International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:
3048-7722
Volume 4, Issue 3, May-june--2026, PP: 01-09

Student Performance Prediction Using Machine Learning

Venugopal S[1], Dr. P. Jeyanthi[2]
[1](Student Information Technology
Sri Ramakrishna College of Arts & Science,
Nava India Avinashi Road, Coimbatore – 641006, Tamil Nadu, India
smvenugopal410@gmail.com)
[2](Dr. P. Jeyanthi
Assistant Professor
Department of Information Technology Sri Ramakrishna College of Arts & Science, Coimbatore, India
jeyanthi@srcas.ac.in)

***Abstract*- Predicting student performance has grown in importance as a field of study in education analytics and machine learning. Large volumes of student data, including attendance, grades, assignments, and behavioral records, are produced by educational institutions. Conventional assessment techniques frequently fail to spot weak students early on. A machine learning-based method for forecasting students' academic success is presented in this research. In order to handle missing values and normalize attributes, the system gathers and preprocesses student data. The most pertinent elements influencing student performance are found using feature selection strategies. Prediction is done using machine learning algorithms like Decision Tree, Random Forest, SVM, and ANN. Students are divided into three performance categories by the trained models: High, Medium, and Low. Additionally, the suggested approach produces likelihood scores for predicting academic performance. According to experimental findings, ANN and Random Forest outperform conventional techniques in terms of prediction accuracy. The approach assists teachers in identifying kids who are at danger and in promptly offering academic support. Additionally, it helps schools raise student success rates and overall educational quality.**

Keywords: **Artificial Neural Network (ANN), Random Forest, Predictive Analytics, Machine Learning, Educational Data Mining, and Student Performance Prediction.**

I. INTRODUCTION

Large volumes of student-related data have been produced in recent years due to the quick expansion of digital education systems [1]. Attendance data, test scores, assignment performance, classroom involvement, study habits, and socioeconomic information are all gathered by educational institutions. Effective data analysis can help teachers better understand how students learn and enhance academic performance. However, traditional methods of assessing student performance mostly rely on physical observation and periodic exams, which are frequently time-consuming and less accurate in identifying academically poor kids at an early stage[2].Educational institutions are progressively implementing intelligent technologies to automate academic analysis and prediction duties as machine learning advances [3].Machine learning methods can examine past student data, spot hidden trends, and make highly accurate predictions about future academic achievement. Teachers and administrators can identify at-risk kids early on and offer appropriate academic support, such as counseling, mentorship, and individualized learning plans, with the use of these prediction systems [4].

Because it enhances decision-making in contemporary educational institutions, student performance prediction is a significant application of educational data mining [5]. Academic achievement is greatly influenced by a number of factors, including attendance, prior grades, study habits, family history, classroom interaction, and assignment completion. Machine learning algorithms can categorize students into several performance groups, such as High,

International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:
3048-7722
Volume 4, Issue 3, May-june--2026, PP: 01-08

Medium, and Low performers, by using these parameters as input features [6].

In order to accurately forecast student performance, this research suggests a machine learning-based method that employs feature selection, data preprocessing, and classification techniques. The system analyzes student data and produces prediction results using methods including Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN) [7]. By eliminating unnecessary and duplicated attributes from the dataset, feature selection techniques increase efficiency.

The goal of the suggested approach is to help teachers increase student success rates and improve their teaching methods. Institutions can take preventive action before students face serious academic challenges thanks to early prediction of academic performance [8]. In comparison to conventional statistical methods, experimental study demonstrates that sophisticated machine learning models, especially ANN and Random Forest, yield superior prediction accuracy.

II. LITERATURE SURVEY

Due to the quick advancement of machine learning and intelligent educational systems, student performance prediction has received a lot of interest lately [1]. Numerous methods have been put forth by researchers to evaluate student academic data and make precise predictions about future performance. These prediction methods are used by educational institutions to increase overall academic quality and detect poor pupils early.

Traditional statistical techniques like logistic regression and linear regression were the mainstay of early student performance analysis research [2]. These methods estimated academic outcomes using variables like attendance, exam scores, and assignment scores. Despite being easy to use, statistical models frequently failed to capture intricate correlations between various student variables, which led to decreased prediction accuracy.

Researchers started using classification algorithms including Decision Tree, Naive Bayes, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) for educational data analysis as machine learning techniques advanced [3]. Decision tree algorithms gained popularity due to their interpretability and ease of usage. However, when trained on big datasets, Decision Trees may experience overfitting issues. Later, Random Forest algorithms were developed to decrease overfitting and increase prediction accuracy [4]. According to a number of studies, Random Forest predicts student academic performance more accurately than conventional classifiers. Because Support Vector Machine (SVM) works well on high-dimensional educational datasets, it has also been employed extensively [5].

