## Key Terminology in the Project

### 1. ****Features (Independent Variables)****

These are the input variables used to train the model. In this project, features include:

* **Gender**: Male or Female
* **Ethnic Group**: Categorical group (e.g., Group A, B, C...)
* **Parental Level of Education**: Highest education achieved by parents
* **Lunch**: Standard or Free/Reduced
* **Test Preparation Course**: Completed or None
* **Other optional demographic fields**: Can be customized

### 2. ****Target Variable (Dependent Variable)****

The outcome we are trying to predict:

**Overall-score**: An engineered metric calculated as the average of:

* MathScore
* ReadingScore
* WritingScore

This helps convert a multi-output regression problem into a single-output regression task.

### 3. ****Machine Learning Model****

You are using **Random Forest Regressor**, a robust ensemble learning method that:

1. Builds multiple decision trees
2. Aggregates their results for better accuracy and stability

### 4. ****Preprocessing Steps****

* **Imputation**: Handling missing data using SimpleImputer (median for numerical, most frequent for categorical)
* **Scaling**: Standardizing numerical features using StandardScaler
* **Encoding**: Converting categorical variables using OneHotEncoder

### 5. ****Evaluation Metrics****

To assess how well the model performs:

* **MSE (Mean Squared Error)**: Average squared difference between actual and predicted values
* **RMSE (Root MSE)**: Square root of MSE, easier to interpret
* **R² Score (Coefficient of Determination)**: Indicates the proportion of variance explained by the model (1 = perfect fit)

### 6. ****Model Persistence****

* **Joblib** is used to save the trained model to disk so it can be reused without retraining.

### 7. ****Streamlit Interface****

* **Sidebar**: Navigation across modules (Data Overview, Analysis, Model Training, Prediction)
* **Main Panel**: Dynamic forms, plots, prediction results, metrics