Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

A Machine Learning-Based Approach for Student Performance and Risk Prediction

[1]Megha Sahu, [2]Pawan Kumar
[1]Student, [2]Assistant Professor
[1, 2]Amity University Raipur, Chhattisgarh, India
[1]meghasahu1066@gmail.com
[2]pkumar@rpr.amity.eduAbstract

Abstract:

Student performance prediction has become an important focus in educational data mining, as it helps in identifying students who may be at risk at an early stage and supports better academic decision-making. In many traditional systems, student evaluation mainly depends on exam scores and attendance, which often fail to capture a student's overall performance. To address this limitation, this research proposes a machine learning-based approach that considers multiple factors such as academic performance, behavioral engagement, attendance, and skill-based attributes. The system applies several classification algorithms, including Decision Tree, Random Forest, Logistic Regression, Perceptron, and Multi-Layer Perceptron (MLP) Neural Network, to analyze student data and generate predictions. Before applying these models, the data is preprocessed through steps such as cleaning, encoding, normalization, and feature selection to improve its quality and reliability. The models are then evaluated using standard performance metrics like accuracy, precision, recall, and F1-score. The results show that the MLP Neural Network performs the best among all models, as it can effectively capture complex and non-linear relationships in the data. The proposed system is implemented as a web-based application, allowing users to enter student information and receive predictions instantly. The interface is designed to be simple and user-friendly, and it presents results through visual elements such as dashboards, graphs, and reports. In addition to prediction, the system also provides recommendations based on the results, which makes it more useful for practical applications. Another important aspect of the system is the inclusion of skill-based evaluation. By considering technical skills, academic abilities, and behavioral factors, the system provides a more detailed and realistic analysis of student performance. This multi-dimensional approach improves prediction accuracy and helps in identifying patterns and risk factors more effectively. Moreover, the system is designed to be efficient and scalable, allowing it to handle larger datasets and multiple users without affecting performance. The use of open-source tools and a lightweight design also makes it cost-effective and suitable for educational institutions. The results obtained demonstrate that the system is capable of providing accurate predictions along with meaningful insights. Overall, the proposed approach offers a reliable and practical solution for predicting student performance and identifying at-risk students. It supports early intervention and promotes data-

ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

driven decision-making, which can contribute to improving academic outcomes in educational environments.

Keywords: Student Performance Prediction, Machine Learning, Educational Data Mining, Risk Detection, Classification Algorithms, Data Analysis

1. Introduction

In recent years, the use of data-driven approaches in education has increased significantly, especially with the rise of machine learning and educational data mining. Educational institutions generate large amounts of data related to student performance, behavior, and learning patterns. However, much of this data remains underutilized in traditional systems, where evaluation is often limited to exam scores and attendance records. This creates a gap in understanding the complete picture of a student's performance and makes it difficult to identify students who may be at risk at an early stage.

Student performance prediction has therefore become an important area of research. By analyzing historical and real-time data, it is possible to detect patterns that indicate whether a student is likely to perform well or face difficulties. Early prediction can help educators take timely actions such as providing additional support, improving engagement, or guiding students toward better learning strategies. This not only improves academic outcomes but also enhances overall student development.

The proposed system focuses on addressing these challenges by developing a machine learning-based model that considers multiple aspects of student performance. Unlike traditional approaches, this system does not rely solely on academic marks. It also incorporates behavioral factors such as classroom participation, resource usage, and interaction levels, along with skill-based attributes. The inclusion of skill evaluation makes the system more realistic and capable of analyzing student performance in a comprehensive manner.

To achieve this, various classification algorithms are implemented, including Decision Tree, Random Forest, Logistic Regression, Perceptron, and Multi-Layer Perceptron (MLP) Neural Network. These models are trained on a dataset containing different student attributes and are evaluated using standard performance metrics. The comparison of multiple models ensures that the most suitable algorithm is selected for accurate prediction.

