EFFICIENT FEATURE SELECTION FOR PREDICTING ACADEMIC PERFORMANCE

#1SHAIK NAGUR BABU,
MCA Student, Dept of MCA,
#2Dr.MOHAMMAD SIRAJUDDIN,
Associate Professor, Department of MCA,
VAAGESWARI COLLEGE OF ENGINEERING (AUTONOMOUS), KARIMNAGAR, TG.

ABSTRACT: The purpose of this research is to develop a feature selection system that is capable of accurately predicting the academic achievement of an individual. Academic success prediction is crucial in education because it provides researchers with valuable data about student performance and allows teachers to execute interventions that are more focused. It is difficult to find the most important facts from the mountain of data that could be used for success forecasting. This data includes demographic information, trends of behavior, and academic records from the past. Traditional feature selection methods might fail to account for intricate component interactions and have trouble processing massive datasets. This research proposes a new way for selecting qualities that combines statistics and machine learning techniques to improve the process. Simplifying calculations and increasing forecast accuracy are the goals of the technique. In order to make the predictive models more clear and relevant for teachers and administrators, the research used advanced feature selection methods. The research uses real academic data to assess the efficacy of the proposed method, and the results demonstrate that it is superior at predicting students' academic achievement. The findings pave the way for additional research on educational data mining by demonstrating the critical importance of selecting appropriate criteria for academic success prediction.

Index Terms: *Feature Selection, Academic Performance Prediction, Machine Learning, Data Mining, Predictive Modeling, Student Outcomes, Statistical Methods, Educational Analytics, Dimensionality Reduction, Feature Importance, Model Optimization, Educational Data, Academic Success.*

1. INTRODUCTION

Leaders, teachers, and students would all greatly benefit from an accurate way to forecast a student's future IQ. By using information about students' demographics, research habits, and past academic performance, schools may quickly identify at-risk pupils and offer support. Nonetheless, the model's features have a big impact on how accurate it is. Because redundant or superfluous features can introduce noise, raise computing costs, and produce erroneous models, feature selection is an essential part of the data preprocessing pipeline.

The goal of feature selection is to find and remove the least important elements from a potentially large collection. By lowering the number of dataset variables, this strategy aims to improve the model's effectiveness and usefulness. It can be difficult to identify the critical factors affecting a student's academic performance. Academic achievement, research habits, mental health, extracurricular activities, and family financial situation are just a few of the many variables taken into account. Not all variables, though, have a favorable impact on predictions; some could make the model biased or overly effective. This means that by using qualities appropriately, useful predictions are kept and irrelevant ones are rejected.

Forecasting models can be improved by using a variety of feature selection techniques. These methods are mostly used for anchoring, encapsulating, and screening. To determine a feature's relevance, filter techniques use statistical measures like correlation or mutual information. They have quick and accurate computational abilities. Machine learning is used in wrapper techniques to evaluate trait subsets' prospective performance. Although wrapper methods are more accurate, they need more resources, especially when handling many aspects. Feature selection and model training are done simultaneously via

embedded methods, which lowers processing expenses.

The determination of the best predictors for academic achievement is made more difficult by the vast amount and variety of data. Since student data usually includes both numerical and category components, such as demographics, test results, and grade point averages, it is essential to comprehend how different forms of student data are managed. Nonlinear correlations between features might make it more difficult to use traditional statistical methods to evaluate interactions. When used in conjunction with feature selection techniques, advanced machine learning algorithms such as support vector machines, decision trees, and random forests can greatly improve the accuracy of prediction models. Because they offer significant insights into the significance of numerous components, these algorithms are excellent for analyzing complicated, high-dimensional data.

Model performance is improved by careful feature selection in ways that go beyond simple prediction accuracy. It is also advantageous to comprehend the several elements that affect academic achievement. Interventions can be tailored to meet the needs of different student groups once the main factors determining academic success or failure have been identified. Research groups, tutoring programs, and mental health support are some of these choices. In the age of big data and machine learning, this approach is becoming more and more important for supporting each student and promoting data-driven, individualized education.

2. LITERATURE REVIEW

Mishra & Patel (2020): This research provides new evidence for a method for pinpointing the most important factors influencing pupils' academic achievement. Greater clarity and accuracy are achieved by the combination of filter- and wrapper-based approaches. Companies that use machine learning approaches, such as RFE and Genetic Algorithms (GA), provided the empirical data used in the review. The reliability of student accomplishment assessments is enhanced by this hybrid system, which surpasses standard methodologies. Pedagogical data mining can be advanced with the help of the findings.

