

Report on

**Predicting Student Exam Score through Regression**

**Submitted By:**  Sahil Keswani

**Introduction:**

Academic performance is influenced by multiple factors such as study habits, attendance, prior academic success, and family support. Predicting exam scores using these variables can help in early identification of students at risk and guide data-driven academic interventions.

This project focuses on building predictive models to estimate students’ exam scores and analysing which factors have the most significant impact.

**Problem Statement:**

**Aim:**

To develop and evaluate predictive models that estimate students' exam scores based on key academic and behavioural factors, with the objective of identifying the most influential variables that contribute to student performance and enabling data-driven educational interventions.

**Target Variable :** Exam\_Score

**Predictor Variables :** [ Hours\_Studied, Previous\_Scores, Attendance, Tutoring\_Sessions, Parental\_Involvement ]

**Dataset Description:**

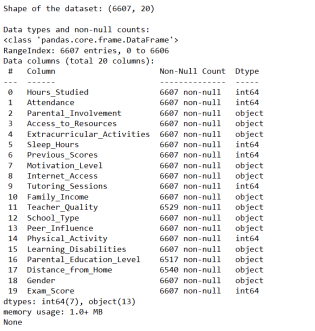
This Dataset is about Student Performance factors. The dataset contains academic, behavioral, and socio-economic information about students, aimed at analyzing factors that influence their exam performance.

**Source**: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors

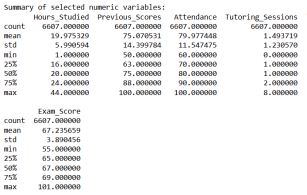
The Dataset consists of around 6607 records and 20 Variables. The nature of each variable is shown in the image below.

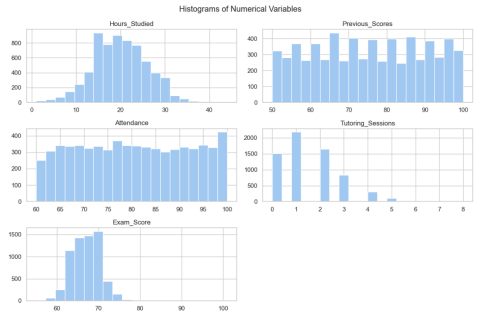
There are also some missing values in the dataset but our Selected variables did not have any missing values. From the selected predictor variables all are **numerical** except for **“Parental\_Involvement”.**

The **“Parental\_Involvement”** is categorical variable and the unique values in it are [‘Low’, ‘Medium’, ‘High’ ]



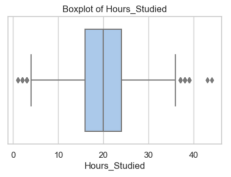
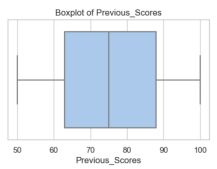
**Performing EDA (Exploratory Data Analysis):** The Summary and Histograms of all the numeric variables of the dataset.

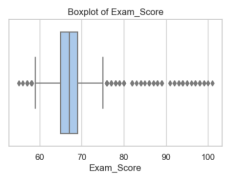
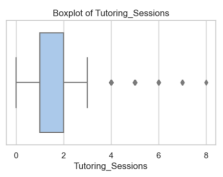
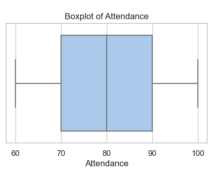




The Histogram gave us a visual representation to understand the distribution of each variable where we found that the graph of “Hours\_Studied” is normally distributed and the Graph of “Exam\_Score” is very Concentrated.

**BOX PLOTS:**

****

****

The Boxplot is used to identify Outliers in our Selected variables of the dataset and the spread of our data. It helped us visualize central tendency and dispersion across the variables.

**Data Preprocessing:**

**Missing Values:**

● There are not many missing values but if the missing values are there, they are handled by using mean imputation.

● Mean imputation is used because the mean is suitable for numerical data and avoids loss of records.



**Encoding:**

● “Parental Involvement ” was encoded using label encoding.

● Model requires numerical inputs. Label encoding helps preserve order if any.



**Model Building:**

After completing data preprocessing, the next step was to build predictive models to estimate students' Exam Scores based on relevant academic and behavioural factors. Two regression models were developed and compared: **Linear Regression** and **Random Forest Regression**.

