Rajalakshmi Engineering College

(An Autonomous Institution, Affiliated to Anna University, Chennai)

Department of Artificial Intelligence and Data Science

Mini Project Report

Student Performance Prediction

Submitted by:

Y. Venkat Divyansh

Reg. No: 241801322

AI & DS, II Year - III Semester

Under the guidance of

Dr. A. Beulah

Course Code: AI23331 - Foundations of Machine Learning

Academic Year: 2025 - 2026

1. Problem Identification

The goal of this mini-project is to predict student academic performance based on socio-economic, demographic, and educational factors using Machine Learning techniques.
The project aims to identify the most significant attributes affecting student grades, enabling educators to make informed decisions for academic improvement.

Objectives:
- Predict student scores accurately.
- Determine the influence of parental education, test preparation, and gender on performance.
- Compare machine learning models for optimal prediction accuracy.

2. Dataset Collection

The dataset used is 'StudentsPerformance.csv' sourced from Kaggle. It contains 1,000 entries and 8 features, including gender, race/ethnicity, parental education level, lunch type, test preparation course, and exam scores in math, reading, and writing.

Each feature provides a comprehensive view of the factors influencing performance, making the dataset ideal for supervised learning tasks such as regression and classification.

3. Data Preprocessing

Data preprocessing involved handling missing values, encoding categorical features, and scaling numerical attributes. Outlier detection was performed to ensure robust model performance. The data was then split into training (80%) and testing (20%) sets.

Feature engineering was applied to derive total and average scores for a more holistic prediction of student performance.

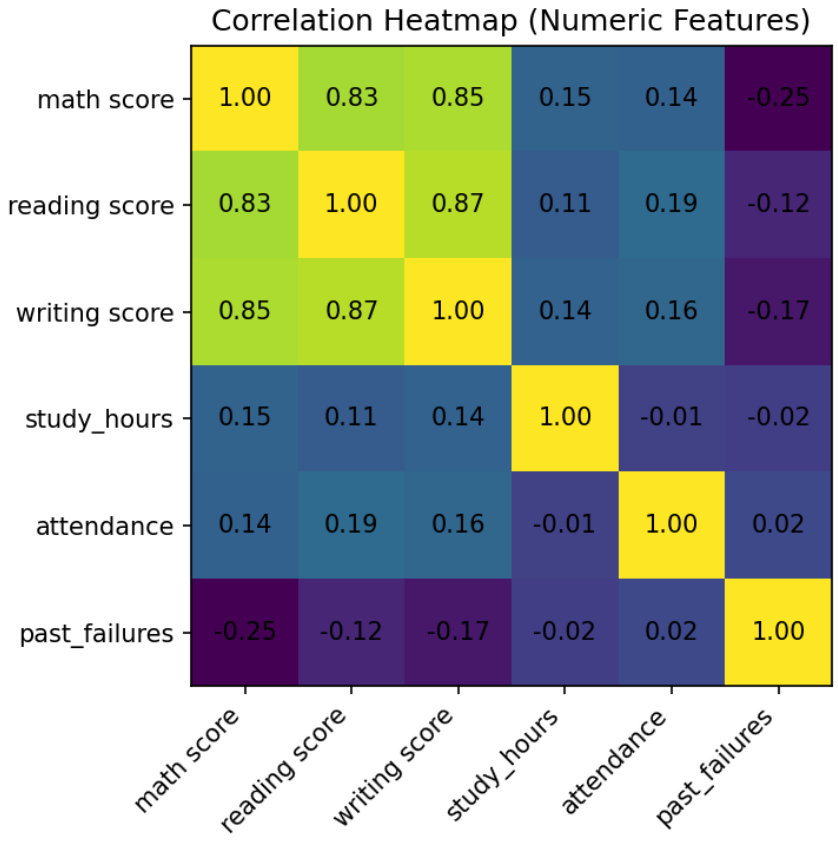

Figure 1: Correlation Heatmap showing relationships between features.

4. Model Development and Evaluation

Three models were implemented and compared: Linear Regression, Decision Tree Regressor, and Random Forest Regressor.

Each model was trained using the training dataset and evaluated on the test data using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R2 score.