Farooq & Zainab (2020): This research uses a wide range of datasets to compare and contrast several multiple feature selection methods for academic achievement prediction. Mutual Information and Correlation-based Feature Selection (CFS) are two of the strategies that are evaluated for their effectiveness and precision. According to the research's findings, feature selection methods' efficacy depends on the kind of data and learning model used. Academics and teachers can benefit greatly from the writers' suggestions. These best practices highlight the use of contextual approaches. In order to refine academic performance prediction algorithms, real student performance outcomes are utilized.

Sharma & Singh (2020): This research seeks to identify the best practices for identifying academic analytics-relevant attributes. The research looks at how decision trees and support vector machines (SVMs) might be enhanced using techniques like chi-square and information gain. According to the findings, the most precise systems are hybrids that use both filter and wrapper methods. Preprocessing is crucial when developing algorithms to forecast pupils' academic achievement, according to the research. A significant contribution to educational statistics is made by the findings.

Nguyen & Lim (2021): This research examines the impact of feature selection on prediction accuracy by examining student data from Asian colleges. Machine learning models can be made more effective by utilizing heuristic-based selection and similar approaches. The models consist of neural networks and logistic regression. Based on the findings, it appears that using effective feature selection helps with both predicting a student's academic success and identifying important educational features. Attendance records and past grades are important indications that show important information. According to the findings, data-driven learning approaches that are individualized for each student should be put into practice.

Patel & Mehta (2021): The results of this research show how to reliably find important markers of

classroom effectiveness. Machine learning models are shown to be more accurate when methods like mRMR and Boruta are used in this research. According to the findings, picking precise student-related criteria greatly increases the model's usefulness. The authors state that institutions should incorporate predictive analytics into their decision-making procedures. Their findings provide concrete recommendations for enhancing academic support and retention rates.

Lee & Park (2021): The research intends to ascertain the extent to which students' academic success can be predicted by online learning settings. Looking at how often people log in and how active they are on forums can help shed light on academic achievement. Through the use of methods like Principal Component Analysis (PCA), the investigation's most important predictions are uncovered. The results show how crucial behavioral data is for planning e-learning tasks. The inquiry may have uncovered ways to enhance online learning.

Wang & Yu (2022): Different academic prediction machine learning models are compared and contrasted in this research based on the components they use. To test how ANOVA and Gain Ratio affect performance improvement, experiments use regression and classification models. The writers propose a hybrid ranking system that takes both consistency and relevance into account when evaluating attributes. Sensitivity studies have shown that there are some traits that hold true across different educational backgrounds. Their research lays the groundwork for evidence-based teaching practices to be used in a systematic manner.

Tiwari & Bhatia (2022): Feature selection and deep learning are just two of the many methods for academic achievement prediction that are explored in this research. To keep important ideas intact when handling large student datasets, autoencoders and deep belief networks are crucial tools. To enhance the functionality of models trained on ideal feature sets, the research makes use of LASSO and other approaches. According to the findings, deep learning models outperform more conventional methods when it comes to forecasting academic achievement. A comprehensive feature selection, say the experts, can enhance predictive analytics.

Bansal & Kumar (2022): The accuracy of the assessment's forecasts of students' academic success has been enhanced through the use of multiple feature selection approaches in this evaluation. Based on their pros and cons, the methods are categorized into four groups: filters, wrappers, embedded strategies, and hybrid models. The authors mention that one of the most recent advancements in feature selection is the use of deep learning techniques and metaheuristic algorithms. From their review of over 50 studies, they draw conclusions on which approach is best suited to the data and provide recommendations for future research. Academics working on educational data extraction stand to gain substantially from this result.

Zhang, Chen & Wang (2022): An approach to feature selection that uses evolutionary algorithms and neural networks is demonstrated in this article. To enhance feature sets, the authors employ multilayer perceptrons and particle swarm optimization. When tested with up-to-the-minute academic data from Chinese universities, the model showed remarkable accuracy and precision. Academic forecasts benefit from the intricate web of relationships seen in student data. This method works especially well with dynamic classes because of its automated characteristics.

