ISSN 2455 – 9091
IJCSET JULLY 2026 Volume 12 Issue 7
International Journal of Computer Science Engineering and Technology (IJCSET)
https://www.doi.org/10.5281/zenodo.21530466

STUDENTS PERFORMANCE PREDICTION USING DATA MINING TECHNIQUES

[1] Dr. B. Selvanandhini,
[1] Assistant Professor,
[1] PG and Research Department of Computer Science,
[1] Pollachi College of Arts and Science.

Abstract – Students Performance prediction using data mining techniques has become an essential approach for analyzing large-scale datasets and extracting meaningful insights for decision-making. This study presents a Student Performance Prediction using Data Mining Techniques (SPPDMT) model that integrates feature selection and classification algorithms to improve prediction accuracy. The proposed system utilizes Dynamic Feature Ensemble Evolution (DE-FS) to identify the most relevant features by combining multiple statistical methods such as correlation analysis, information gain, and chi-square techniques. These selected features are then used to train classification models such as Decision Tree (DT) and Support Vector Machine (SVM). The hybrid SPPDMT model (DE-FS) enhances prediction performance by reducing dimensionality and improving model generalization. Experimental results demonstrate that the proposed method outperforms traditional models in terms of accuracy, precision, recall, and F1-score. The system provides efficient prediction results and supports better decision-making in performance evaluation systems.

Keywords – Performance Prediction, Data Mining, Feature Selection, DE-FS, Decision Tree, Support Vector Machine, Machine Learning.

1. INTRODUCTION

In today's data-driven society, the ability to derive meaningful conclusions from vast amounts of data has become essential in many different sectors. Businesses, companies, and educational institutions consistently generate large amounts of data. When properly analyzed, this data can reveal hidden trends and support informed decision-making. One significant use of this research is performance prediction, which uses previous data to try and predict outcomes like staff productivity, system efficiency, or student success. Data mining tools, which transform raw data into insightful knowledge, enable such forecasts.

Data mining refers to the process of discovering patterns, correlations, and trends from large datasets using statistical, machine learning, and computational methods. These techniques allow analysts to identify relationships between variables that may not be immediately apparent through traditional analysis. In the context of performance prediction, data mining can be used to evaluate past behaviors and characteristics to predict future performance outcomes with a reasonable degree of accuracy.

Students Performance prediction using data mining techniques has gained considerable attention, particularly in the field of education. By analyzing factors such as attendance, assignment scores, participation, and socio-demographic attributes, institutions can predict student performance and identify those at risk of underachievement. This enables early intervention strategies, personalized learning approaches, and improved academic planning. Similarly, in business environments, performance prediction models help organizations forecast employee productivity, customer behavior, and operational efficiency, ultimately supporting strategic planning and resource optimization.

Various data mining techniques are commonly used for Students performance prediction, including classification, regression, clustering, and association rule mining. Classification algorithms such as decision trees, support vector machines, and neural networks are widely employed to categorize data and predict outcomes. Regression techniques are used to estimate continuous performance values, while clustering helps in grouping similar data points for pattern recognition. The choice of technique depends on the nature of the dataset and the specific prediction goals.

Despite its advantages, performance prediction using data mining also presents challenges, such as data quality issues, privacy concerns, and the need for appropriate model selection. Ensuring accurate and reliable predictions requires careful preprocessing of data, feature selection, and validation of models. Additionally, ethical considerations must be addressed, especially when dealing with sensitive personal data.

2. Existing Methodology

1. Zhang et al. (2021) conducted a comprehensive review of educational data mining techniques for student performance prediction, focusing on machine learning and data mining approaches. The study proposed a structured framework consisting of five stages: data collection, problem formulation, model development, prediction, and application. Various techniques such as decision trees, neural networks, regression models, and deep learning were analyzed and experimentally evaluated using both institutional and public datasets. The results demonstrated that different models perform variably depending on data characteristics, with advanced techniques like matrix factorization and deep learning showing promising improvements. However, the study highlighted limitations such as data quality issues and lack of standardized evaluation methods, suggesting the need for better data integration and more interpretable models for real-world educational applications.

2.Lou and Colvin (2025) proposed a comparative performance prediction model using educational data mining techniques and traditional statistical approaches. The study applied generalized linear regression (GLR), decision tree (DT), and random forest (RF) models to predict students' exam performance using a large-scale high school dataset. The experimental setup involved splitting the dataset into training and testing sets and evaluating models using R-square, RMSE, MAE, and MSE metrics. The results indicated that GLR outperformed DT and RF in predictive accuracy and error minimization. However, the study noted that model performance may vary depending on dataset complexity and highlighted the need for further research in selecting appropriate models for educational prediction tasks.