Linear Regression was chosen as a baseline model due to its simplicity and interpretability. Exploratory Data Analysis suggested that the relationship between the predictors and the target variable (Exam Score) was generally linear, justifying the use of this model.

To capture potential non-linear relationships and interactions between variables, a **Random Forest Regressor** was also implemented. Random Forest is an ensemble learning method that builds multiple decision trees and averages their outputs to improve prediction accuracy and reduce overfitting.

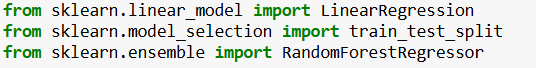
Both models were trained and evaluated on the dataset, allowing for a performance comparison to determine which approach provided better generalisation on unseen data.

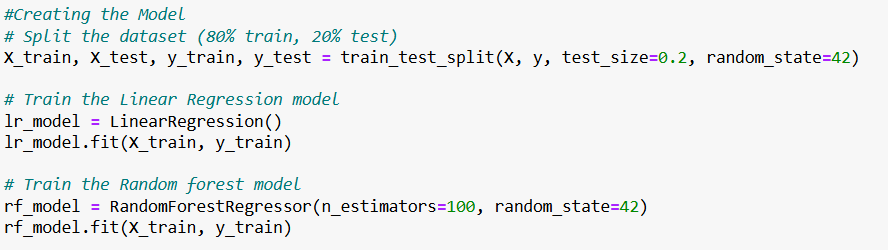
**Splitting the Dataset into Training and Testing Sets:**

The Dataset was split into:

● Training Set (80%)

● Testing Set (20%)





**Model Training:**

Two regression models were trained using scikit-learn: **Linear Regression** and **Random Forest Regressor**.

* For Linear Regression, the **LinearRegression()** class from **sklearn.linear\_model** was used. The model fits a line by minimizing the residual sum of squares between actual and predicted Exam Scores.
* For Random Forest, the **RandomForestRegressor()** from **sklearn.ensemble** was employed. It builds multiple decision trees on random subsets of the data and averages their predictions to improve accuracy and reduce overfitting.

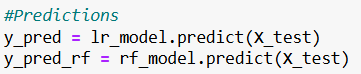
Each model was trained on the pre-processed dataset, and the trained models were then used to generate predictions on the test set.

**Model Interpretation:**

The intercept and coefficients give us insight into the relative importance of each feature and how each feature influences the target.

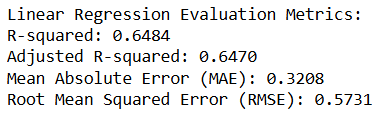
Linear Regression: After training, the intercept and coefficients were extracted to interpret the linear relationships. Each coefficient represents the expected change in Exam Score for a unit change in the corresponding predictor, keeping other variables constant. The intercept indicates the predicted Exam Score when all predictors are zero.

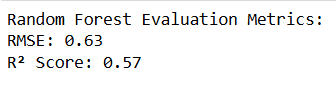
Random Forest Regression: Since Random Forest is a non-linear, ensemble-based model, it does not provide coefficients. Instead, feature importance scores were used to determine how much each variable contributes to the final prediction. These scores help identify the most influential features in a model that captures complex relationships.



**Model Evaluation:**

To assess the performance of the predictive models, multiple evaluation metrics were used:





**R-squared (R² Score)**

● Measures the proportion of variance in the target variable **(Exam\_Score)** explained by the predictors.

● A higher R² indicates a better model fit.

**Mean Absolute Error (MAE)**

● It is a metric that measures the average difference between predicted and actual values. It is used to evaluate the performance of the model.

● Lower MAE indicates more accurate predictions.

**Root Mean Squared Error (RMSE)**

● Measures the square root of the average squared differences between predicted and actual values.

● It indicates how well the model's predictions align with the real data.