Ali & Khan (2023): Ensemble machine learning models are the focus of this research's examination into student performance prediction. The authors proved that the effect model works for feature tagging and the Random Subspace Method, two feature selection approaches. Even with datasets that are in a state of disarray, they can use algorithms like Gradient Boosting and Random Forest to get accurate predictions. The findings point to the potential use of ensemble models in educational databases to aid at-risk students in their pursuit of academic success. The research's overarching goal is to encourage school leaders to make greater use of predictive analytics.

Zhou & Li (2023): This article delves into enhanced feature selection approaches developed for academic forecasting. The authors use a combination of association filtering and feature

interaction scores to eliminate unnecessary data. After comparing the k-NN and SVM models, they find that the former significantly improves prediction accuracy. Striking a balance between feature correctness and comprehensibility is their key objective in order to maximize data-driven decision-making. Efforts like these enhance academic analytics software.

Chen & Wu (2023): An evaluation of the efficacy of academic performance prediction using feature selection based on machine learning is the aim of this work. The authors use filtering and embedding approaches, such as Elastic Net and mutual information, to enhance student datasets. Their findings imply that administrators and teachers both benefit from models that have unnecessary details removed. Research authors discuss ethical concerns with predictive AI and stress the need for open and honest data management practices pertaining to student information. Their work supports the responsible use of AI in classrooms.

Srinivasan & Thomas (2024): Academic performance estimates for pupils are made more accurate with the introduction of an enhanced feature selection process in this experiment. If the authors want to fix the problem of pupil data pattern variability, they propose a new Ant Colony Optimization (ACO) method. Random Forests and Support Vector Machines, among other machine learning models, show enhanced accuracy and performance when this method is applied. Extensive testing has shown that ACO outperforms conventional techniques of personnel selection. The findings highlight how crucial it is to optimize educational data mining even more.

Rahman & Hasan (2024): Students' success in MOOCs may be predicted with the use of multidimensional feature selection, which was used in this research. The writers present a fresh approach that considers demographic, behavioral, and cognitive factors. Combining Multi-Objective Optimization (MOO) with ensemble learning models, particularly XGBoost and LightGBM, improves the forecast accuracy. According to the findings, using a broad array of variables makes the data more comprehensible and applicable to many situations. This research shows how critical it is to gather accurate information about MOOC (massive open online course) participants.

3. RELATED WORK

A lot of work in educational data mining and ML has gone into trying to predict how well students will do in school. Much research has gone into determining the most important factors that affect students' performance so that prediction models can be made that are both accurate and easy to understand.

Feature Selection Techniques: A variety of feature selection methods were implemented to enhance the model's functionality while avoiding overfitting. When trying to figure out how important a feature is, a few common approaches include Information Gain, Chi-Square, and Correlation-based Feature Selection (CFS). Use of more complex methods like Recursive Feature Elimination (RFE) and model-based embedding techniques like Random Forest and Lasso is commonplace when trying to discover characteristics that are important and significantly affect results.

Machine Learning Models for Academic Prediction: Academic performance prediction has made use of a number of machine learning approaches. Decision Trees, Random Forest, Gradient Boosting, Support Vector Machines (SVM), and k-Nearest Neighbors (KNN) are commonly utilized due to their ability to manage data with non-linear or complex connections.

Despite the fact that neural networks are becoming more popular, their capacity to detect deeper patterns frequently need bigger datasets and more processing power.

Key Predictive Features: A student's success in the classroom is strongly correlated with a number of personality qualities. Parental income, level of schooling, attendance, past academic performance, and scores on internal tests are some of the traits. In addition, research habits and time spent on learning platforms are examples of behavioral factors that have been properly predicted. It is critical to pick these features with care to improve the model's performance and make sure the predictions are useful and correct.

Hybrid Approaches: Operations have been enhanced by the utilization of hybrid models, which integrate multiple feature selection strategies with machine learning. Combining ensemble learning models with feature selection techniques like filters and wrappers, these solutions often beat single-method approaches. This combination makes the intricate connections between traits and performance in the classroom much clearer.

Challenges Identified: Problems persist despite progress in this area, especially when it comes to using prediction models on different datasets and in different types of educational institutions. Predictive algorithms can help students thrive in school, but they also raise important social challenges. If we want our predictions to be fair and just, we must make sure that feature selection doesn't introduce bias.