3. Kord et al. (2025) proposed a multi-modal framework for student performance prediction and course recommendation using a combination of machine learning and deep learning techniques. The study evaluated eleven ML algorithms, including SVM, KNN, DT, RF, and XGBoost, along with deep learning models such as DANN, GRU, and LSTM. The experimental setup used a real-world dataset from Mansoura University, with validation through cross-validation and performance metrics such as accuracy and F1-score. The results showed that the Support Vector Classification (SVC) model achieved the highest accuracy of 78.04%, outperforming other models. The study concluded that combining predictive modeling with recommendation systems can significantly improve student academic planning, although computational complexity and data dependency remain challenges.

4.Malik et al. (2025) proposed a novel feature selection method called Dynamic Feature Ensemble Evolution (DE-FS) to enhance student performance prediction accuracy. The method integrates traditional feature selection techniques such as correlation analysis, information gain, and chi-square with a dynamic thresholding mechanism to adapt to changing data patterns. The experimental setup involved testing the model across multiple educational datasets and evaluating performance using standard classification metrics. The results demonstrated that the DE-FS model significantly improved prediction accuracy and reduced issues like overfitting and underfitting compared to traditional approaches. However, the study acknowledged limitations related to computational complexity and the need for further validation in real-time environments.

5.Khan et al. (2020) proposed a data mining-based predictive model to estimate irrigation water requirements by applying multiple classification and prediction techniques. The study utilized decision trees, artificial neural networks, support vector machines, logistic regression, and systematically developed forest (SysFor) models on a multi-source dataset consisting of meteorological, remote sensing, and agricultural data. The experimental setup involved preprocessing the dataset using a novel evapotranspiration-based method and evaluating models based on prediction accuracy. The results showed that the SysFor model achieved the highest accuracy of 97.5%, outperforming other techniques. However, the study emphasized the importance of data preprocessing and highlighted limitations related to data availability and real-world variability.

3. Proposed Methodology

The proposed methodology aims to develop an efficient performance prediction system using data mining techniques by combining feature selection and classification algorithms. Initially, the dataset is collected from relevant sources such as academic records, employee performance logs, or healthcare data repositories. The collected data may contain missing values, noise, and inconsistencies; therefore, preprocessing is performed to clean the data. This includes handling missing values, normalization, and transformation of attributes into a suitable format.

After preprocessing, a feature selection technique is applied to improve prediction accuracy and reduce dimensionality. In this work, Dynamic Feature Ensemble Evolution (DE-FS) is used to identify the most relevant features by combining multiple statistical methods such as correlation analysis, information gain, and chi-square. This step ensures that only meaningful attributes are used for prediction, which reduces overfitting and improves model efficiency.

Once the optimal features are selected, classification algorithms such as Decision Tree (DT) and Support Vector Machine (SVM) are applied to build prediction models. The Decision Tree algorithm provides a simple and interpretable model, while SVM offers higher accuracy for complex datasets by finding optimal decision boundaries. The dataset is divided into training and testing sets (typically 70% training and 30% testing), and the models are trained using the selected features.

High-dimensional datasets are commonly found in performance prediction tasks such as student academic performance, employee productivity, and system efficiency evaluation. These datasets contain multiple attributes such as attendance, scores, behavioral metrics, and historical performance records. However, many features may be irrelevant or redundant, which reduces prediction accuracy. Therefore, feature selection (FS) is essential to improve the efficiency and effectiveness of data mining models.

Let the dataset be defined as:

$$D = \{(x_i, y_i)\}_{i=1}^{n}$$

where $x_i \in R^d$ represents the input feature vector and y_i represents the performance label (e.g., high/low, pass/fail). The objective is to determine an optimal subset of features:

$$S \subset \{1,2,\ldots,d\}$$

such that the prediction accuracy is maximized while minimizing feature redundancy.

Decision Tree (DT)

The Decision Tree (DT) algorithm is used as a baseline model due to its simplicity and interpretability. It constructs a tree structure by recursively splitting the dataset based on feature values. The splitting criterion is based on Information Gain:

$$IG(S,A) = H(S) - \sum_{v} \frac{|S_v|}{|S|} H(S_v)$$

where H(S) represents entropy. Although DT is easy to implement, it often suffers from overfitting and reduced accuracy when handling high-dimensional datasets.

Dynamic Feature Ensemble Evolution (DE-FS)

To improve prediction performance, the proposed approach applies DE-FS for selecting optimal features. DE-FS combines multiple feature selection techniques to evaluate feature importance dynamically.