In addition to prediction, the system is designed as a web-based application that provides an interactive interface for users. Teachers or administrators can input student data and instantly obtain prediction results along with insights. The system also presents visual representations such

ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

as graphs and charts, making it easier to understand trends and relationships within the data. This enhances usability and supports better decision-making.

Another important feature of the system is its ability to provide recommendations based on prediction results. Instead of only identifying whether a student is at risk, the system suggests possible actions to improve performance. This makes the system more practical and useful in real-world educational environments.

Overall, this research aims to develop an efficient and reliable system that leverages machine learning techniques to improve student performance analysis. By combining academic, behavioral, and skill-based data, the proposed approach provides a more complete understanding of student performance and contributes to smarter, data-driven educational practices.

Furthermore, the growing importance of personalized learning has increased the need for intelligent systems that can adapt to individual student needs. Every student has a unique learning pattern, and a one-size-fits-all evaluation approach is no longer effective in modern education. By leveraging machine learning, the proposed system can analyze diverse data points and provide insights tailored to each student's performance profile. This supports educators in making informed decisions and encourages a more student-centered approach to teaching and learning. As educational systems continue to evolve, such predictive models can play a crucial role in enhancing both teaching strategies and student outcomes.

2. Literature Review

The use of machine learning in education has gained significant attention in recent years, particularly in the area of student performance prediction. Researchers have explored various techniques to analyze student data and identify patterns that can help in improving academic outcomes. Traditional approaches mainly focused on statistical methods, but with the advancement of machine learning, more accurate and efficient prediction models have been developed.

Several studies have applied classification algorithms such as Decision Trees and Logistic Regression to predict student performance based on academic records and attendance. These methods are simple and easy to interpret, but their performance is often limited when dealing with complex and non-linear data. To overcome these limitations, ensemble methods like Random Forest have been introduced, which combine multiple decision trees to improve prediction accuracy and reduce overfitting.

In recent research, neural network-based models, especially Multi-Layer Perceptron (MLP), have shown promising results in handling complex datasets. These models are capable of learning intricate relationships between features and provide higher accuracy compared to traditional

ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

algorithms. However, they require proper preprocessing and parameter tuning to achieve optimal performance.

Apart from algorithm selection, researchers have also emphasized the importance of feature selection and data preprocessing. Studies show that including multiple parameters such as behavioral data, resource usage, and interaction levels can significantly improve prediction accuracy. However, many existing systems still rely heavily on limited features like marks and attendance, which do not fully capture a student's overall performance.

Another key observation from previous work is the lack of practical implementation in many research studies. While several models achieve good accuracy in theoretical analysis, they are not always integrated into user-friendly systems. This creates a gap between research and real-world application.

The proposed system addresses these gaps by combining multiple machine learning models with a web-based interface and incorporating skill-based attributes along with academic and behavioral data. This approach not only improves prediction accuracy but also enhances usability, making the system more practical for educational institutions.

In addition, recent studies have explored the use of educational data mining techniques to extract meaningful insights from large datasets generated in learning environments. These techniques focus on identifying hidden patterns, correlations, and trends that are not easily visible through traditional analysis. The integration of such techniques with machine learning models has significantly improved the ability to predict student outcomes. However, the effectiveness of these models largely depends on the quality of data and the selection of relevant features.

Moreover, there is a growing interest in developing systems that not only predict student performance but also provide actionable insights. Many existing approaches focus only on classification results without offering guidance for improvement. This highlights the need for systems that combine prediction with recommendation mechanisms. The proposed work contributes to this area by integrating prediction results with meaningful suggestions, thereby bridging the gap between analysis and practical application in educational environments.

This approach enhances the overall usefulness of the system by enabling educators to take timely and informed decisions. As a result, such systems can play a significant role in improving academic performance and supporting student success in modern education.

ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

3. Problem Statement

In traditional educational systems, student performance is primarily evaluated based on examination scores and attendance records. However, these factors alone do not provide a complete understanding of a student's overall performance. Important aspects such as classroom participation, resource utilization, behavioral engagement, and skill development are often ignored. As a result, it becomes difficult to accurately identify students who are at risk of poor academic performance.

Another major issue is the lack of early prediction mechanisms. In many cases, students are identified as underperforming only after their results decline significantly, leaving limited opportunity for improvement. This delayed identification affects both academic outcomes and student confidence. Additionally, manual analysis of student data is time-consuming and prone to human error, making it inefficient for large-scale educational environments.

Existing systems that attempt to predict student performance often rely on limited datasets and do not integrate multiple influencing factors. Many of these systems also lack user-friendly interfaces and practical implementation, which reduces their usability in real-world scenarios. Furthermore, they focus only on prediction without providing actionable insights or recommendations for improvement.

Therefore, there is a need for an efficient and intelligent system that can analyze multiple parameters, predict student performance accurately, and identify at-risk students at an early stage. The system should also provide meaningful insights and recommendations to support better decision-making. Addressing these challenges is essential for improving academic performance and enabling data-driven educational practices.

4. Proposed Methodology / Model

The proposed system is designed to predict student performance and identify at-risk students using machine learning techniques. The methodology follows a structured pipeline that transforms raw student data into meaningful predictions and insights. The system integrates data preprocessing, feature engineering, model training, evaluation, and real-time prediction within a unified framework.

The overall workflow begins with data collection, where student-related information is gathered. This data includes academic performance, behavioral engagement, attendance, and skill-based attributes. Unlike traditional systems that rely only on marks, the proposed approach incorporates multiple dimensions of student performance, making the prediction more comprehensive and

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

accurate.