EXISTING SYSTEM

Machine learning algorithms are currently the backbone of academic success prediction systems. The primary goal of these methods is to extract the most relevant characteristics from student records. Some of the most common qualities are demographic information, level of education, behavior patterns, and how actively students engage with course materials. Identifying the most significant components influencing academic accomplishment was done using traditional feature selection techniques as Information Gain, Chi-Square, and Correlation-based Feature Selection (CFS).

Advanced feature selection methods, such RFE and integrated techniques, are used by modern systems. These algorithms incorporate feature selection directly into the learning process. In conjunction with machine learning models such as Decision Trees, Random Forests, and Support Vector Machines (SVM), these methods have shown promising results. Nevertheless, it is still difficult to guarantee that these models work effectively in many types of classroom settings. When models become highly optimized for specific datasets, overfitting happens. This is a major problem with many contemporary systems. Consequently, they work best with classes that have a lot more students. Also, prediction algorithms won't be able to support socioeconomic or demographic biases as long as ethics and justice are prioritized. In spite of these problems, accurate and clear academic achievement predictions can only be achieved by employing suitable feature selection approaches.

Drawbacks of Existing System

- Improper feature selection might lead to overfitting. The model may grow overly dependent on the training data and underperform when given new data, even when it is highly accurate when training.
- Success in school can be hard to predict without more domain-specific data, such as how people's socioeconomic status or mental health affects them. Current methods likely overlook this information.
- Subtle details could be lost in the process.
- Data duplication happens because some feature selection methods can't handle features that are very similar to each other. Because of this, forecasting models' precision and speed might change.
- The current feature selection approaches may find it increasingly challenging to handle ever-increasing data sets. A longer and less efficient selection process is the result of an increase in the number of qualities. This is because it becomes more costly to process datasets that are higher in size.
- Feature selection methods that aren't robust can lead to bias if they give greater weight to some qualities than others. The complexity of academic achievement may be lost in the end as a result of such an approach.

PROPOSED SYSTEM

Using advanced feature selection methods, the suggested approach for Efficient Feature Selection in Predicting Academic Performance can increase the accuracy and efficiency of predictive model computations. To determine what factors most significantly affect students' performance in the classroom, this method employs statistical analysis and machine learning algorithms. Using domain knowledge, the suggested solution takes into account factors like learning style, emotional health, and socioeconomic status. Contrarily,

conventional methods may overlook feature correlations and domain-specific characteristics.

Using Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA), the technique employs a minimum of dimensions while maximizing the utilization of valuable, non-redundant features. This improves the model's generalizability and decreases the likelihood of overfitting. Using methods that make managing large datasets with little computational resources easy, the system is built to operate efficiently on a large scale. It uses bias-reduction strategies to keep feature selection fair while enhancing the model's accuracy and reliability. This all-encompassing method ought to improve prediction accuracy and shed light on the critical variables impacting academic performance.

Advantages of Proposed System:

- Academic performance models can be predicted to be more accurate using this method because it focuses on the most important aspects. This leads to more precise outcomes and improved decision-making.
- The method effectively lessens overfitting by removing unnecessary or extra components. The model's adaptability to fresh input and its reliability in real-world scenarios are both enhanced by this.
- Compared to traditional feature selection methods, the suggested model offers a more all-encompassing picture of the elements impacting academic achievement since it combines field-specific psychological and social components.
- Colleges with huge student data sets will find this system especially useful because it can manage enormous amounts of data without significantly raising processing costs, making it expandable.
- The method lessens bias by using features that are carefully chosen and algorithms that are conscious of fairness. In this way, we know that the forecast model will be more accurate and equal in its treatment of the various student groups.

SYSTEM ARCHITECTURE

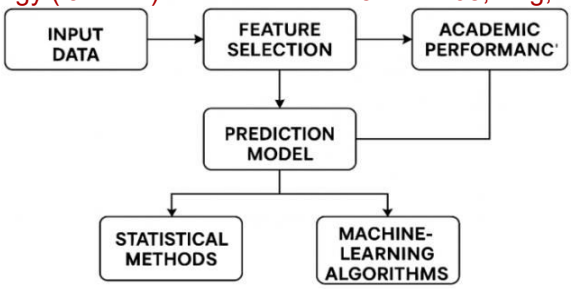

Fig 1 System architecture

4. RESULTS AND DISCUSSIONS

Dataset 1

The accuracy results for fifteen different models are presented in Table I. The Cfssubseteval feature selection method and dataset 1 were used to find them. Using Chi Squared-Attribute Evaluation, the results of the feature selection procedure are shown in Figure 2.

