Machine Learning Model Development

1. Model Goal

The goal of the model is to **predict students' performance** based on multiple academic and socio-demographic factors.
It predicts **two target variables simultaneously**:

Overall Performance

Academic Index (Academic Average)

To achieve this, a **Multi-Output Regression** approach was adopted.

---

2. Data Preprocessing

Before model training:

The dataset was loaded from DATA.csv.

Unnecessary columns (student_id, student_name) were dropped.

The following categorical features were label-encoded:

TransportMeans

ParentEduc

LunchType

TestPrep

ParentMaritalStatus

PracticeSport

IsFirstChild

NrSiblings

A **LabelEncoder** was applied for each feature to convert text categories into numeric values suitable for training.
The mappings were stored in a file named **label_mappings.json** for use during prediction.

---

3. Feature and Target Selection

Features (X): All independent variables except the target columns.

Targets (y):

OverallPerformance

AcademicIndex

---

4. Model Architecture

The system uses a **MultiOutputRegressor** wrapping a **RandomForestRegressor**.
This allows the model to predict multiple continuous target variables at once.

Parameters:

n_estimators = 150

max_depth = 12

random_state = 42

This configuration balances **accuracy and efficiency**, preventing overfitting.

---

5. Training Process

The dataset was split into:

Training set: 80%

Testing set: 20%

The model was trained on the training data and evaluated on unseen test data.

---

6. Model Evaluation

After training, the model achieved the following performance metrics:

Metric	Overall Performance	Academic Index
R^2 Score	0.971	1.000
Mean Absolute Error (MAE)	1.551	0.166

The **R^2** values close to 1 indicate the model fits the data extremely well.

The **low MAE** values demonstrate that the predictions are very accurate.

---

7. Feature Importance

The most influential features (based on average feature importances across all estimators) include:

Rank	Feature	Impact
1	Study Time	High
2	Parent Education	High
3	Absences	High
4	Test Preparation	Moderate
5	Practice Sport	Moderate

The model analysis showed that **students who spend more time studying and whose parents have higher education levels tend to perform better academically**.

---

8. Saving and Deployment

After successful training:

The model was saved as **student_multi_model.pkl**

The label mappings were saved as **label_mappings.json**
This ensures that the exact encoding scheme used during training can be reused during deployment and prediction.

---

9. Example Prediction

After training, the model produced the following sample output (using test data):

Example Prediction:

Predicted OverallPerformance: 69.02

Predicted AcademicAverage: 81.15

These values align with real-world performance levels, confirming that the model generalizes well.

10. Integration with Streamlit Dashboard

The trained model is integrated into a **Streamlit web application**:

Users input student details (via sliders and dropdown menus).

The app reads **label_mappings.json** to correctly encode inputs.

Predictions for both target variables are displayed instantly.

Results are visualized in the **AI Predictions tab** and compared with real data through charts.