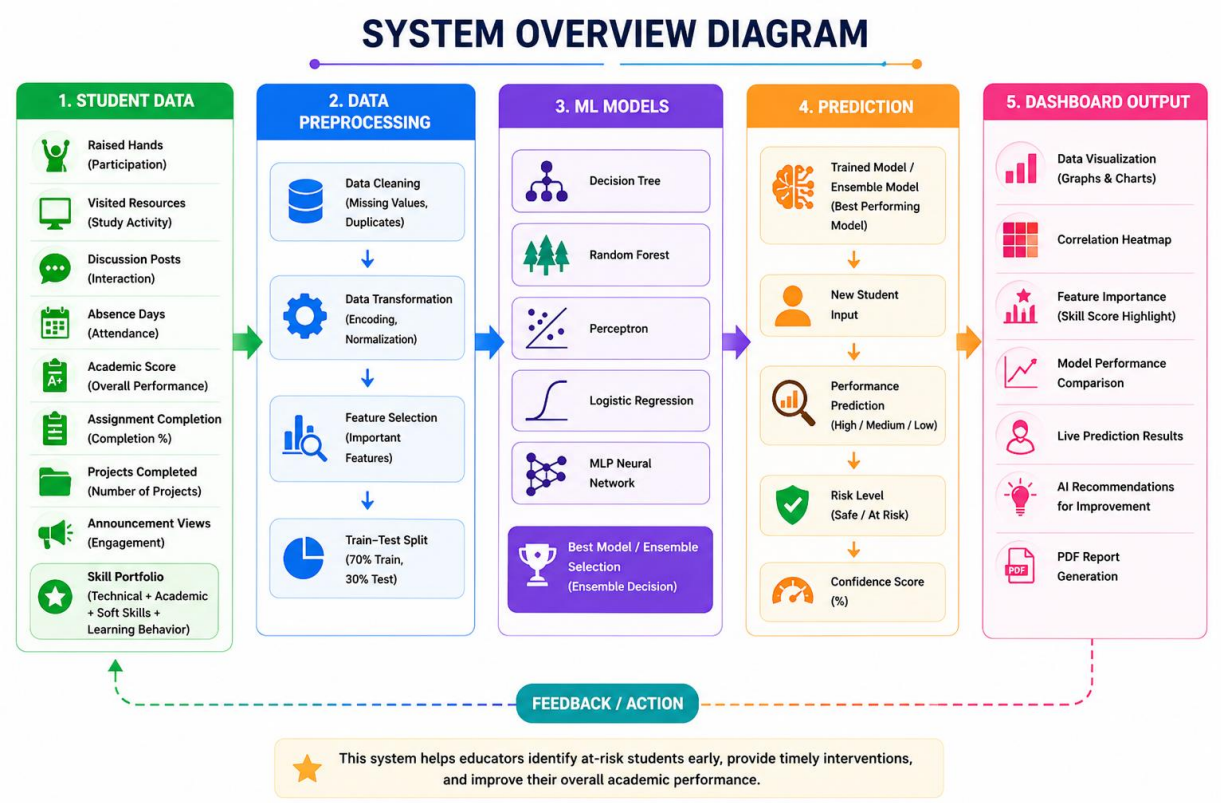


Figure 4.1: System Architecture / Methodology Flow Diagram

4.1 Data Preprocessing

Data preprocessing is performed to ensure data quality and consistency. Raw data may contain missing values, inconsistencies, or non-uniform formats, which can affect model performance.

The preprocessing steps include:

- Data cleaning to remove invalid or incomplete entries
- Encoding categorical variables into numerical form
- Normalization to scale features
- Feature selection to identify important attributes

ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

These steps improve model accuracy and efficiency.

4.2 Feature Engineering

Feature engineering enhances the predictive capability of the system.

- Skill-based attributes are converted into numerical values
- Behavioral features such as participation and interaction are combined
- Relevant features are selected based on their importance

This ensures that the model focuses on meaningful data.

4.3 Model Development

The system uses multiple classification algorithms to analyze student data:

- Decision Tree
- Random Forest
- Logistic Regression
- Perceptron
- MLP Neural Network

Using multiple models allows performance comparison and selection of the best-performing model.

4.4 Model Training and Evaluation

The dataset is divided into training (70%) and testing (30%) sets. Models are trained and evaluated using:

- Accuracy
- Precision
- Recall
- F1 Score

Cross-validation is applied to ensure reliability. The MLP Neural Network achieves the highest accuracy among all models.

ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

4.5 Prediction and System Integration

The trained model is used to generate predictions based on user input. The system outputs:

- Performance Level (High, Medium, Low)
- Risk Status (Safe or At Risk)
- Confidence Score

The system is integrated into a web-based application using Flask, allowing users to input data and receive real-time predictions along with visual insights.

5. Implementation

The implementation of the proposed system focuses on developing a web-based application that integrates machine learning models with an interactive user interface. The system is implemented using Python as the primary programming language, with Flask as the backend framework. Various libraries such as Pandas, NumPy, and Scikit-learn are used for data processing and model development, while Matplotlib and Seaborn are used for visualization.

The dataset is loaded from a CSV file and preprocessed to ensure data quality. Preprocessing includes handling missing values, encoding categorical variables, and normalizing feature values. The processed data is then divided into training and testing sets for model development.

Multiple machine learning models, including Decision Tree, Random Forest, Logistic Regression, Perceptron, and MLP Neural Network, are implemented using Scikit-learn. Each model is trained on the dataset and evaluated based on performance metrics. The best-performing model is selected for generating predictions.

The system is integrated into a Flask-based web application that allows users to input student data through a form. The backend processes the input data, applies the trained model, and generates prediction results in real time. The output includes performance level, risk status, and confidence score.

The frontend of the application is developed using HTML, CSS, and JavaScript to provide a user-friendly interface. The system also includes a dashboard that displays visual insights such as graphs and performance comparisons, making it easier to interpret results.