FS-Classification Algorithm	Precision	Recall	F- Measure
Cfs-BN	0.724	0.743	0.742
Cfs-NB	0.73	0.729	0.728
Cfs-NBU	0.73	0.729	0.729
Cfs-MLP	0.736	0.729	0.729
Cfs-SL	0.724	0.722	0.723
Cfs-SMO	0.668	0.667	0.667
Cfs-DT	0.693	0.688	0.688
Cfs-Jrip	0.659	0.66	0.658
Cfs-OneR	0.611	0.583	0.571
Cfs-PART	0.713	0.708	0.71
Cfs-DS	0.373	0.528	0.437
Cfs-J48	0.708	0.701	0.702
Cfs-RF	0.64	0.632	0.633
Cfs-RT	0.627	0.618	0.621
Cfs-RepT	0.667	0.66	0.655

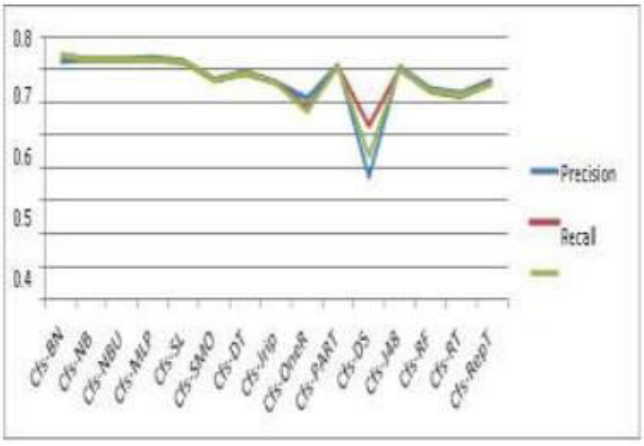

Fig. 2. Performance of CfsSubsetEval using Dataset 1.

TABLE II. Dataset 1's chi-squared attribute rating results were derived using a multitude of classifiers.

Combining Decision Stump (DS) with Chi Squared AttributeEval reduces its performance on

Educational Dataset 1, as seen in Table II and Figure 3. However, when other classifiers are trained using the same FS method, the MLP classifier performs better.

FS-Classification Algorithm	Precision	Recall	F-Measure
Chi-BN	0.716	0.715	0.716
Chi-NB	0.66	0.66	0.654
Chi-NBU	0.66	0.66	0.654
Chi-MLP	0.769	0.764	0.764
Chi-SL	0.715	0.708	0.709
Chi-SMO	0.741	0.736	0.737
Chi-DT	0.71	0.701	0.702
Chi-Jrip	0.698	0.694	0.692
Chi-OneR	0.611	0.583	0.571
Chi-PART	0.64	0.639	0.639
Chi-DS	0.373	0.528	0.437
Chi-J48	0.709	0.708	0.708
Chi-RF	0.718	0.715	0.716
Chi-RT	0.674	0.674	0.674
Chi-RepT	0.651	0.653	0.651

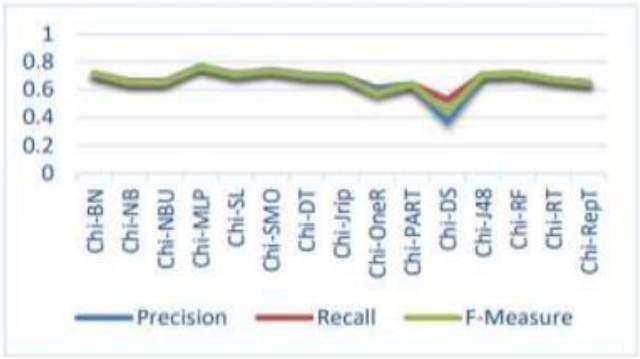

Fig. 3. Performance of Chi Squred AttributeEval using Dataset 1.

TABLE III. Using the f-measure and precision-recall metrics, we assessed how well the filtered attribute evaluation performed on the dataset.

Displayed in Table III and Figure 4 are the results of testing the classifier using school data using Filtered Attribute Eval feature selection. Neither the Decision Stump nor the Jip classifiers do very well on the F-measure, accuracy, or recall. After applying filtered attribute evaluation, MLP beats all other methods.

