IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Comparative Analysis of Supervised Machine Learning Algorithms for Student Performance Assessment

Dr.K.Gayathri[1], M.Sc.,M.Phil.,Ph.D.,
Associate Professor,
Department of Computer Science,
Nirmala College for Women,
Red Fields, Coimbatore – 641118.

D.Abisha[2]
Department of Computer Science,
Nirmala College for Women, Red Fields,
Coimbatore – 641118.

Abstract : Accurate prediction of student academic achievement plays an important role in enhancing educational quality and enabling early academic support. Conventional assessment approaches often require significant manual effort and may fail to identify struggling students at the right time. This study develops a supervised machine learning framework to estimate final examination scores using selected academic and demographic attributes. The input features include Gender, Attendance percentage, Internal Test marks, and a derived Internal Assessment score. Three regression algorithms—Linear Regression, Decision Tree, and Random Forest—were developed and compared using performance measures such as RMSE, MAE, and R^2 score. Among the evaluated models, Random Forest produced the lowest prediction error, demonstrating better accuracy and stability. An interactive dashboard was also designed to support real-time score prediction, data visualization, and academic decision-making. The results highlight the effectiveness of machine learning techniques in monitoring student performance and enabling timely intervention.

Keywords : Student Performance Prediction, Academic Achievement, Machine Learning, Random Forest, Regression Models, Data Visualization, Early Academic Intervention, Predictive Analytics

Introduction

Educational institutions often struggle to recognize students who are likely to score poorly before final examinations. Early performance forecasting can support targeted academic assistance and improve student success rates. By utilizing structured academic records, machine learning techniques provide an effective way to uncover performance patterns and estimate final exam results. In this work, attributes such as Gender, Attendance percentage, Internal Test marks, and a derived Internal Assessment score are used to train supervised regression models.

Linear Regression, Decision Tree, and Random Forest algorithms are implemented and comparatively analyzed to identify the most reliable model. Additionally, a Streamlit-based interactive dashboard enables administrators to generate real-time predictions, monitor overall class trends, and support data-driven decisions. The developed framework combines preprocessing, feature construction, model

assessment, and visualization within a unified system, making it practical for real-world academic environments.

Objectives of the Study

1. Examine significant academic and demographic variables that influence student achievement.
2. Design and compare three supervised learning model sv Linear Regression, Decision Tree, and Random Forest.
3. Assess the effectiveness of each model using performance measures such as RMSE, MAE, and R^2 score.
4. Determine the most accurate algorithm for early identification of students at academic risk.
5. Create an interactive visualization dashboard to support performance monitoring and data-driven decision-making.

Review of Literature

Machine learning approaches have been extensively explored for forecasting student academic outcomes in the domain of educational data analytics. Research reported in IEEE Explore compared several predictive models for estimating student performance and observed that ensemble techniques generally achieve higher consistency and improved accuracy than single-model approaches. The study also stressed the significance of using appropriate evaluation measures such as RMSE and R^2 to validate model effectiveness.

Ahmed (2024), in work published through Wiley Online Library, examined supervised learning methods including Decision Tree and Random Forest for predicting academic results. The study demonstrated that careful preprocessing of data and appropriate feature selection play a crucial role in enhancing predictive performance. A publication in the International Journal of Environmental Sciences investigated structured student datasets and identified attendance levels and internal assessment marks as influential indicators of final examination scores. Further related studies identified via Google Scholar suggest that although many researchers emphasize classification-based outcome prediction, relatively fewer investigations integrate regression modeling with interactive visualization frameworks for real-time academic monitoring.

Methodology

Model Selection

1. Choose supervised learning models for prediction.
2. Use **Linear Regression, Random Forest, and Decision Tree**.
3. Compare which model works best for this dataset.
4. Keep models simple and easy to implement.

Model Training: Linear Regression, Random Forest, and Decision Tree.

Evaluation: Metrics include RMSE, MAE, R^2, and confusion matrix for performance categories (Excellent / Average / Poor).

Prediction: Dashboard predicts final exam marks for individual students, provides visualization, and allows filtered dataset download.

Dataset Description

The dataset used in this study consists of student academic, demographic, and behavioral records, collected from publicly available educational repositories. The data is structured and suitable for supervised machine learning models.

Student_ID – Unique identifier for each student.

