a comprehensive category that includes several filter-based feature selection techniques, which evaluate the significance of each feature individually using statistical metrics such as Information Gain or Chi-Squared. Characteristics are prioritized by rating, facilitating the selection of the top k characteristics or those above a specified threshold. The Correlation Ranking Filter employs this methodology, using correlation as its principal indicator of efficiency and efficacy in locating non-redundant features, often functioning as a first step in hybrid feature selection approaches that combine filtering and wrapper techniques.

3.3.2. CFS Subset Evaluator with Greedy Stepwise Method (CFSSE_ GSM)

The CFS Subset Evaluator using the Greedy Stepwise Method (CFSSE_GSM) is a well-recognized filter-based feature selection approach in machine learning. It integrates the Correlation-based Feature Selection (CFS) heuristic for assessing the value of a feature subset with a Greedy Stepwise search method to explore the feature space. The fundamental tenet of CFSSE_GSM is that "an effective feature subset comprises features that are strongly correlated with the class but uncorrelated with one another". The Greedy Stepwise technique is a heuristic search algorithm that investigates the domain of potential feature subsets. It works by making locally optimum decisions at each stage, aiming to get a globally optimal or near-optimal solution.

3.3.3. Information Gain Ranking Filter with Attribute Ranking (IGRF_AR)

The Information Gain Ranking Filter with Attribute Ranking (IGRF_AR) is a filter-based feature selection technique used in ML and data mining. It uses the Information Gain (IG) principle to assess the significance of each feature (attribute) individually, then ranks them to identify the most relevant subset for model training. The core premise of IGRF_AR is based on information theory, particularly entropy minimization:

- **Entropy:** A quantification of the disorder or unpredictability within a dataset's class distribution. A dataset containing only examples of a single class exhibits zero entropy, while a fully balanced dataset demonstrates maximum entropy.

- **Information Gain (IG):** Quantifies the anticipated decrease in entropy resulting from the segmentation of data according to a certain property. A higher information gain value indicates a greater ability to distinguish among classes.

3.3.4. Relief-Based Feature Selection Methods

Relief-based feature selection techniques constitute a robust category of filter algorithms that evaluate feature quality by assessing their ability to distinguish between proximate examples within the instance space.

Table 2: Attribute selection using different feature selection methods

Feature Selection Method	Selected Features	Number of Features (N=17)
Correlation Ranking Filter with Attribute Ranking (CRF_AR)	FINC, PSM, ATT, Sex, LOC, GP, Cat, CTG, FQUAL, HSG_Grade, FSIZE, SEM_P	12
CFS Subset Evaluator with Greedy Stepwise Method (CFSSE_ GSM)	Cat, SSG_Grade, FINC, PSM, GP, ATT	6
Information Gain Ranking Filter with Attribute Ranking (IGRF_AR)	PSM, FINC, ATT, SSG_Grade, CTG, Sex, Cat, SEM_P, FQUAL, HSG_Grade	10
Relief Ranking Filter with Attribute Ranking (RRF_AR)	FINC, PSM, ATT, SSG_Grade, CTG, Sex, LOC, SG_Grade, SEM_P, Cat	10

They are distinguished by their ability to identify feature interactions and non-linear connections without explicitly evaluating feature combinations. A summary of the features selected by each feature selection method is presented in Table 2.

3.4. Classification Model for Student Academic Performance Prediction

The predictive accuracy of the different features selected through different feature selection algorithms can be assessed using classification methods. Presently, more than 15 classification algorithms are available in ML tools; however, owing to spatial constraints, only 5 classification techniques were focused on in this work. To find the best classifier that effectively generalizes the data with greater accuracy, this study analyzed five classifiers from each of four broad classification families. Thus, the Naive Bayes classifier from the Bayes family, the SMO classifier from the Functions family, the IBK classifier from the Lazy Classifiers family, the JRip classifier from the Rules family, and the J48 classifier from the Trees family of classifiers were selected and implemented in Python. The reason for choosing these four classifiers (Naive Bayes, SMO, IBK, JRip, and J48) is that they achieved the best performance compared with other classifiers in the same family.

