INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 2, No 08, 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-760 PAGE NO: 015-019

A MACHINE LEARNING APPROACH TO STUDENT ACADEMIC PERFORMANCE ANALYSIS USING LINEAR REGRESSION

Radeesh Velayudhan
Department of Computer Science, Sri Krishna Adithya College of Arts & Science (Co-Education), Coimbatore, Tamil Nadu, India.

E. K. Girisan
Department of Computer Science, Sri Krishna Adithya College of Arts & Science (Co-Education), Coimbatore, Tamil Nadu, India.

Abstract:

Student academic performance analysis is an important application of machine learning in the education sector. Educational institutions collect academic data such as attendance percentage, internal assessment marks, assignment scores, and previous semester results. Analyzing these attributes can help identify performance patterns and estimate students' future academic outcomes. Traditional result analysis methods mainly rely on manual calculations and descriptive statistics, which may not provide sufficient predictive insights. This paper proposes a student performance analysis system using Linear Regression, a supervised machine learning algorithm. The system uses historical student data to predict final examination marks based on selected academic attributes. The methodology includes data collection, preprocessing, exploratory data analysis, model training, prediction, and evaluation. The model is evaluated using Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and the coefficient of determination. A sample student dataset is used to illustrate the proposed implementation. The system can assist teachers in identifying academic trends, estimating expected performance, and planning suitable learning interventions. The study highlights the usefulness of linear regression as an interpretable baseline model for educational data analysis.

Keywords: Data mining, linear regression, machine learning, student performance.

1. Introduction

Education is an essential factor in individual and social development. Student academic performance is generally evaluated using internal assessments, assignments, attendance, practical examinations, and final examinations. Educational institutions maintain records of these activities to evaluate learning outcomes and identify areas that require improvement. Traditional student result analysis involves calculating total marks, averages, pass percentages, and subject-wise performance. Although these

All Rights Reserved ©2026 International Journal of Computer Science (IJCS Journal)

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 2, No 08, 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-760 PAGE NO: 015-019

techniques are useful, they mainly describe existing performance. They may not provide a systematic method for estimating future examination marks from multiple academic attributes.

Machine learning offers techniques for identifying patterns in historical data and making predictions for new observations. Linear Regression is one of the simplest supervised machine learning algorithms. It models the relationship between one or more independent variables and a continuous dependent variable. In student result analysis, Linear Regression can estimate final examination marks using attendance percentage, internal marks, assignment scores, and previous semester marks. The predicted marks can help teachers identify students who may need additional academic support. The proposed system applies Linear Regression to student academic data. It aims to demonstrate how historical examination records can be converted into a predictive model that supports academic monitoring and educational decision-making.

2. Objectives

The main objectives of the proposed system are:

- To analyze student academic performance using machine learning.
- To identify relationships between academic attributes and final examination marks.
- To predict final examination marks using Linear Regression.
- To evaluate the model using suitable regression metrics.
- To support teachers in planning academic guidance and improvement activities.

3. Methodology

The proposed system follows a sequence of steps: data collection, preprocessing, exploratory data analysis, feature selection, model training, prediction, and evaluation.

A. Data Collection

The dataset may be collected from institutional examination records, student information systems, or anonymized academic spreadsheets. Each row represents one student, and each column represents an academic attribute. The following attributes are considered:

- **Student ID:** A unique identifier used to distinguish student records.
- **Attendance Percentage:** The percentage of classes attended by the student.
- **Internal Marks:** Marks obtained in internal assessments.
- **Assignment Score:** Marks obtained through assignments and coursework.
- **Previous Semester Marks:** Marks obtained in the previous semester.
- **Final Examination Marks:** The final examination score, which is the target variable.

Student ID is excluded from model training because it does not represent academic performance. Sensitive personal

All Rights Reserved ©2026 International Journal of Computer Science (IJCS Journal)

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 2, No 08, 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-760 PAGE NO: 015-019

information should also be excluded unless its use is necessary and appropriately authorized.

B. Data Preprocessing

Raw student data may contain missing values, duplicate records, invalid marks, or inconsistent formats. Data preprocessing improves the quality of the dataset before training the model.

The training data is used to learn the model coefficients, while the testing data is used to evaluate predictions on unseen observations. Any preprocessing parameters must be learned from the training data to avoid data leakage.

C. Feature Selection

In this study, four independent variables are selected: attendance percentage, internal assessment marks, assignment score, and previous semester marks. The dependent variable is final examination marks.