Deep learning techniques like Long Short-Term Memory (LSTM) models and Artificial Neural Networks (ANN) are the subject of recent research [6]. ANN algorithms produce extremely accurate predictions and can recognize intricate nonlinear correlations among student variables. When enough training data is available, researchers discovered that ANN-based systems perform better than many conventional machine learning techniques.The significance of feature selection methods in enhancing prediction performance has also been highlighted by a number of researchers [7]. To eliminate unnecessary features and lower computing complexity, techniques including correlation analysis, Principal Component Analysis (PCA), and Genetic Algorithms are employed. Even with these developments, many systems continue to have drawbacks such poor interpretability, problems with real-time prediction, and large computing demands [8].

III. METHODOLOGY:

The suggested system for predicting student success is based on a structured machine learning framework that is intended to evaluate educational data and provide precise predictions about students' academic performance. Data collection, data preparation, feature selection, model training, prediction, and evaluation are some of the steps that make up the technique. Every step is essential to raising the prediction system's effectiveness and precision.

A. Information Gathering

Gathering student-related data from educational institutions or publicly accessible academic databases is the first stage in the suggested process [1]. The dataset includes a number of characteristics that affect students' academic achievement. These qualities consist of:

- The percentage of attendance
- Internal evaluation scores
- The results of the assignment

International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:
3048-7722
Volume 4, Issue 3, May-june--2026, PP: 01-08

• Grades from prior semesters
• Hours spent studying
• Engagement in the classroom
• Family history
• Access to the internet and educational materials
• Extracurricular pursuits

For additional analysis, the gathered data is kept in an organized format like database tables or CSV.

B. **Preprocessing Data**

Incomplete, inconsistent, or noisy data are frequently found in educational datasets. Preprocessing is therefore required prior to using machine learning techniques [2]. The following tasks are part of the preparation stage:

1. Managing Missing Data

Statistical techniques like the following are used to find and replace missing values:
• Average replacement
• The median substitute
This procedure reduces prediction errors and enhances data consistency.

2. Cleaning Data

To enhance the quality of the dataset, duplicate records and unnecessary attributes are eliminated. Additionally, outliers that have a detrimental impact on model performance are identified and removed.

3. Transformation of Data

Encoding techniques like these are used to translate categorical attributes like gender, parental education, and school type into numerical form: Label Encoding One-Hot Coding

4. Normalization of Data

To guarantee that every attribute contributes equally during model training, numerical features are normalized using Min-Max Scaling or Standardization [3].

C. Feature Selection

Finding the most pertinent characteristics affecting student performance is a crucial stage in the feature selection process [4]. Prediction accuracy is increased and computational complexity is decreased by choosing significant features.The proposed system uses:

- Correlation Analysis
- Principal Component Analysis (PCA)
- Information Gain Method

While duplicate attributes are eliminated, highly connected and important aspects are kept. Attendance, internal grades, and study time are among the most important variables influencing academic success, according to experimental studies.

Fig. 1. Student Performance Prediction System Workflow

D. Dataset Splitting

The dataset is split into two sections following feature selection and preprocessing: • 70%–80% of the training dataset • 20%–30% of the testing dataset Machine learning models are trained using the training dataset, while model performance and prediction accuracy are assessed using the testing dataset.

E. Machine Learning Model Training

Several machine learning methods are used by the suggested system to compare prediction performance and determine which model performs the best.

International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:
3048-7722
Volume 4, Issue 3, May-june--2026, PP: 01-08

1. Decision Tree Classifier

Students are categorized using a tree-like structure of decision criteria by the Decision Tree algorithm [5]. It is simple to comprehend and analyze. Based on significant characteristics like attendance and grades, the algorithm divides the dataset into branches.

Advantages:

- Simple implementation
- Fast prediction
- Easy visualization

Limitations:

- Overfitting problem on large datasets

2. Random Forest Algorithm

Several decision trees are combined in the Random Forest ensemble learning technique to increase prediction accuracy and decrease overfitting [6].

Advantages:

- High prediction accuracy
- Robust against noisy data
- Better generalization capability

The final prediction is obtained using majority voting among multiple decision trees.

3. Support Vector Machine (SVM)

By identifying the best hyperplane to divide various student performance categories, Support Vector Machines are employed for classification [7]. Benefits • Suitable for datasets with many dimensions • Excellent categorization results Utilizing kernel functions, it effectively handles nonlinear data. Based on academic characteristics, the SVM model divides students into High, Medium, and Low performers.