4. Results and Discussion

The train-test split method was used to assess the efficacy of different ML algorithms in predicting student outcomes using data withheld from model training. The k-fold cross-validation approach is one of the popular methods for splitting the training and testing datasets for model prediction. In this method, the dataset is typically split into k mutually exclusive, equal-sized subsets (folds), and the classifier is trained on each subset. Each time, one of the k subsets is utilized for testing, while the remaining k-1 subsets are employed for training. Accuracy is calculated k times across k tests, and the average of the k resulting accuracies provides a forecast of the classification task's accuracy. Cross-validation uses all observations in the given dataset for testing, ensuring that test sets are independent and enhancing the reliability of the findings. This research used $k = 10$, randomly partitioning the data into 10 equal segments, with one segment designated as the test dataset and the other nine segments utilized as training sets.

4.1. Performance Accuracy of the Classifiers without Feature Selection

The performance accuracy of the classifiers with all seventeen input attributes is shown in Table 3.

Table 3: Results of the classification methods with all attributes classifier performance accuracy

Classifier	Performance Accuracy
Naïve Bayes	75.85%
SMO	78.49%
IBK	72.08%
JRip	75.47%

J48	76.60%
Average Accuracy	75.70%

SMO classifier recorded the best accuracy of 78.49%, followed by J48 (76.60%) and JRip (75.47%) when all the features were included. The IBK classifier achieved the lowest accuracy of 72.08% without feature selection.

4.2. Results of the Classification Methods with Feature Selection

The average performance accuracy of the classifiers with feature selection is shown in Table 4 and Figure 3.

Table 4: Results of the classification methods with feature selection

Classifier	Accuracy with Different Feature Selection Methods				Average Accuracy of Classifiers
	Correlation Ranking Filter With Attribute Ranking (12)	CFS Subset Evaluator with Greedy Stepwise Method (6)	Information Gain Ranking Filter with Attribute Ranking (10)	Relief Ranking Filter with Attribute Ranking (10)	
NaiveBayes	76.23%	78.49%	77.74%	77.74%	77.55%
SMO	78.49%	78.87%	78.87%	79.25%	78.87%
IBK	70.19%	76.98%	75.09%	75.47%	74.43%
JRip	76.23%	76.98%	76.98%	78.49%	77.17%
J48	77.74%	78.87%	78.49%	78.11%	78.30%
Average Accuracy	75.78%	78.04%	77.43%	77.81%	

Figure 3 shows the “classification accuracy” of five machine learning algorithms (Naïve Bayes, SMO, IBK, JRip and J48) using four different feature selection methods: “CRF_AR, CFSSE_GSM, IGRF_AR and RRF_AR”. Generally, in most feature selection methods, “SMO and J48” are more accurate, but “IBK” is comparatively less accurate.

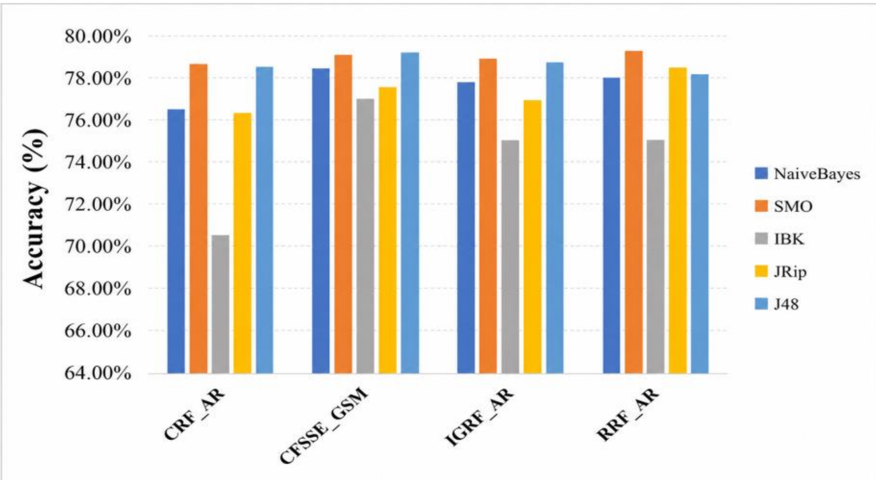

Figure 3: Results of the classification methods with feature selection

The results show that the selection of the feature selection technique is critically influential on classification accuracy, with the highest overall performance achieved by “RRF_AR” and “CFSSE_GSM”.