These features are selected because they represent different aspects of a student's academic history. Their usefulness should be verified using actual data. Feature selection must also reflect the intended prediction time: only information available before the final examination should be used as an input.

4. Linear Regression

Linear Regression is a supervised machine learning algorithm used to predict continuous numerical values. It estimates the relationship between input features and an output variable by fitting a linear equation to historical observations. For a single independent variable, the equation is:

$$\hat{y} = b_0 + b_1 x \quad (1)$$

Where ŷ is the predicted output, x is the independent variable, b0 is the intercept, and b1 is the regression coefficient.

For student result analysis, multiple academic attributes are used. Therefore, Multiple Linear Regression is applied.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \quad (2)$$

Here, x1 represents attendance percentage, x2 represents internal marks, x3 represents assignment score, and x4 represents previous semester marks. The coefficients indicate the estimated contribution of each feature while the other included features are held constant.

The algorithm estimates the coefficients by minimizing the sum of squared differences between actual and predicted marks:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3)$$

Where yi is the actual final examination mark, ŷi is the predicted mark, and n is the number of observations.

The fitted model can then estimate final examination marks for students whose input attributes are known.

All Rights Reserved ©2026 International Journal of Computer Science (IJCS Journal)

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 2, No 08, 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-760 PAGE NO: 015-019

5. Results and Discussion

The performance of the regression model is assessed using four evaluation metrics.

A. Evaluation Metrics

Mean Absolute Error (MAE): Measures the average absolute difference between actual and predicted marks. A smaller value indicates that predictions are closer to the observed scores on average.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (4)$$

Mean Squared Error (MSE): Measures the average squared prediction error. Larger errors receive greater weight because the differences are squared.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (5)$$

Root Mean Squared Error (RMSE): Calculates the square root of MSE. It is expressed in the same units as the examination marks.

$$RMSE = \sqrt{MSE} \quad (6)$$

Coefficient of Determination (R^2): Measures how well the model explains variation in the target values relative to a baseline that predicts the test-set mean. An R^2 value closer to 1 generally indicates a better fit. However, R^2 may be negative on test data if the model performs worse than that baseline.

B. Interpretation of Results

After the model is executed, the evaluation metrics can be used to assess the quality of the predictions. For example, a lower MAE means that the model's predictions are closer to actual examination marks on average. RMSE helps identify the effect of larger prediction errors, while R^2 provides information about the model's explanatory performance on the test set.

A comparison between predicted and actual marks can reveal whether the model consistently overestimates or underestimates student performance. A scatter plot of actual marks against predicted marks can also help identify systematic errors.

The coefficients may indicate associations between the selected academic features and final marks. However, they do not establish causation. For instance, an association between attendance and examination marks does not prove that attendance alone causes improved performance.

6. Conclusion and Future Scope

This paper presents a machine learning-based approach for student academic performance analysis using Multiple Linear Regression. The proposed system uses attendance percentage, internal assessment marks, assignment scores, and previous semester marks to predict final examination marks. The methodology includes data

All Rights Reserved ©2026 International Journal of Computer Science (IJCS Journal)

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 2, No 08, 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-760 PAGE NO: 015-019

collection, preprocessing, feature selection, model training, and evaluation using MAE, MSE, RMSE, and R^2. Python libraries provide a practical way to implement the model and analyze its predictive performance.

Linear Regression offers an interpretable baseline for educational prediction tasks. Its practical value must be established through experiments on real student data and comparison with suitable alternative models. Future work may include developing a student performance dashboard, analyzing results across multiple semesters, applying cross-validation, and comparing Linear Regression with Random Forest Regression and other nonlinear methods.

In conclusion, Linear Regression can support systematic student result analysis and academic planning when it is implemented using reliable data, rigorous evaluation, and responsible human oversight.

References

1. Romero, C. and Ventura, S. 2010. "Educational Data Mining: A Review of the State of the Art." IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 40(6), pp. 601–618.
2. Cortez, P. and Silva, A. 2008. "Using Data Mining to Predict Secondary School Student Performance." In Proceedings of the 5th Annual Future Business Technology Conference (FUBUTEC 2008), pp. 5–12.
3. Baker, R. S. and Inventado, P. S. 2014. "Educational Data Mining and Learning Analytics." In Larusson, J. A. and White, B. (eds.), Learning Analytics: From Research to Practice. Springer, pp. 61–75.
4. James, G., Witten, D., Hastie, T., and Tibshirani, R. 2021. An Introduction to Statistical Learning: With Applications in R, 2nd ed. Springer.

All Rights Reserved ©2026 International Journal of Computer Science (IJCS Journal)