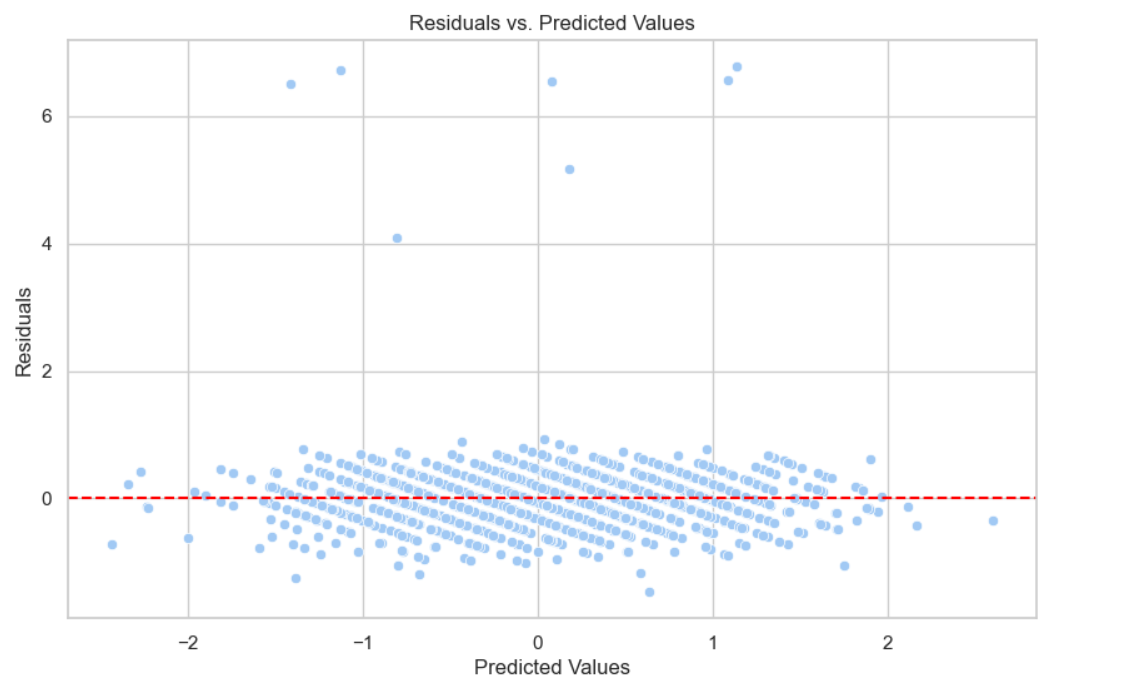
Both **Linear Regression** and **Random Forest Regression** were evaluated on the test dataset using these metrics.

**Residual Analysis:**

A **residual plot** and **distribution of residuals** were also used to:

● Check for **linearity** and **homoscedasticity** (equal spread of errors), normality of residuals, and absence of major outliers.

● Identify any skew or violations of model assumptions.



This is the Graph of **Residuals** vs **Predicted Values** This plot is used to evaluate **linearity** and **homoscedasticity.**

**No Clear Pattern**

● The residuals appear **randomly scattered** around the horizontal line at 0 (the red dashed line), with no strong systematic shape (like a curve or trend).

● This suggests that the **relationship between the predictors and the target is linear**, satisfying the linearity assumption.

**Most Residuals are close to 0**

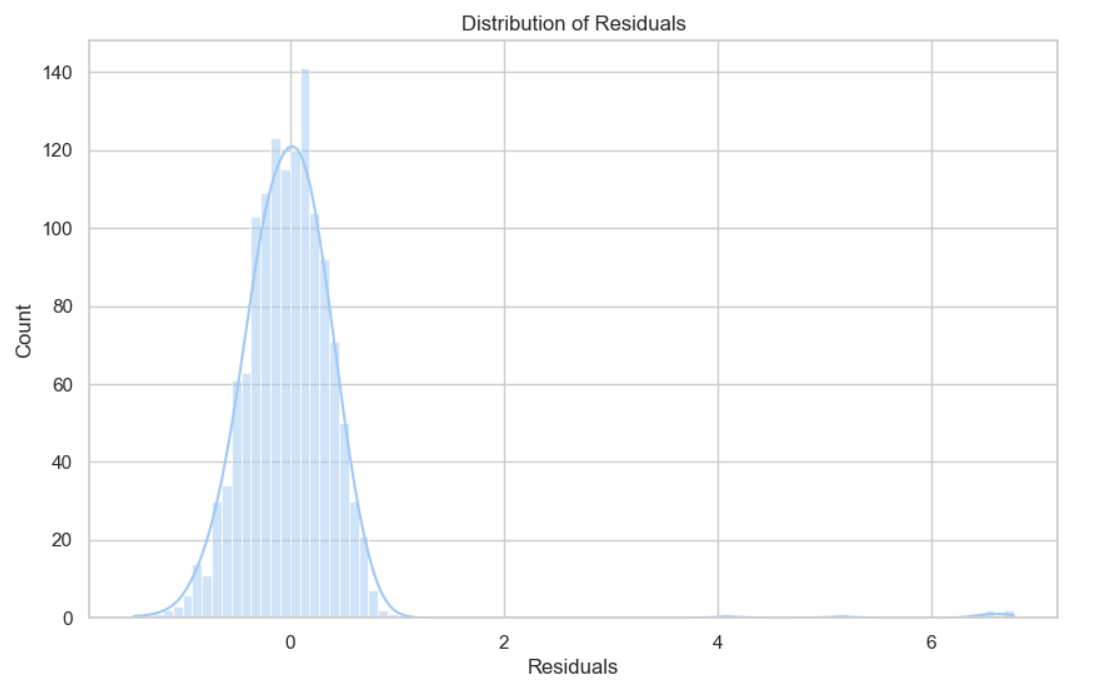
● A large cluster of points is near the red line, meaning that the **majority of predictions are close to the actual values**, indicating decent model performance.