4. Artificial Neural Network (ANN)

Artificial neural networks use interconnected neurons to mimic how the human brain functions [8]. ANN models are very good at finding intricate patterns in data related to education. This system's ANN architecture includes: • The Input Layer
• Undiscovered Layers • The Output Layer

The network is trained using the backpropagation algorithm to minimize prediction error.

Advantages:

- High accuracy
- Learns nonlinear relationships
- Suitable for large datasets

Experimental results show that ANN achieves better performance compared to traditional machine learning algorithms.

F. Performance Prediction

After model training, the trained classifiers predict student academic performance categories such as: • Outstanding Performer
• The Average Performer
• Poor Performance
Teachers and administrators can identify academically disadvantaged pupils early on and offer suitable academic advice thanks to the prediction results.

G. Performance Evaluation Metrics

Standard evaluation metrics are used to assess machine learning models' performance [5]:

1. Correctness

The percentage of accurately predicted cases is known as accuracy.

Precision = (TP+TN)/(TP+TN+FP+FN)

2. Accuracy

The accuracy of optimistic forecasts is measured by precision.
Precision is equal to TP/FP.

3. Remember

Recall quantifies the model's capacity to recognize good examples.

Recall = TP/TP + FN

4. The F1-Score

The precision and recall harmonic mean is given by the F1-Score.

F1"-" Score = ×(Precision×Recall)/(Precision+Recall)

5. Matrix of Confusion

By contrasting actual and anticipated outcomes, a confusion matrix is utilized to display the prediction model's categorization performance.

International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:
3048-7722
Volume 4, Issue 3, May-june--2026, PP: 01-08

IV. IMPLEMENTATION

Modern software tools and machine learning approaches are used to construct the suggested student performance prediction system. The system is made to effectively gather, preprocess, evaluate, and forecast student academic performance. Dataset preparation, preprocessing, feature selection, model training, prediction generation, and result visualization make up the implementation process.

A. Utilized Tools and Software

The following software tools and technologies are used in the implementation of the suggested system: Python is the programming language.

- Machine Learning Libraries: TensorFlow, Keras, and Scikit-learn
- Data Processing Libraries: NumPy, Pandas
- Visualization Tools: Seaborn and Matplotlib Flask is the web framework.
- Development Environment: Visual Studio Code and Jupyter Notebook

Python was chosen due of its ease of use and broad support for data analysis and machine learning packages.

B. Dataset Preparation

The dataset utilized in this study includes behavioral and academic data about students gathered from academic records. Important characteristics consist of:

• The percentage of attendance
• Results from prior exams
• The results of the assignment
• Hours spent studying
• Engagement in the classroom
• Socioeconomic status

The Pandas library is used to load the CSV-formatted dataset into the system. To comprehend data distribution and find missing values, exploratory data analysis is carried out after the dataset has been loaded.

C. Data Preprocessing Implementation

Prior to training machine learning models, data preparation is used to enhance dataset quality. Included in the preprocessing phase are:

1. Managing Missing Data
Depending on the kind of attribute, mean, median, or mode values are used to replace missing values in the dataset.

2. Categorical Data Encoding
Label Encoding or One-Hot Encoding procedures are used to translate categorical information, like gender or parental education, into numerical values.

3. Normalization of Data
Min-Max Scaling is used to standardize numerical numbers in order to preserve consistent feature ranges and enhance model performance.

4. Splitting Datasets
An 80:20 ratio is used to separate the dataset into training and testing sets. Machine learning models are trained using the training dataset and evaluated using the testing dataset.

D. Feature Selection Implementation

The most significant characteristics influencing student performance are found using feature selection techniques. The link between attributes and the intended output is measured using correlation analysis. Key features that have been chosen are:

• Participation

• Prior grades

• Hours spent studying

• Completing the assignment

• The degree of participation

Prediction accuracy is increased and computational complexity is decreased by eliminating unnecessary features.

E. Machine Learning Model Implementation

For performance prediction, several machine learning methods are put into practice and contrasted.
1. The Decision Tree
The Scikit-learn library is used to implement the Decision Tree classifier. It creates comprehensible prediction criteria and categorizes students according to academic characteristics.

International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:
3048-7722
Volume 4, Issue 3, May-june--2026, PP: 01-08

2. The Random Forest
By combining several Decision Trees, Random Forest increases the accuracy of predictions. It works effectively on multidimensional educational datasets.

3. Support Vector Machine (SVM):
SVM accurately separates various student performance categories and is used for high-dimensional data classification.