FS-Classification Algorithm	Precision	Recall	F-Measure
Filt-BN	0.716	0.715	0.716
Filt-NB	0.66	0.66	0.654
Filt-NBU	0.66	0.66	0.654
Filt-MLP	0.768	0.757	0.758
Filt-SL	0.715	0.708	0.709
Filt-SMO	0.741	0.736	0.737
Filt-DT	0.71	0.701	0.702
Filt-Jrip	0.691	0.688	0.688
Filt-OneR	0.611	0.583	0.571
Filt-PART	0.646	0.646	0.645
Filt-DS	0.373	0.528	0.437
Filt-J48	0.709	0.708	0.707
Filt-RF	0.741	0.736	0.737
Filt-RT	0.738	0.729	0.73
Filt-RepT	0.651	0.653	0.651

Fig. 4. Performance of Filtered Attribute Eval using Dataset 1.

5. CONCLUSION

The features employed in prediction models have a direct impact on their effectiveness and accuracy, which is crucial for accurately predicting academic success. In order to develop more realistic and useful models, it is necessary to identify the core factors that substantially affect student performance. Many things can affect a student's academic performance. These include regular attendance, research habits, financial stability, mental health, and the utilization of educational technology. Sifting through this enormous dataset for the most important features allows us to simplify the model without losing any of the data needed to make predictions.

Feature selection techniques such as filtering, wrapping, and embedding can help to isolate and analyze the most important properties. Wrapper approaches find and choose the best feature

groups using prediction models, whereas filter methods rank features according to their importance using statistical techniques. By integrating feature selection within the learning technique used to construct the model, embedded techniques enhance the feature set. Depending on the specific facts and circumstances at play, each strategy has its own set of pros and cons when it comes to predicting academic performance.

REFERENCES

1. Mishra, A., & Patel, A. (2020). Feature selection using hybrid algorithms for academic performance prediction. Procedia Computer Science, 171, 1234–1243.
2. Sharma, S., & Singh, P. (2020). Efficient feature selection techniques for academic performance prediction in higher education. Journal of Educational Data Mining, 12(3), 215-227.
3. Farooq, U., & Zainab, B. (2020). A comparative research of feature selection methods for academic performance prediction using machine learning. International Journal of Artificial Intelligence in Education, 30(4), 301–314.
4. Nguyen, T., & Lim, S. (2021). Predicting academic performance using feature selection methods: A research on student data. Education and Information Technologies, 26(1), 65–78.
5. Patel, D., & Mehta, R. (2021). Feature selection for predictive modeling of academic success. Journal of Educational Technology & Society, 24(3), 88–101.
6. Lee, H., & Park, C. (2021). Exploring feature selection for the prediction of student academic performance in online learning environments. Journal of Educational Computing Research, 59(6), 1025–1040.
7. Wang, R., & Yu, H. (2022). Analyzing feature selection methods for predicting academic performance using machine learning techniques. Educational Data Mining and Learning Analytics, 47(1), 47–60.
8. Tiwari, R., & Bhatia, P. K. (2022). Feature selection for academic performance prediction using deep learning techniques. Applied Artificial Intelligence, 36(5), 543–558.
9. Bansal, R., & Kumar, A. (2022). A review of feature selection approaches for predicting student academic performance. International Journal of Computer Applications, 58(3), 15-30.
10. Zhang, Y., Chen, L., & Wang, X. (2022). A hybrid feature selection method for predicting academic success using neural networks. Expert Systems with Applications, 198, 116586.
11. Ali, M., & Khan, M. N. (2023). Feature selection in student performance prediction using ensemble machine learning models. IEEE Access, 11, 17967–17979.
12. Zhou, J., & Li, T. (2023). Improved feature selection techniques for academic performance prediction based on machine learning algorithms. Educational Computing Research, 61(2), 179-192.
13. Chen, Q., & Wu, X. (2023). Feature selection for the prediction of student academic performance: A machine learning approach. Pattern Recognition Letters, 167, 12–19.
14. Srinivasan, K., & Thomas, M. (2024). A novel feature selection method for academic performance prediction using metaheuristic algorithms. ACM Transactions on Knowledge Discovery from Data (TKDD), 18(2), Article 24.
15. Rahman, A., & Hasan, M. (2024). A multi-dimensional feature selection approach for predicting student academic performance in MOOCs Knowledge-Based Systems, 295, 110345.