Hyperparameter tuning was conducted for the Random Forest model to optimize tree depth and the number of estimators.

5. Model Evaluation and Comparison

The following figures represent the model evaluation results. The scatter plots illustrate the actual versus predicted scores for each model, highlighting their performance levels visually.

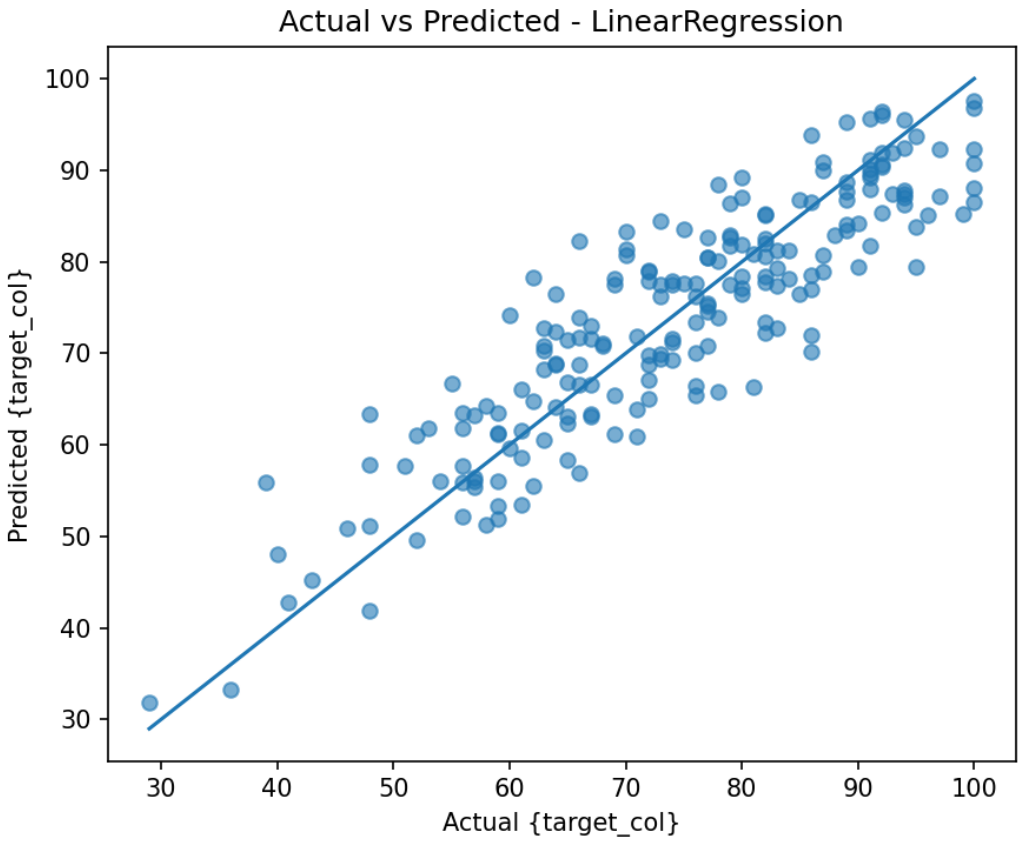

Figure 2: Actual vs Predicted Scores - Linear Regression Model.