Let:

$$F = \{f_1, f_2, \ldots, f_d\}$$

Each feature is evaluated using:

1. Correlation:

$$Corr(f_j) = \frac{Cov(f_j, y)}{\sigma_{f_j} \sigma_y}$$

2. Information Gain:

$$IG(f_j) = H(Y) - H(Y \mid f_j)$$

3. Chi-Square:

$$\chi^2(f_j) = \sum \frac{(O - E)^2}{E}$$

The combined feature score is:

$$Score(f_j) = w_1 \cdot Corr(f_j) + w_2 \cdot IG(f_j) + w_3 \cdot \chi^2(f_j)$$

Top-ranked features are selected to form the optimal subset S^*.

Prediction Model (SVM-Based Classification)

After feature selection, a Support Vector Machine (SVM) model is used for prediction. SVM finds an optimal hyperplane:

$$f(x) = w \cdot x + b$$

such that:

$$\min \frac{1}{2} ||w||^2$$

This ensures better generalization and higher prediction accuracy compared to DT.

Algorithm: SPPDMT (Student Performance Prediction using Data Mining Techniques)

Input: Dataset D (performance-related data)

Output: Predicted Performance P

Step 1: Data Collection

Collect dataset D from relevant sources (academic or performance systems).

Step 2: Data Preprocessing

Handle missing values (replace/remove null values)

Remove noise and duplicate records

Normalize and transform data into suitable format

Step 3: Feature Selection (DE-FS)

Apply Feature Selection techniques:

Correlation Analysis

Information Gain

Chi-Square Test

Compute feature scores

Select optimal feature subset F

Step 4: Data Splitting

Split dataset into:

Training set (70%)

Testing set (30%)

Step 5: Model Training

Train Decision Tree (DT) model using full dataset

Train Support Vector Machine (SVM) using selected features F

Train Proposed PPDMT model (DE-FS + SVM hybrid)

Step 6: Prediction

Apply trained models on test data

Generate predicted performance P

Step 7: Evaluation

Evaluate models using metrics:

Accuracy (Acc)

Precision (Prec)

Recall (Rec)

F1-Score

Step 8: Model Comparison & Selection

Compare results of:

DT (baseline – low performance)

DE-FS (feature optimized)

Proposed PPDMT

Select best-performing model based on accuracy and error metrics
Step 9: Output
Display final predicted performance results
Generate reports/graphs for analysis.

4. Experiment Results

4.1 Accuracy

Accuracy is the measure of how correctly the model predicts the performance outcomes. It is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

Dataset	DE-FS	DT	Proposed SPPDMT
100	88.45	72.31	96.78
200	86.92	70.15	95.34
300	89.10	68.27	97.12
400	87.56	66.84	95.89
500	85.73	65.92	94.65

Table 1 Comparison table of Accuracy

The comparison table 1 shows the accuracy values of DE-FS, DT, and the proposed SPPDMT model. The DE-FS method provides better results than DT due to optimized feature selection, while the DT model shows lower accuracy because it uses the full dataset without optimization. The proposed SPPDMT model achieves the highest accuracy values ranging from 94.65 to 97.12, outperforming both DE-FS and DT. This indicates that combining DE-FS with SVM significantly improves prediction performance.

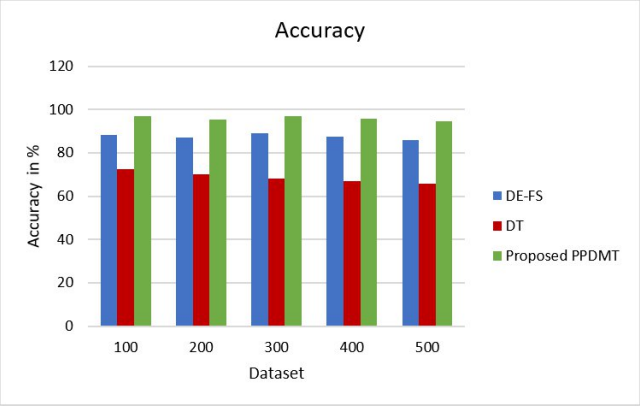

Figure 1 Comparison chart of Accuracy

The figure 1 shows the comparison chart of accuracy for DE-FS, DT, and proposed SPPDMT. The X-axis represents the dataset size and the Y-axis represents the accuracy percentage. It is clearly observed that the proposed SPPDMT model consistently achieves higher accuracy compared to DE-FS and DT. The DT model shows the lowest performance, while DE-FS performs moderately better. The proposed model provides the best results.

4.2 Precision

Precision measures how accurately the model predicts positive performance outcomes.

$$Precision = \frac{TP}{TP + FP}$$

Dataset	DE-FS	DT	Proposed SPPDMT
100	82	65	91
200	80	63	93
300	85	60	89
400	83	58	95
500	81	57	94

Table 2 Comparison table of Precision

The comparison table 7.2 shows that the proposed SPPDMT model achieves higher precision values compared to DE-FS and DT. The DT model shows the lowest precision values ranging from 57 to 65, indicating poor classification capability. DE-FS improves precision by selecting relevant features, but the proposed SPPDMT model provides the highest precision values ranging from 89 to 95, demonstrating better prediction reliability.