4.3. Average Performance Accuracy of Different Feature Selection Methods

The classification results show that SMO achieved the highest accuracy (79.25%) with the Relief Ranking Filter and attribute ranking methods. Likewise, the highest accuracy was obtained with the CFS Subset Evaluator and the Greedy Stepwise Search Method for the SMO classifier (78.87%) and JRip (78.87%). The Information Gain Ranking Filter with Attribute Ranking Search Method achieved the highest accuracy (78.87%) with the SMO classifier. The results indicate that the SMO algorithm has the highest accuracy (79.25%), followed by the J48 algorithm (78.87%), and the Naïve Bayes and JRip algorithms (78.49% each). The results demonstrate the effectiveness of the SMO classifier as compared to the other classifiers tested. Additional studies could further investigate the feasibility of combining these techniques to improve classification accuracy on other datasets (Figure 4).

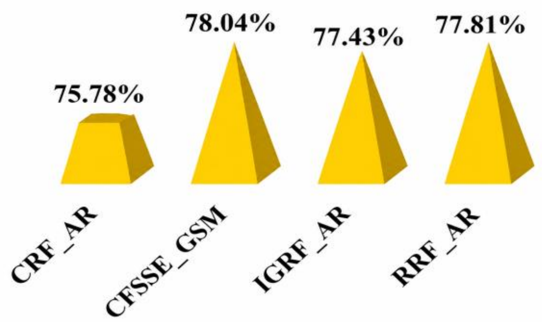

Figure 4: Average performance accuracy of different feature selection methods

The average performance accuracy of different Feature selection techniques showed that the CFSSE_GSM technique achieved the best accuracy of 78.04%. The technique was then followed by Relief Ranking Filter with Attribute Ranking (RRF_AR) and Information Gain Ranking Filter with Attribute Ranking (IGRF_AR), which yielded 77.81% and 77.43%, respectively. The average performance without a feature selection method is 75.70%, while the feature selection method, namely Correlation Ranking Filter with Attribute Ranking (CRF_AR), achieves a sub-optimal accuracy of 75.78%. The findings indicate the importance of feature selection in improving model predictive performance. The use of advanced techniques such as CFSSE_GSM can further enhance the accuracy, unlocking the potential for more accurate academic performance prediction analyses in student applications.

4.4. Average Performance Accuracy of Different Classifiers

The average performance of different classifiers showed that SMO achieved the best accuracy of 78.87%. J48 and Naïve Bayes were then run on the technique, achieving accuracies of 78.30% and 77.55%, respectively. JRip achieved 77.17% accuracy, while IBK's average accuracy was 74.43% (Figure 5).

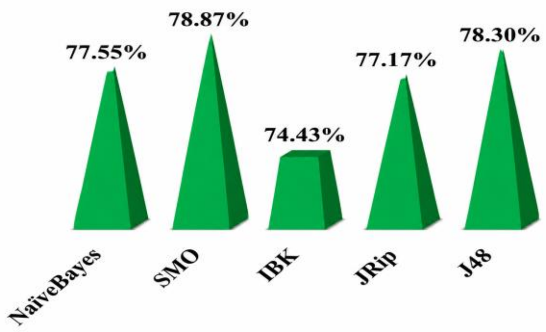

Figure 5: Average performance accuracy of different classifiers

These results show that there was a significant difference in the performance of the classifiers used. Additional investigation may be needed to determine what may be causing this difference in their performance.

4.5. Comparison of Average Performance Accuracy between Different Classifiers with and Without Feature Selection

Significant insights emerged from comparing the average performance accuracy of different classifiers with and without feature selection methods. Across classifiers, the average performance with feature selection methods has outperformed that without feature selection. The results revealed that SMO achieved the highest accuracy without feature selection (78.49%) and with feature selection (78.87%). The IBK classifier achieved the lowest average accuracy of 72.08% without feature selection; with feature selection, the accuracy was 74.43%. IBK, a distance-based method, could be adversely affected by noise and irrelevant features, as they might distort distance metrics, such as the Euclidean distance, used to identify the closest neighbors, leading to misclassifications. The results suggest that feature selection can enhance model performance, especially in SMO, where selected features are effectively used (Figure 6).

Figure 6: Comparison of average performance accuracy among different classifiers with and without feature selection

A more detailed examination of what might be causing these accuracy improvements and whether they are relevant across different data sets could also be conducted.

Table 5: Prediction results of different classifiers with CFS subset evaluator (Greedy Stepwise Method) feature selection

Prediction Method	Precision	Recall	Accuracy	F-measure	Kappa Statistic
NaiveBayes	0.766	0.785	0.784	0.772	0.603
SMO	0.760	0.789	0.788	0.773	0.604
IBK	0.739	0.770	0.769	0.747	0.555
JRip	0.726	0.770	0.769	0.746	0.569
J48	0.781	0.789	0.788	0.764	0.585

The overall prediction results of different classifiers, considering the CFS Subset Evaluator (Greedy Stepwise Method) Feature Selection, are presented in Table 5.