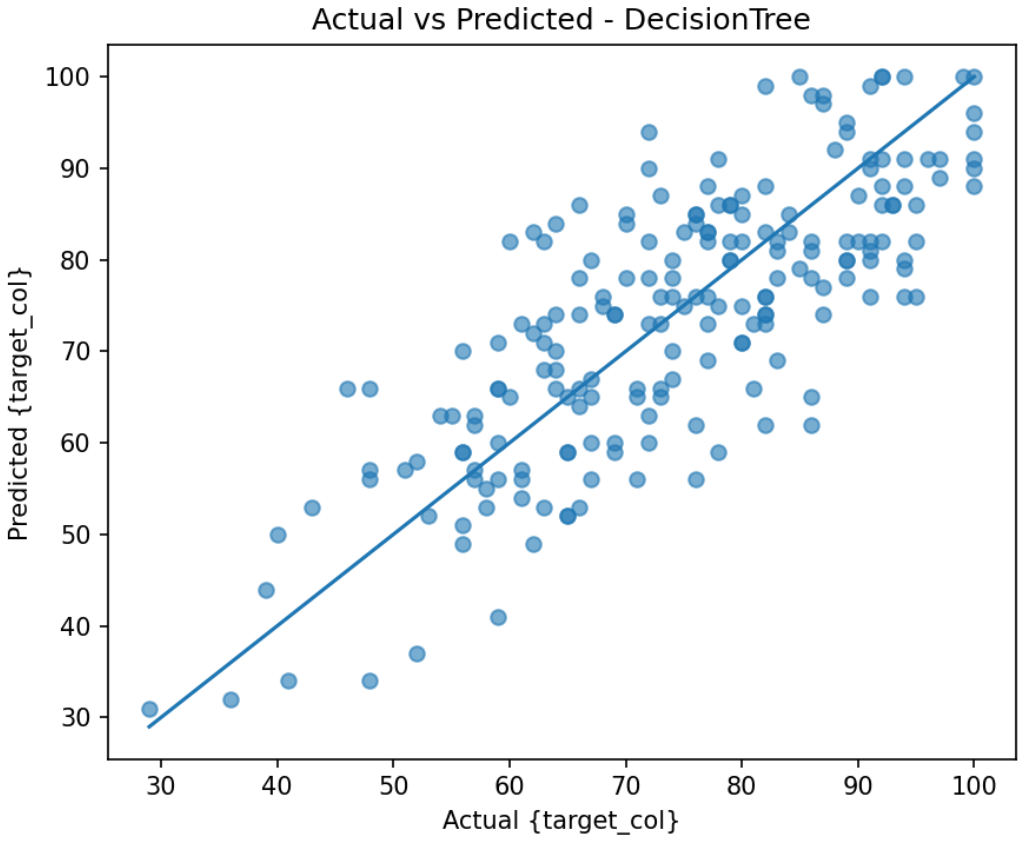

Figure 3: Actual vs Predicted Scores - Decision Tree Regressor.