4. ANNs, or artificial neural networks TensorFlow and Keras libraries are used to implement ANN.
The network includes:
•The input layer

• ReLU-activated hidden layers

• Sigmoid/Softmax activation in the output layer

The ANN model generates extremely accurate predictions by learning intricate patterns from student data.

F. Prediction Module
The following student performance levels are predicted using the trained models:
• Excellent Performance

• Moderate Efficiency

• Poor Performance

Additionally, the system generates probability scores that indicate the possibility of academic risk or student success. These forecasts assist teachers in spotting struggling pupils early on.

G. Web Application Implementation
A web application built with Flask is created to offer an intuitive prediction interface. The online interface enables users to:
• Add student datasets.

Enter the student's information.

• Examine the prediction outcomes

Produce reports on performance.

For improved visualization, the prediction results are shown graphically utilizing dashboards and charts.

H. Performance Evaluation
Performance measures like these are used to assess the implemented models:

• Precision

• Accuracy

• Remember

• The F1-Score

• Matrix of Confusion According to experimental findings, Random Forest and ANN outperform conventional machine learning techniques in terms of prediction accuracy. The implementation effectively shows how machine learning techniques may be used to support intelligent educational decision-making systems and forecast students' academic achievement.

V. RESULTS AND DISCUSSION

The proposed student performance prediction system was tested using a variety of machine learning techniques to evaluate prediction accuracy and overall system performance. A dataset of students with behavioral, For experimental analysis, socioeconomic and academic traits were utilized. After the dataset was divided into training and testing sets, 80% of the data was used for training and 20% was used for testing.

Among the machine learning models that have been implemented are decision trees, Random Forests, Support Vector Machines (SVM), and Artificial Neural Networks (ANN). Each model was trained using the preprocessed dataset, and it was evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix.
The results of the experiment show how accurately machine learning algorithms can predict pupils' academic success. The ANN-based prediction model beat all other applicable models due to its ability to comprehend complex nonlinear relationships between student variables. Additionally, Random Forest produced reliable results since it can learn in groups and has less overfitting features.
The performance comparison of different models is shown below:

International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:
3048-7722
Volume 4, Issue 3, May-june--2026, PP: 01-08

Algorithm	Accuracy
Decision Tree	78.45%
Random Forest	88.62%
SVM	84.30%
ANN (Proposed)	92.26%

The results showed that ANN had the best forecast accuracy of around 92.26%, while Random Forest had an accuracy of 88.62%. Decision Tree produced somewhat lower accuracy because of overfitting issues on complex datasets. SVM required more processing power and parameter tweaking despite producing subpar results.

The confusion matrix analysis showed that the proposed ANN model successfully classified most students into the High, Medium, and Low performance categories. Misclassification errors were significantly reduced once preprocessing and feature selection techniques were applied. Student performance was found to be significantly influenced by attendance %, preceding exam results, assignment completion rate, and study hours.
Feature selection techniques greatly improved the system's performance by removing superfluous attributes from the dataset. As a result, prediction accuracy increased and training time was reduced. The use of normalizing techniques during training substantially enhanced the model's stability ,convergence.

The developed Flask web interface efficiently presented prediction results and probability ratings for student performance analysis. Teachers can utilize the data to identify students who have academic difficulties and provide early intervention strategies such as tailored learning support, counseling, and mentorship.

The suggested approach shows how decision-making in education can be greatly enhanced by machine learning-based prediction models. Institutions may lower dropout rates, enhance academic performance, and facilitate data-driven instructional planning by accurately predicting student performance.

Overall, the experimental findings support the effectiveness, dependability, and suitability of thesuggested student performance prediction system for intelligent academic monitoring applications. The proposed method demonstrates how machine learning-based prediction models can significantly improve educational decision-making. By precisely forecasting student success, institutions can improve academic achievement, reduce dropout rates, and enable data-driven instructional planning.

All things considered, the experimental results validate the efficacy, reliability, and appropriateness of the proposed student performance prediction system for intelligent academic monitoring applications.

VI. CONCLUSION AND FUTURE SCOPE

A. Conclusion

The proposed methodology for predicting student achievement demonstrates how machine learning methods may be applied successfully to evaluate educational data and predict academic results for students. The method uses a range of student-related variables, such as attendance records, past exam scores, assignment performance, study hours, involvement level, and socioeconomic background, to identify trends that impact academic progress. The proposed system combines preprocessing, feature selection, and classification algorithms to provide accurate and reliable prediction results.

