Students Marks Prediction Model

(Data Science Project Report)

Instructor:

Dr. Muhammad Ismail Mangrio

Submitted by:

Sarfraz Ali Katpar	**023-23-0045**
Hassan Ali	**023-23-0046**
Faizanullah	**023-23-0047**

ABSTRACT

Academic performance prediction has become a crucial application of Data Science in the education sector. This project presents a **Student Marks Prediction Model** based on **supervised machine learning**, specifically **Linear Regression**, to estimate students' academic outcomes using measurable academic and behavioral attributes.

The proposed system predicts student marks by analyzing key parameters such as **study hours, class attendance, assignment completion, sleep duration, and previous academic performance**. These features were selected due to their established relevance in educational research and their strong influence on student learning outcomes.

The system is implemented as a **web-based application using the Flask framework**, enabling users to input student-related data and receive instant predictions. The machine learning model is trained on a real-world dataset and evaluated using standard regression performance metrics including **Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R^2 score**.

Experimental results demonstrate that the model achieves satisfactory predictive performance and provides interpretable results, making it suitable for academic analysis. The system can serve as a **decision-support tool for educators and institutions**, helping identify students who may require early academic intervention and enabling data-driven educational planning.

INTRODUCTION

With the rapid growth of data-driven decision-making, **predicting student academic performance** has gained significant attention in educational institutions. Traditional evaluation methods often detect academic issues only after poor performance has already occurred, limiting opportunities for timely intervention.

Machine learning techniques offer the ability to analyze historical student data and **predict future academic outcomes proactively**. Such predictive models help identify critical factors that influence student performance and enable educators to design personalized learning strategies.

This project focuses on the development of a **Student Marks Prediction System** using **supervised machine learning techniques**. The core objective is to build a predictive model capable of estimating student marks based on previously observed academic and behavioral data. A **Linear Regression algorithm** is employed due to its simplicity, interpretability, and suitability for modeling linear relationships between dependent and independent variables.

Additionally, the project integrates the trained model into a **Flask-based web application**, ensuring usability for non-technical users and demonstrating real-world deployment of a machine learning solution.

LITERATURE REVIEW

Numerous studies have explored the application of **machine learning algorithms** in predicting student academic performance. Commonly used techniques include **Linear Regression, Decision Trees, Random Forests, Support Vector Machines (SVMs), and Artificial Neural Networks**.

Linear Regression has been widely adopted in educational prediction tasks due to its **transparent mathematical formulation and ease of interpretation**. Several studies confirm that features such as **attendance rate, study hours, and previous academic performance** exhibit strong correlations with student outcomes.

More complex models, such as Random Forests and Neural Networks, have also shown improved accuracy in some cases; however, they often sacrifice interpretability, which is a critical requirement in academic environments. As a result, Linear Regression remains a preferred baseline model for educational analytics.

Furthermore, recent research highlights the importance of **web-based predictive systems** to ensure accessibility for educators and students. Frameworks such as **Flask and Django** are frequently used to deploy machine learning models due to their lightweight nature and flexibility.

Based on the reviewed literature, **Linear Regression combined with a web-based interface** was selected as an appropriate and effective approach for this project.

PROPOSED METHODOLOGY

The proposed system follows a **systematic machine learning pipeline**, ensuring data quality, model reliability, and deployment readiness.

1. Data Collection

A structured dataset containing **academic and non-academic student attributes** is used. The dataset includes variables such as:

- Study hours
- Attendance percentage
- Assignment completion
- Sleep hours
- Previous academic performance
- Final marks (target variable)

2. Data Preprocessing

Data preprocessing is a critical step to improve model performance and reliability. The following techniques were applied:

- **Missing Value Handling:** Missing values were treated using **median imputation**, which is robust against outliers.
- **Duplicate Removal:** Duplicate records were identified and removed to prevent bias.
- **Invalid Value Handling:** Logically impossible values (e.g., negative study hours) were eliminated.
- **Data Type Conversion:** All features were converted into numerical format to ensure compatibility with machine learning algorithms.

3. Exploratory Data Analysis (EDA)

EDA was conducted to understand data distribution and relationships:

- **Distribution Analysis:** Histograms were used to examine feature distributions.
- **Correlation Analysis:** Pearson correlation coefficients were calculated to identify relationships between variables.
- **Visualization:** Scatter plots and heatmaps were used to visually inspect linear trends and feature interactions.

4. Feature Selection and Engineering

Features demonstrating strong correlation with the target variable were retained. Redundant or weakly contributing features were excluded to reduce noise and improve generalization.

5. Model Training

A **Linear Regression model** was trained using the processed dataset. The dataset was split into **training and testing sets** to evaluate generalization performance.

6. Model Evaluation

The trained model was evaluated using standard regression metrics:

- **MAE (Mean Absolute Error)** – measures average absolute prediction error
- **MSE (Mean Squared Error)** – penalizes larger errors
- **RMSE (Root Mean Squared Error)** – interpretable error in original units
- **R^2 Score** – measures variance explained by the model

7. Model Deployment

The trained model was serialized using **Joblib** and deployed through a **Flask-based web application**, enabling real-time predictions through a user-friendly interface.