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

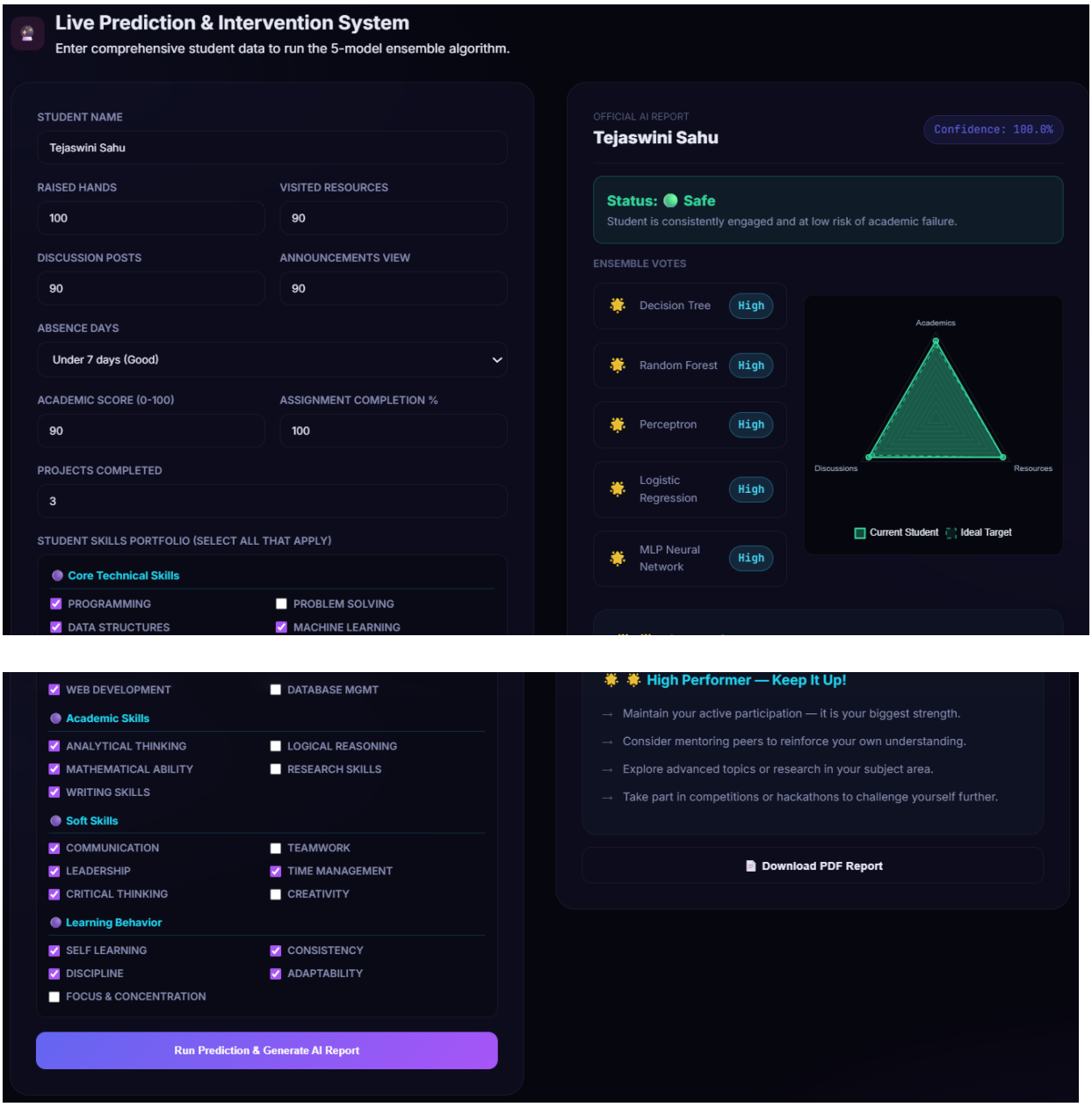


Figure 5.1: System Interface showing input form and prediction output

6. Results and Discussion

The results of the proposed system demonstrate the effectiveness of machine learning techniques in predicting student performance. Multiple classification models were implemented and evaluated using standard performance metrics to identify the most suitable model.