4.6. Findings

The empirical analysis revealed that the overall classification accuracy of the classifiers ranged from 72.08% to 78.49% when all attributes were included. Overall, all the classifiers reported optimum classification accuracies above 70%. The average classifier accuracy ranged from 74.43% to 78.87% when feature selection methods were applied. The feature selection method, namely the CFS Subset Evaluator with Greedy Stepwise Method, with six attributes, exhibited the best accuracy of 78.04%. The SMO classifier achieved the best accuracy of 78.87% with feature selection. The findings are supported by several studies, including Ahmed [25], which reported the highest accuracy for SMO classifiers compared with other classifiers such as Naïve Bayes, IBK, JRip, and J48. The improved performance of SVM arises from its margin maximization, its capacity to handle non-linear relationships among attribute categories, and its resilience to overfitting. Consistent with the present study, Kabakchieva [26] also found that Bayes classifiers have lower accuracy than alternatives. The results of this study differ from those of Menaka and Kesavaraj [27], who reported that IBK outperformed other classifiers by a large margin.

5. Conclusion

Determining the factors that impact students' academic performance can enhance intervention strategies and support services for those struggling in their studies, enabling timely assistance. Given that educational data is often skewed and sparse, significant effort is required in the pre-processing stages to ensure that the filtering process yields an effective model. The extra effort is required to model the prediction outputs as valuable information for identifying underperforming students. The results clearly showed that classifier algorithms are effective in predicting students' academic performance in the final exam using different input attributes identified in the study. In addition, incorporating feature selection methods has marginally improved the performance of the prediction methods. Even though feature selection methods improve classifier performance, the overall performance remains below 80%. Therefore, there is significant potential to enhance the existing feature selection and classification methods through alternative approaches such as parameter tuning and hybridization. Exploring these alternatives could lead to significant advancements in predictive accuracy. Additionally, incorporating ML techniques such as ensemble methods might yield promising results in addressing the limitations currently observed. Furthermore, the performance of the classifiers could be greatly enhanced by incorporating hybrid feature selection methods that incorporate meta-heuristic approaches. Thus, future work will focus on developing advanced feature selection methods that extract the best features that significantly contribute to the outcome (students' end-of-semester performance).

Acknowledgment: The authors sincerely thank Vels Institute of Science, Technology, and Advanced Studies for its encouragement, support, and academic assistance throughout this study.

Data Availability Statement: No new data were generated or analyzed during this study; therefore, data availability does not apply to this paper.

Funding Statement: No external funding or financial assistance was received for conducting this research or preparing this manuscript.

Conflicts of Interest Statement: The authors declare that they have no conflicts of interest related to this research.

Ethics and Consent Statement: This study was conducted in accordance with established ethical standards, and informed consent was obtained from all participants.