EXPERIMENTAL SETUP

Programming Languages

- Python
- JavaScript
- HTML
- CSS

Libraries and Frameworks

- NumPy – numerical computations
- Pandas – data manipulation and analysis
- Scikit-learn – machine learning algorithms
- Matplotlib & Seaborn – data visualization
- Flask – web application framework
- Joblib – model serialization

Development Tools

- VS Code / Jupyter Notebook
- Web Browser (Chrome / Firefox)

FINDINGS

The Linear Regression model performed satisfactorily in terms of predicting data for the test dataset. It can be seen that the error margin and the correlation coefficient in the results were satisfactory.

Key Observations

- **Study hours and attendance** showed a strong positive correlation with final marks.
- **Previous academic performance** had a significant influence on predictions.
- **Sleep hours and assignment completion** exhibited a moderate impact.
- The deployed web application successfully generated **real-time predictions** based on user inputs.

These findings align with existing educational research and validate the selected feature set.

DISCUSSION

The results confirm that **machine learning techniques can effectively predict student academic performance** when relevant features are available. Linear Regression proved to be a suitable choice due to its interpretability and simplicity, making it ideal for academic environments where understanding the influence of features is essential.

However, the model assumes linear relationships between variables, limiting its ability to capture complex patterns. Additionally, the **quality and size of the dataset** directly affect prediction accuracy. Despite these limitations, the system provides meaningful insights and serves as a strong foundation for educational analytics.

CONCLUSION

This project successfully demonstrates the design, implementation, and deployment of a **Student Marks Prediction Model** using machine learning and web technologies. The integration of **Linear Regression with a Flask-based interface** results in a practical, user-friendly system for academic performance prediction.

The system can assist educators in identifying students who may require additional support and promote **data-driven academic decision-making**.

Future Enhancements

- Incorporating advanced models such as **Random Forests or Gradient Boosting**
- Expanding the dataset for improved generalization
- Adding feature normalization and cross-validation
- Integrating dashboards for performance visualization
- Deploying the application on cloud platforms

GROUP MEMBERS CONTRIBUTION

This project was collaboratively developed by all group members, with each member contributing to specific phases of the Data Science lifecycle based on their technical expertise. The responsibilities were clearly distributed to ensure efficiency, accountability, and high-quality outcomes.

Faizanullah – Dataset Gathering, Data Cleaning, and Preprocessing

Faizanullah was primarily responsible for the **data acquisition and preparation phase**, which forms the foundation of any machine learning project. His contributions included collecting and organizing the student-related dataset containing both academic and behavioral attributes.

He performed comprehensive **data cleaning and preprocessing**, which involved:

- Identifying and handling missing values using appropriate imputation techniques (median-based imputation)
- Removing duplicate and inconsistent records to maintain data integrity
- Detecting and eliminating invalid or logically impossible values
- Converting categorical or inconsistent data formats into suitable numerical representations
- Ensuring the dataset was structured and ready for exploratory analysis and model training

These preprocessing steps significantly improved data quality and ensured reliable input for downstream machine learning tasks.

Sarfraz Ali Katpar – Exploratory Data Analysis, Model Building, Training, and Testing

Sarfraz Ali Katpar handled the **analytical and modeling phase** of the project. His primary responsibility was to explore the dataset and extract meaningful insights through **Exploratory Data Analysis (EDA)**.

Key contributions included:

- Performing statistical analysis to understand data distributions and trends
- Conducting correlation analysis to identify influential features affecting student marks
- Creating visualizations such as histograms, scatter plots, and heatmaps to interpret relationships among variables

- Selecting relevant features based on analytical findings
- Designing and implementing the **Linear Regression model**
- Splitting the dataset into training and testing sets to evaluate generalization
- Training the model and assessing performance using standard regression metrics including MAE, MSE, RMSE, and R^2 score

His work ensured that the model was both **technically sound and interpretable**, aligning with academic evaluation requirements.

Hassan Ali – Model Deployment, Graphical User Interface, and Flask Integration

Hassan Ali was responsible for the **deployment and application development phase**, transforming the trained machine learning model into a usable system.

His contributions included:

- Serializing the trained machine learning model using **Joblib** for efficient reuse
- Developing a **Flask-based backend** to handle user input, model inference, and result generation
- Designing and implementing a user-friendly **Graphical User Interface (GUI)** using HTML, CSS, and JavaScript
- Integrating the frontend with the Flask API to enable seamless communication between the user interface and the prediction model
- Testing the deployed application to ensure accurate predictions and smooth user interaction

This deployment work enabled **real-time student marks prediction**, making the project practical and accessible to non-technical users.

Video Link

https://drive.google.com/file/d/1L5DQa2Qno_pv4rzbNr31JI7Z72PvhDzN/view?usp=sharing