Several machine learning methods, including Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN), were developed and analyzed using standard evaluation criteria, including accuracy, precision, recall, and F1-score. According to an experimental investigation, the

ANN-based model performed better in terms of prediction accuracy than all other algorithms that were investigated. Additionally, Random Forest produced outstanding results since it is resistant to overfitting and can learn in an ensemble. The results show that educational institutions can identify students who struggle academically early on with the aid of machine learning algorithms.The method has several practical benefits for modern educational environments. By anticipating student performance early on, teachers

can take proactive measures before students encounter significant academic difficulties. Teachers and academic advisers can use prediction results to provide tailored counseling, mentorship, remedial instruction, and learning recommendations. The suggested web-based interface greatly improves usability by allowing users to enter datasets, see prediction results, and efficiently monitor student progress.Furthermore, the proposed framework reduces manual labor in educational record analysis and promotes intelligent academic decision-making processes. Feature selection tactics increased model efficiency by removing unnecessary features, while preprocessing techniques improved dataset quality and prediction stability. The system can therefore be a helpful tool for applications combining smart education and educational data analytics.

There are still other approaches to enhance prediction performance and usefulness in further research, even though the recommended approach yielded promising findings.

B. Scope for future development

The future enhancements of the proposed system include:

Future improvements to the suggested system include:

• Integration with online learning environments and real-time educational administration systems.

• The application of sophisticated deep learning models for sequential learning analysis, such as Transformer-based architectures and Long Short-Term Memory (LSTM).

• The creation of personalized recommendation systems that provide students individualized study schedules and educational materials.

• Explainable AI (XAI) methods are used to increase transparency and aid instructors in comprehending prediction decisions.

• The addition of dashboards for real-time monitoring to track student performance continuously.

• Including mobile applications so that academic notifications and prediction reports are easily accessible.

• Using cloud computing technology to effectively manage massive educational datasets.

• Incorporating behavioral and emotional analysis based on student interaction data to improve forecast accuracy.

• Improving privacy and security measures to safeguard sensitive

In conclusion, the proposed student performance prediction system provides an intelligent, scalable, and efficient solution for academic performance analysis. The integration of machine learning techniques into educational systems can significantly improve student success rates, reduce drorisks, and support data-driven educational planning in modern smart learning environments.

REFERENCES

[1] Romero, C., & Ventura, S., “Educational Data Mining: A Review of the State of the Art,” IEEE Transactions on Systems, Man, and Cybernetics, 2010.

[2] Han, J., Kamber, M., & Pei, J., Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2011.

[3] Tom Mitchell, Machine Learning, McGraw-Hill Education, 1997.

[4] Breiman, L., “Random Forests,” Machine Learning Journal, vol. 45, no. 1, pp. 5–32, 2001.

[5] Vapnik, V., Statistical Learning Theory, Wiley Publications, 1998.

[6] Goodfellow, I., Bengio, Y., & Courville, A., Deep Learning, MIT Press, 2016.

[7] Guyon, I., & Elisseeff, A., “An Introduction to Variable and Feature Selection,” Journal of Machine Learning Research, 2003.

[8] Baker, R., & Inventado, P., “Educational Data Mining and Learning Analytics,” Springer International Publishing, 2014.

[9] Cortez, P., & Silva, A., “Using Data Mining to Predict Secondary School Student Performance,” Proceedings of the 5th Future Business Technology Conference, 2008.

[10] Kotsiantis, S., Pierrakeas, C., & Pintelas, P., “Predicting Students’ Performance in Distance Learning Using Machine Learning Techniques,” Applied Artificial Intelligence, vol. 18, no. 5, pp. 411–426, 2004.

[11] Pandey, M., & Taruna, S., "Towards the Integration of Multiple Classifier Pertaining to the Student's Performance Prediction," Perspectives in Science, vol. 8, pp. 364–366, 2016.
[12] Al-Barrak, M., & Al-Razgan, M., "Predicting Students Final GPA Using Decision Trees: A Case Study," International Journal of Information and Education Technology, vol. 6, no. 7, pp. 528–533, 2016.
[13] Shahiri, A. M., Husain, W., & Rashid, N., "A Review on Predicting Student's Performance Using Data Mining Techniques," Procedia Computer Science, vol. 72, pp. 414–422, 2015.
[14] Dekker, G., Pechenizkiy, M., & Vleeshouwers, J., "Predicting Students Drop Out: A Case Study," International Working Group on Educational Data Mining, 2009.
[15] Hussain, S., Dahan, N., Ba-Alwi, F., & Ribata, N., "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA," Indonesian Journal of Electrical Engineering and Computer Science, vol. 9, no. 2, pp. 447–459, 2018.

Fig. 1: Student Performance Prediction System Workflow