1. Gender – Encoded as 0 (Female) and 1 (Male); included to examine any demographic influence on performance.
2. Attendance (%) – Represents the overall class attendance; a higher percentage is correlated with better academic performance.
3. Internal Test 1 (out of 40) – Marks obtained in the first internal test; reflects mid-term knowledge and preparation.
4. Internal Test 2 (out of 40) – Marks obtained in the second internal test; indicates progress and consistency.

5. Assignment Score (out of 10) –Marks obtained in assignments; demonstrates continuous learning and effort.

6. Internal Assessment (out of 100) –Combined score of Internal Test 1, Internal Test 2, and Assignment Score, scaled to 100; used as a key predictor for final marks.

7. Final Exam Marks (out of 100) –Target variable; represents the student's final performance.

Data Preprocessing

- Handling Missing Values: All incomplete or null entries were cleaned or removed to maintain dataset integrity.
- Feature Selection: Age, Daily Study Hours, and original Assignment Score were dropped to reduce noise and redundancy.
- Encoding: Categorical variables such as Gender were label-encoded.
- Feature Scaling: Numerical features were normalized using Standard Scalar to ensure uniformity across all models.

Algorithm Description

1. Linear Regression (LR)

Linear Regression is a supervised machine learning algorithm that models the relationship between a continuous target variable and one or more input features by fitting a linear equation. In this project, LR was applied to predict students' final xam marks using features such as Gender, Attendance percentage, Internal Test scores, and a combined Internal Assessment score. The model assumes a linear relationship between the input features and the predicted marks and provides a baseline for comparing more complex algorithms.

2. **Decision Tree (DT)**

Decision Tree is a supervised learning algorithm that predicts a continuous output by recursively splitting the dataset into subsets based on feature values. Each internal node represents a decision rule on a feature, and each leaf node represents a predicted score. In this study, the Decision Tree used academic and demographic attributes to learn patterns in student performance, capturing non-linear relationships and feature interactions that linear models cannot easily identify.

3. Random Forest (RF)

Random Forest is an ensemble learning technique that builds multiple Decision Trees using bootstrap samples of the dataset and combines their predictions to improve accuracy and reduce over fitting. For predicting student exam marks, RF leverages features such as Gender, Attendance, Internal Tests, and Internal Assessment scores to create a robust model that generalizes well across the dataset. The ensemble approach reduces variance and enhances the reliability of predictions compared to a single Decision Tree.

Model Evaluation

The effectiveness of the developed regression models was examined to ensure precise prediction of students' final examination scores. The models were evaluated using standard performance metrics suitable for continuous value prediction.

1. **Root Mean Square Error (RMSE):** RMSE was used to quantify the overall prediction deviation by computing the square root of the mean of squared differences between actual exam marks and predicted marks. Since it gives higher weight to larger errors, it helps in understanding how well the model handles significant prediction gaps. Lower RMSE values indicate better accuracy in estimating student performance.

2. **Mean Absolute Error (MAE):** MAE was calculated to determine the average magnitude of errors between predicted and actual final marks. Unlike RMSE, it treats all errors equally and provides a straightforward interpretation of the typical prediction difference. Smaller MAE values reflect improved model reliability.

3. **R^2 Score :**

The R^2 metric was applied to measure how effectively input features such as Attendance, Internal Test scores, Gender, and Internal Assessment explain the variation in final exam results. Values closer to 1 indicate that the model captures a higher proportion of performance variability.

Experimental result

Login page

The login page securely validates users to restrict access to the dashboard and prediction tools.

Actual vs Predicted (RF)

The graph shows a comparison between the actual and predicted student scores using the Random Forest model. It demonstrates the model's accuracy in predicting performance trends.

Performance Metrics

This table presents the performance metrics of different models, highlighting their accuracy in predicting student scores. It helps identify the most effective model for the system.

Confusion Matrix

The confusion matrix visualizes the correct and incorrect predictions of the model, showing how well it classifies student performance. It helps evaluate prediction accuracy

Conclusion

The study demonstrates that Random Forest is the most effective algorithm for predicting student academic performance, achieving the lowest RMSE and highest R^2. Linear Regression and Decision Tree models provide reasonable accuracy but are cless reliable. The Streamlit dashboard facilitates interactive visualization, early identification of at-risk students, and data-driven academic planning. Future enhancements could include deep learning models, larger datasets, and real-time cloud deployment for broader applicability.

References

1. Evaluation of Machine Learning Models in Student Academic Performance Prediction. IEEE Xplore.

https://ieeexplore.ieee.org/document/1096 3104

2. Student Performance Prediction Using Machine Learning Algorithms, Ahmed (2024), Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1155 /2024/4067721