**A Few Outliers:**

● Some points lie far from the red line, especially above residual value 6.

● These are **potential outliers** or high-error predictions and may influence the model. Investigating them could help improve the model further.

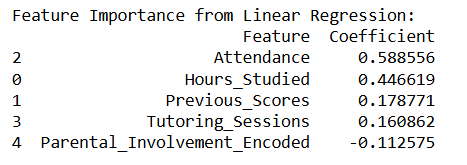
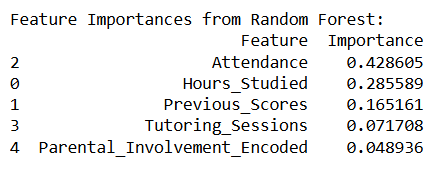
**Distribution of Residuals**

****

This plot helps to assess the **normality of residuals**, which is another key assumption in linear regression. The peak is around 0, meaning **most prediction errors are small and centered around zero**, which confirms the model is generally accurate.

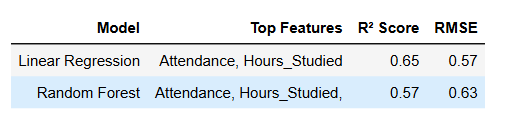
The comparison revealed that **Linear Regression performed slightly better**, suggesting that a linear relationship was sufficient for this dataset. However, the Random Forest model also showed competitive results, capturing potential non-linear patterns and interactions.

**Metrics Prediction:**



We Found that **“Attendance” & Hours Studied** were the most important and the most influential factor of all.

**Conclusion:**

****

This project aimed to predict students' exam scores using various academic and behavioural factors. After thorough exploratory data analysis and data preprocessing, both Linear Regression and Random Forest Regression models were developed and evaluated.

The analysis revealed that **Attendance Rate** and **Hours Studied** were among the most influential predictors of exam performance. Linear Regression performed slightly better in terms of R-squared and RMSE, indicating a predominantly linear relationship between the predictors and the target variable.

Key observations include:

● **Study Hours and Attendance Rate** showed a strong positive correlation with Exam Scores.

● **Linear Regression** provided interpretable coefficients that helped understand the individual impact of each factor.

● **Random Forest Regression** offered insights through feature importance scores and accounted for potential non-linear interactions.

● The residual plots confirmed that model assumptions for Linear Regression were reasonably satisfied.

A few **Limitations** which were observed such as:

**1.** **Limited Feature Set**:

●The dataset, while rich in academic and behavioural variables, may not fully capture socio-economic, psychological, or environmental factors that also influence exam performance.

**2. Assumption of Linearity:**

● Linear Regression assumes a linear relationship between predictors and the target, which may oversimplify real-world dynamics.

**3. Model Generalisability:**

● The models were trained on a specific dataset. Their effectiveness on different student populations or education systems may vary.

**Future Enhancement could include:**

* Expanding the dataset with more diverse student backgrounds and additional features like teacher effectiveness or peer influence.
* Exploring other machine learning models like Gradient Boosting or Neural Networks for potentially better accuracy.
* Deploying the model into an interactive dashboard or app to support real-time academic monitoring.