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

6.1 Model Performance Analysis

The performance of different machine learning models is compared based on accuracy. The results show that the MLP Neural Network achieves the highest accuracy among all models, followed by Random Forest and Logistic Regression. Decision Tree and Perceptron show comparatively lower performance.

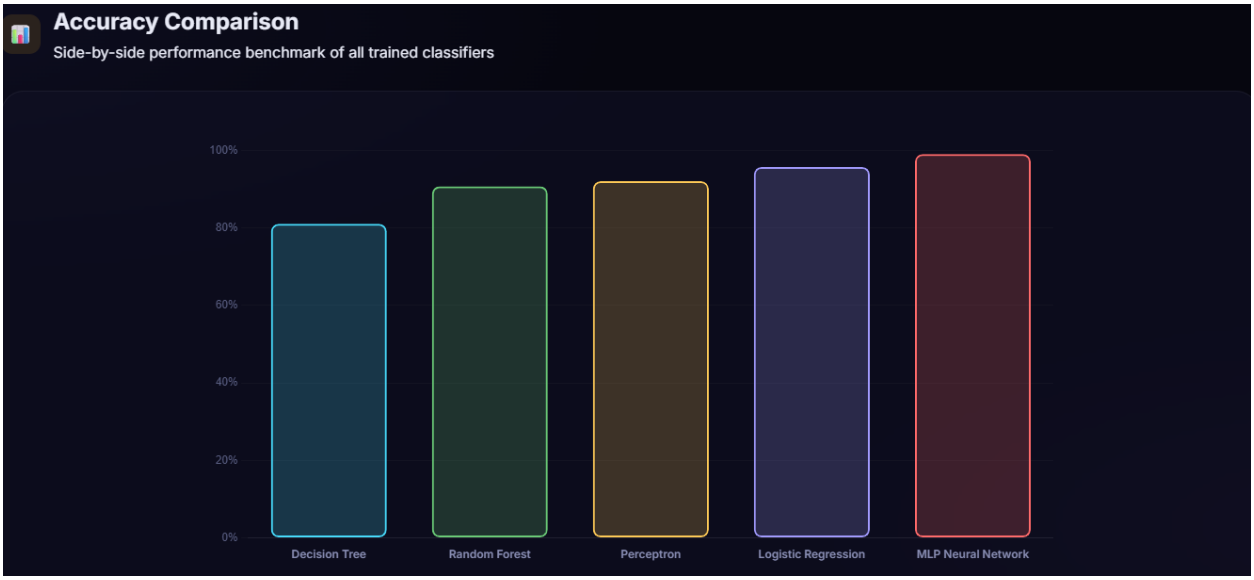
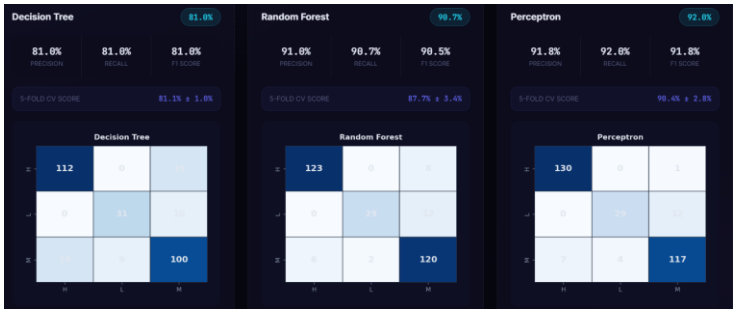


Figure 6.1.1: Model Performance Comparison

This indicates that advanced models such as neural networks are more effective in capturing complex relationships in student data.

6.2 Confusion Matrix Analysis

The confusion matrix is used to evaluate the classification performance of the model by comparing actual and predicted values. The results show that most predictions fall along the diagonal, indicating correct classification.



Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

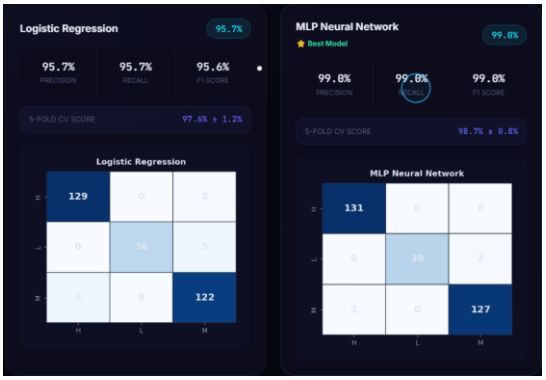


Figure 6.2.1: Confusion Matrix

This confirms that the model produces accurate and reliable predictions with minimal misclassification.

6.3 Cross-Validation Results

Cross-validation is applied to ensure that the model performs consistently across different subsets of data. The results show stable accuracy across multiple folds, indicating that the model is reliable and not overfitting.

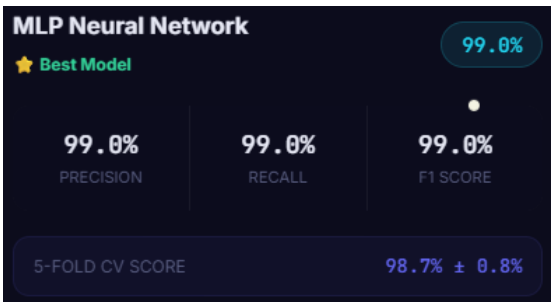


Figure 6.3.1: Cross-Validation Performance

6.4 Data Visualization

The system includes visual analysis tools such as correlation heatmaps and graphs. These visualizations help in understanding relationships between different features and identifying patterns in student data.

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

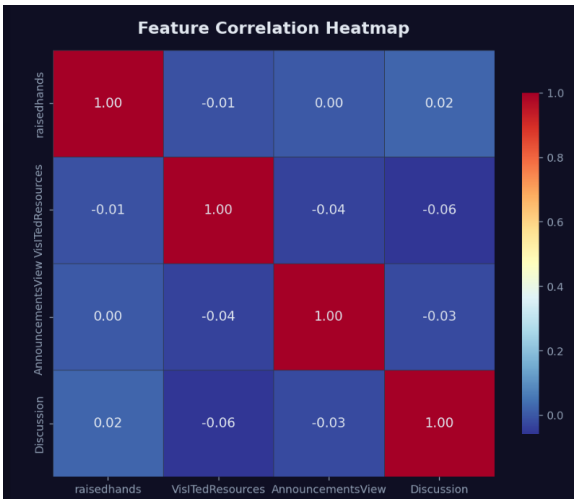


Figure 6: Data Visualization (Heatmap / Graphs)

6.5 Discussion

The results show that the proposed system effectively predicts student performance using multiple input parameters. The inclusion of academic, behavioral, and skill-based features improves prediction accuracy. The MLP Neural Network performs best due to its ability to model complex patterns.

The system also provides real-time predictions and visual insights, making it useful for practical applications. Overall, the results confirm that the system is accurate, efficient, and suitable for educational environments.

7. Testing and Validation

Testing and validation are performed to ensure that the proposed system works correctly, efficiently, and produces accurate predictions. This phase is essential to verify that the system meets its objectives and performs reliably under different conditions. Both functional testing and machine learning validation techniques are applied to evaluate the system.

7.1 Testing Approach

The system is tested at multiple levels to ensure proper functionality:

ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

• **Functional Testing:** Verifies that all modules such as data input, preprocessing, prediction, and output display operate correctly
• **Integration Testing:** Ensures that different components of the system work together without errors
• **User Testing:** Confirms that the system is user-friendly and provides meaningful results

These testing methods ensure that the system behaves as expected and delivers correct outputs for various inputs.