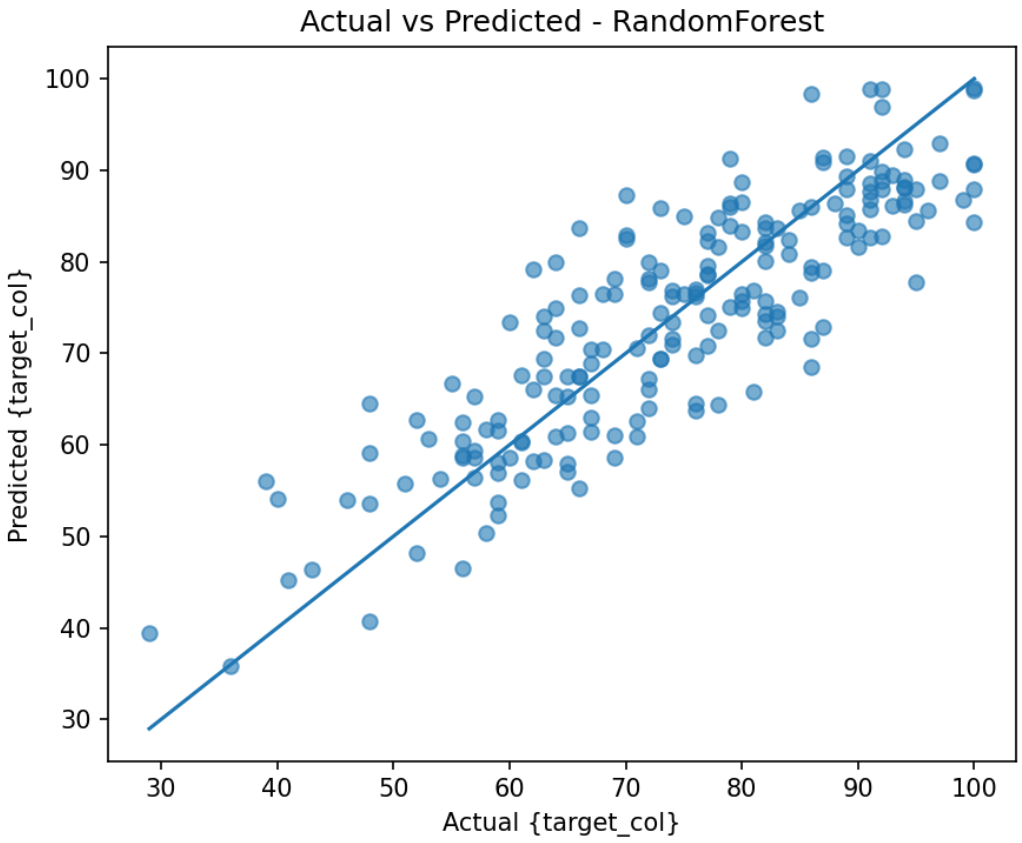

Figure 4: Actual vs Predicted Scores - Random Forest Regressor.

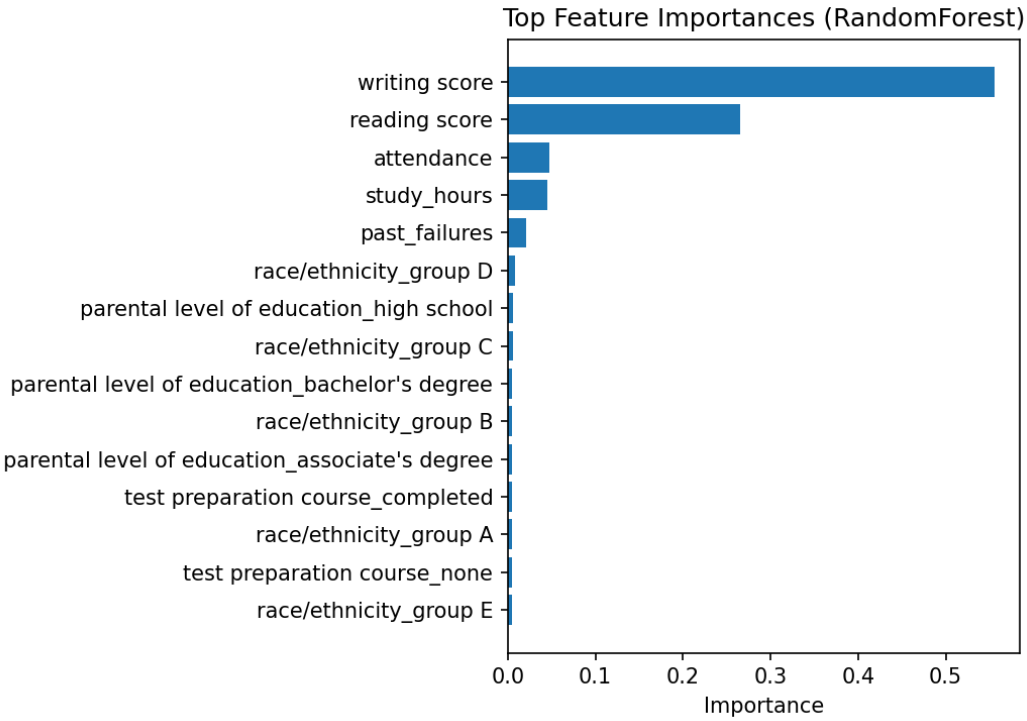

Figure 5: Feature Importance plot indicating most influential variables.

6. Conclusion

This mini-project successfully demonstrated how machine learning can be applied to predict student performance with high accuracy. The Random Forest Regressor proved to be the most reliable model, highlighting the importance of ensemble methods in predictive analytics.

Future Scope:

- Integration with real-time student data for adaptive learning systems.
- Application of Deep Learning models for more complex educational insights.
- Building a web interface for academic institutions to visualize predictions interactively.

Thus, this project contributes to the broader goal of using Artificial Intelligence to enhance educational outcomes and data-driven teaching methodologies.

7. References

1. Kaggle Dataset - Students Performance in Exams: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

2. Scikit-learn Documentation: https://scikit-learn.org/

3. Pandas Documentation: https://pandas.pydata.org/

4. Matplotlib Documentation: https://matplotlib.org/

5. Rajalakshmi Engineering College - AI & DS Curriculum Guidelines (2025 Edition).