References

1. A. Angeioplastis, J. Aliprantis, M. Konstantakis, and A. Tsimpiris, "Predicting student performance and enhancing learning outcomes: A data-driven approach using educational data mining techniques," *Computers*, vol. 14, no. 3, p. 83, 2025.
2. G. Deeva, J. D. Smedt, and J. D. Weerdt, "Educational sequence mining for dropout prediction in MOOCs: Model building, evaluation, and benchmarking," *IEEE Trans. Learn. Technol.*, vol. 15, no. 6, pp. 720–735, 2022.
3. A. Khan and S. K. Ghosh, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," *Educ. Inf. Technol.*, vol. 26, no. 1, pp. 205–240, 2021.
4. S. Batool, J. Rashid, M. W. Nisar, J. Kim, H. Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," *Educ. Inf. Technol*., vol. 28, no. 1, pp. 905–971, 2023.
5. Z. Xu, H. Yuan, and Q. Liu, "Student performance prediction based on blended learning," *IEEE Trans. Educ.*, vol. 64, no. 1, pp. 66–73, 2021.
6. F. Wang, Z. Huang, Q. Liu, E. Chen, Y. Yin, and J. Ma, "Dynamic cognitive diagnosis: An educational priors-enhanced deep knowledge tracing perspective," *IEEE Trans. Learn. Technol*., vol. 16, no. 3, pp. 306–323, 2023.
7. Y. Chen and L. Zhai, "A comparative study on student performance prediction using machine learning," *Educ. Inf. Technol.*, vol. 28, no. 9, pp. 12039–12057, 2023.
8. M. E. Webb, A. Fluck, J. Magenheim, J. Malyn-Smith, J. Waters, M. Deschênes, and J. Zagami, "Machine Learning for human learners: Opportunities, issues, tensions and threats," *Educ. Technol. Res. Dev*., vol. 69, no. 4, pp. 2109–2130, 2021.
9. J. H. Yu, D. Chauhan, R. A. Iqbal, and E. Yeoh, "Mapping academic perspectives on AI in education: Trends, challenges, and sentiments in educational research (2018–2024)," *Educ. Technol. Res. Dev*., vol. 73, no. 1, pp. 199–227, 2025.
10. P. M. Arsad, N. Buniyamin, and J. L. A. Manan, "A neural network students' performance prediction model (NNSPPM)," in *Proc. IEEE Int. Conf. Smart Instrum. Meas. Appl. (ICSIMA)*, Kuala Lumpur, Malaysia, 2013.
11. E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining educational data to predict students' academic performance using ensemble methods," *Int. J. Database Theory Appl.*, vol. 9, no. 8, pp. 119–136, 2016.
12. M. Zaffar, M. A. Hashmani, K. S. Savita, and S. S. H. Rizvi, "A study of feature selection algorithms for predicting students' academic performance," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 5, pp. 541–549, 2018.
13. S. Hirokawa, "Key attribute for predicting student academic performance," in *Proc. 10th Int. Conf. Educ. Technol. Comput*., Tokyo, Japan, 2018.
14. G. Gunawan, H. Hanes, and C. Catherine, "Information systems students' study performance prediction using data mining approach," in *Proc. 4th Int. Conf. Informat. Comput. (ICIC)*, Semarang, Indonesia, 2019.
15. C. Li, M. Hains, J. Wallin, and Q. Wu, "Applying data science methods for early prediction of undergraduate student retention," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Las Vegas, Nevada, United States of America, 2019.
16. S. D. A. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, and H. Fujita, "Multiclass prediction model for student grade prediction using machine learning," *IEEE Access*, vol. 9, no. 6, pp. 95608–95621, 2021.

17. G. Gunawan, F. Halim, and D. Djoni, "Students' timely graduation attributes prediction using feature selection techniques, case study: Informatics Engineering Bachelor study program," in *Proc. IEEE Int. Conf. Comput. Sci. Inf. Technol. (ICOSNIKOM)*, Laguboti, North Sumatra, Indonesia, 2022.
18. A. L. Hemdanou, M. L. Sefian, Y. Achtoun, and I. Tahiri, "Comparative analysis of feature selection and extraction methods for student performance prediction across different machine learning models," *Comput. Educ. Artif. Intell.*, vol. 7, no. 12, p. 100301, 2024.
19. H. Setiadi, K. Sanjaya, A. Wijayanto, D. W. Wardhani, and H. D. Cahyono, "Comparative analysis of classification algorithms using feature selection techniques to predict on-time student graduation," *Information Systems Engineering*, vol. 29, no. 4, pp. 1365–1379, 2024.
20. S. K. Singh and R. K. Dwivedi, "Data mining: Dirty data and data cleaning," *SSRN*, 2020. [Accessed by 17/11/2024].
21. J. García, J. Entrena, and Á. Alesanco, "Empirical evaluation of feature selection methods for machine learning based intrusion detection in IoT scenarios," *Internet Things*, vol. 28, no. 12, p. 101367, 2024.
22. B. Kathirgamanathan and P. Cunningham, "Correlation based feature subset selection for multivariate time-series data," *arXiv preprint arXiv:2112.03705*, 2021. [Accessed by 22/11/2024].
23. S. Jadhav, H. He, and K. Jenkins, "Information gain directed genetic algorithm wrapper feature selection for credit rating," *Appl. Soft Comput.*, vol. 69, no. 8, pp. 541–553, 2018.
24. R. J. Urbanowicz, M. Meeker, W. L. Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *J. Biomed. Inform.*, vol. 85, no. 9, pp. 189–203, 2018.
25. E. Ahmed, "Student performance prediction using machine learning algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2024, no. 1, p. 4067721, 2024.
26. D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybern. Inf. Technol.*, vol. 13, no. 1, pp. 61–72, 2013.
27. S. Menaka and G. Kesavaraj, "Predicting student performance using data mining techniques: A survey of the last 5 years," *Int. J. Adv. Sci. Res. Manage.*, vol. 4, no. 1, pp. 98–102, 2019.