📅 Day 9: Project Report Preparation

Project Title: Student Performance Prediction using Machine Learning

---

1️⃣ Problem Statement

Student performance prediction is an important task in the education domain.
The objective of this project is to build a machine learning model that predicts whether a student will **Pass or Fail** based on demographic and academic factors such as gender, parental education, lunch type, test preparation course, and exam scores.

This prediction can help educational institutions identify students who may need additional support.

2️⃣ Dataset Description

- **Dataset Name:** Students Performance in Exams
- **Source:** Kaggle
- **Total Records:** 1000 students
- **Total Features:** 8 original features

Dataset Features:

- Gender
- Race/Ethnicity
- Parental level of education
- Lunch type
- Test preparation course

- Math score
- Reading score
- Writing score

A new feature **average_score** was created using subject scores, and a target variable **result** was defined:

- `1` → Pass (average score ≥ 40)
- `0` → Fail (average score < 40)

3️⃣ Exploratory Data Analysis (EDA)

EDA was performed to understand data distribution and relationships between variables.

Key Findings:

- Average scores follow a near normal distribution.
- Students who completed the test preparation course performed better.
- Students with standard lunch showed higher performance.
- Reading and writing scores were highly correlated.
- No significant class imbalance was found in the target variable.

Visualizations such as histograms, boxplots, and correlation heatmaps were used for analysis.

4️⃣ Data Preprocessing

The following preprocessing steps were applied:

- Categorical variables were converted into numerical form using **Label Encoding**.

- Numerical features were standardized using **StandardScaler**.
- The dataset was split into training and testing sets using an **80:20 ratio**.

This ensured that the data was suitable for machine learning models.

5️⃣ Model Building

Two machine learning models were trained:

Models Used:

1. Logistic Regression (Baseline Model)
2. Random Forest Classifier

After initial evaluation, Random Forest performed better.
Hyperparameter tuning was applied using **GridSearchCV** to improve model performance.

6️⃣ Model Evaluation & Results

Models were evaluated using:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

Results Summary:

- Random Forest achieved higher accuracy compared to Logistic Regression.

- Tuned Random Forest showed improved performance after hyperparameter optimization.
- The final optimized Random Forest model was selected.

7️⃣ Final Model

The optimized Random Forest model was saved using **Pickle** and **Joblib** for future use.

```
student_performance_model.pkl
student_performance_model.joblib
```

The saved model was successfully reloaded and tested.

8️⃣ Conclusion

This project successfully demonstrated the application of machine learning techniques to predict student performance.
The Random Forest model provided the best results after tuning.
Such predictive models can help educators identify students at risk and take early corrective actions.

9️⃣ Key Learnings

- Data preprocessing plays a critical role in model performance.
- Exploratory Data Analysis helps understand feature importance.
- Ensemble models like Random Forest handle complex relationships well.
- Hyperparameter tuning significantly improves accuracy.
- Proper documentation is essential for real-world projects.