7.2 Model Validation

The machine learning models are validated using standard techniques to ensure reliability and generalization:

• Train-Test Split (70% training, 30% testing)
• Cross-validation to evaluate performance consistency
• Confusion matrix to analyze classification accuracy

Cross-validation plays an important role in checking whether the model performs consistently across different subsets of data.

7.3 Performance Metrics

The system performance is evaluated using widely accepted metrics:

• Accuracy
• Precision
• Recall
• F1 Score

These metrics provide a comprehensive evaluation of the model by measuring correctness, error rate, and balance between predictions.

7.4 Validation Results

The validation results indicate that the proposed system performs effectively:

• The model achieves high accuracy and reliable predictions
• Cross-validation results show stable performance across different folds

ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

• The system handles multiple inputs efficiently
• Response time is fast and suitable for real-time prediction

7.5 Discussion

The testing process confirms that the system is accurate, stable, and efficient. The use of cross-validation ensures that the model is not overfitting and can generalize well to new data. The combination of multiple testing methods and evaluation metrics strengthens the reliability of the system.

Overall, the validation results demonstrate that the system is suitable for practical use in educational environments and can effectively support student performance analysis and risk detection.

10. Conclusion

This research presents a machine learning-based system for predicting student performance and identifying at-risk students using multiple input parameters. Unlike traditional approaches that rely mainly on academic scores, the proposed system incorporates behavioral and skill-based attributes, providing a more comprehensive analysis of student performance.

Various machine learning models, including Decision Tree, Random Forest, Logistic Regression, Perceptron, and MLP Neural Network, were implemented and evaluated. The comparative analysis shows that the MLP Neural Network achieves the highest accuracy, demonstrating its effectiveness in handling complex and non-linear data patterns. The use of multiple models improves the reliability of the system and ensures better prediction performance.

The system is implemented as a web-based application, enabling users to input student data and obtain real-time predictions. The integration of visualization tools such as graphs and charts enhances the understanding of results, while the recommendation feature provides practical guidance for improving student performance.

The results and validation confirm that the system is accurate, efficient, and reliable. Cross-validation ensures model stability, and performance metrics demonstrate strong predictive capability. The system also shows fast response time and user-friendly operation, making it suitable for practical use in educational environments.

ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

Overall, the proposed approach provides an effective solution for student performance prediction by combining machine learning techniques with real-world applicability. It supports early identification of at-risk students and enables data-driven decision-making, contributing to improved academic outcomes and better educational management.

11. Future Scope

The proposed system can be further enhanced in several ways to improve its performance and applicability in real-world educational environments. One of the key improvements is the integration of real-time data from institutional databases, which would enable continuous monitoring of student performance and dynamic prediction updates. The system can also be extended by incorporating additional features such as psychological factors, learning behavior, and socio-economic background to provide a more comprehensive analysis. The use of advanced techniques such as deep learning and ensemble models can further improve prediction accuracy and handle more complex datasets. Another important area of improvement is the development of a mobile or cloud-based version of the system, which would allow access from multiple devices and support large-scale deployment. Enhancing the recommendation system to provide personalized learning strategies can also increase the practical usefulness of the system. Additionally, implementing advanced security mechanisms and role-based access control can ensure data privacy and safe usage of the system. With these improvements, the system has the potential to become a powerful tool for educational analytics and intelligent decision-making.

References

[1] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python*, O'Reilly Media, 2016.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[3] Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[4] McKinney, W., "Data Structures for Statistical Computing in Python," *Proceedings of the 9th Python in Science Conference*, 2010.

ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 35-50

[5] Hunter, J. D., "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[6] Flask Documentation,

[7] Pandas Documentation,

[8] NumPy Documentation,

[9] Seaborn Documentation,

[10] Wikipedia, "Machine Learning,"

[11] GeeksforGeeks, "Machine Learning Concepts,"

[12] Research articles on student performance prediction and educational data mining.