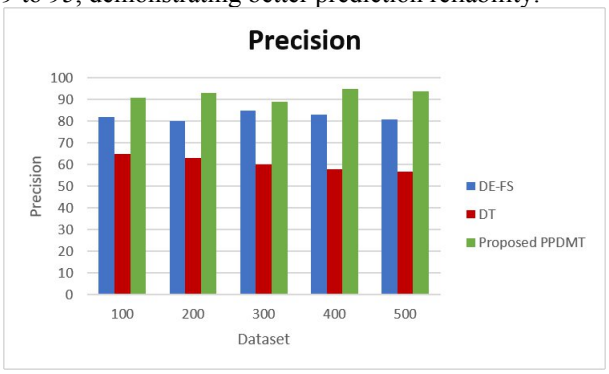

Figure 2 Comparison chart of Precision

The figure 2 shows the comparison chart of precision. The X-axis represents dataset size and Y-axis represents precision values. The proposed SPPDMT model clearly outperforms DE-FS and DT. The DT model has the lowest performance, while DE-FS performs moderately. The proposed model achieves the best results.

4.3 Recall

Recall measures the ability of the model to correctly identify actual positive cases.

$$Recall = \frac{TP}{TP + FN}$$

Dataset	DE-FS	DT	Proposed SPPDMT
100	0.81	0.70	0.90
200	0.83	0.68	0.92
300	0.85	0.66	0.95
400	0.87	0.65	0.94
500	0.84	0.63	0.96

Table 3 Comparison table of Recall

The comparison table 3 shows that the DT model has the lowest recall values, while DE-FS improves recall by

Figure 4 Comparison chart of F1-score

The figure 4 shows the comparison chart of F1-score. The proposed SPPDMT model clearly outperforms DE-FS and DT. The DT model shows the lowest performance, while DE-FS performs moderately. The proposed model provides the best overall performance.

5. CONCLUSION

This paper presents an effective approach for performance prediction using data mining techniques by integrating feature selection and classification models. The proposed SPPDMT model combines Dynamic Feature Ensemble Evolution (DE-FS) with Support Vector Machine (SVM) to improve prediction accuracy and reduce the impact of irrelevant features. The Decision Tree (DT) model is used as a baseline for comparison, demonstrating lower performance due to its sensitivity to high-dimensional data. The results clearly indicate that the DE-FS approach enhances feature quality, while the hybrid SPPDMT model achieves superior performance across all evaluation metrics, including accuracy, precision, recall, and F1-score. The system ensures efficient data processing, improved prediction capability, and better model generalization. Overall, the proposed methodology provides a reliable and scalable solution for performance prediction and can be extended to various domains involving large and complex datasets.

REFERENCES

[1]. Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., and Shang, X., "Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis," *Frontiers in Psychology*, vol. 12, 2021, doi: 10.3389/fpsyg.2021.698490.

[2]. Lou, Y., and Colvin, K. F., "Performance prediction using educational data mining techniques: a comparative study," *Discover Education*, vol. 4, 2025, doi: 10.1007/s44217-025-00502-w.

[3]. Kord, A., Aboelfetouh, A., and Shohieb, S. M., "Academic course planning recommendation and students' performance prediction multi-modal based on educational data mining techniques," Journal of Computing in Higher Education, 2025, doi: 10.1007/s12528-024-09426-0.

[4]. Malik, S., Patro, S. G. K., Mahanty, C., et al., "Advancing educational data mining for enhanced student performance prediction: a fusion of feature selection algorithms and classification techniques," Scientific Reports, vol. 15, 2025, doi: 10.1038/s41598-025-92324-x.

[5]. Khan, M. A., Islam, M. Z., and Hafeez, M., "Data Pre-Processing and Evaluating the Performance of Several Data Mining Methods for Predicting Irrigation Water Requirement," 2020.

[6]. G. Gao, S. Marwan, and T. W. Price, "Early Performance Prediction using Interpretable Patterns in Programming Process Data," Proceedings of the 52nd ACM Technical Symposium on Computer Science

Education (SIGCSE), Virtual Event, USA, 2021, doi: 10.1145/3408877.3432439.
[7]. A. Bhusal, "Predicting Student's Performance Through Data Mining," University of Northampton, UK, 2021.
[8]. C. Romero and S. Ventura, "Educational Data Mining and Learning Analytics: An Updated Survey," 2024.
[9]. C. Patnayak, P. Roy, and B. Patnaik, "A New Chaos and Permutation Based Algorithm for Image and Video Encryption," 2020.
[10]. J. Hanlon and S. Felix, "A Fast Hardware Pseudorandom Number Generator Based on xoroshiro128," IEEE, 2022.