# 🎓 Student Performance - Pass/Fail Prediction Report

## 1. Introduction

This project builds a **machine learning model** to predict whether a student will **Pass** or **Fail**, based on personal, academic, and socio-economic features.

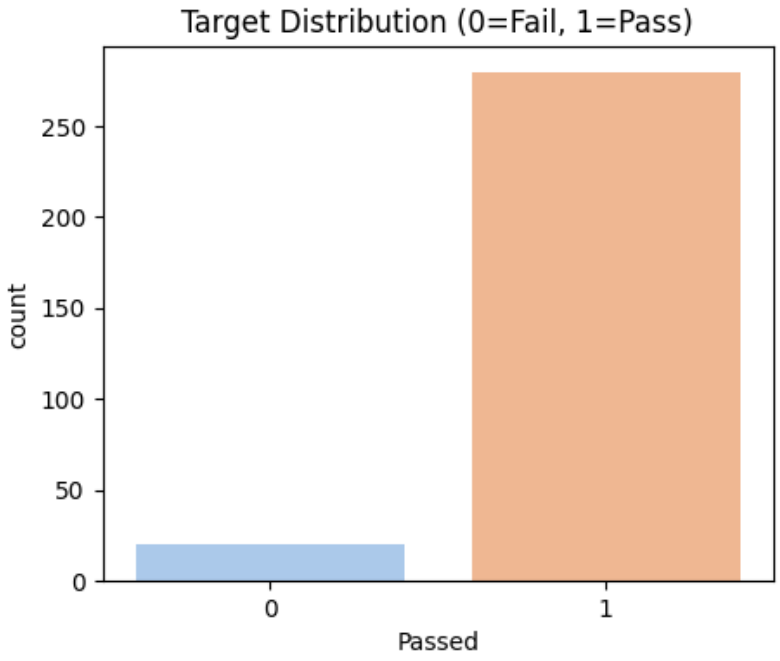
- **Dataset:** `student_performance_dataset.csv`
- **Students:** 300
- **Features:** Gender, Age, Study Hours, Attendance, Parental Education, Internet Access, Extracurricular Activities
- **Target:** Pass (1) / Fail (0)

**Goal:** Help educators identify at-risk students early and provide interventions.

## 2. Dataset Overview

- Total Students: **300**
- Features Used:
  - **Numerical:** Age, Study Hours, Attendance
  - **Categorical:** Gender, Parental Education, Internet Access, Extracurricular
- Target Distribution:
  - **Pass (1): ~280 students**
  - **Fail (0): ~20 students**

📊 *Insert Chart 1: Target Distribution (Bar Plot of Pass vs Fail)*




## 3. Preprocessing Pipeline

1. **Feature Selection:** Removed `Test_Score` (to prevent leakage).
2. **Numeric Features:** Imputed missing values (median), scaled with StandardScaler.
3. **Categorical Features:** Imputed missing values (mode), encoded using OneHotEncoder.
4. **Train-Test Split:** 80% Train, 20% Test (with stratification).

📊 *Insert Diagram 2: Preprocessing Pipeline Flowchart*

## 4. Models Used

### Logistic Regression

- Pros: Simple, interpretable, useful for explaining feature importance.
- Cons: Assumes linearity, sensitive to outliers.

### Random Forest

- Pros: Handles non-linear data, more robust, higher accuracy.
- Cons: Less interpretable, requires more computation.

## 5. Evaluation Metrics

We used multiple metrics to evaluate the models:

- **Accuracy:** % of correct predictions.
- **Precision:** How many predicted “Pass” are actually Pass.
- **Recall (Sensitivity):** How many actual “Pass” were correctly identified.
- **F1-score:** Balance of Precision & Recall.
- **ROC AUC:** How well the model separates Pass vs Fail.

---

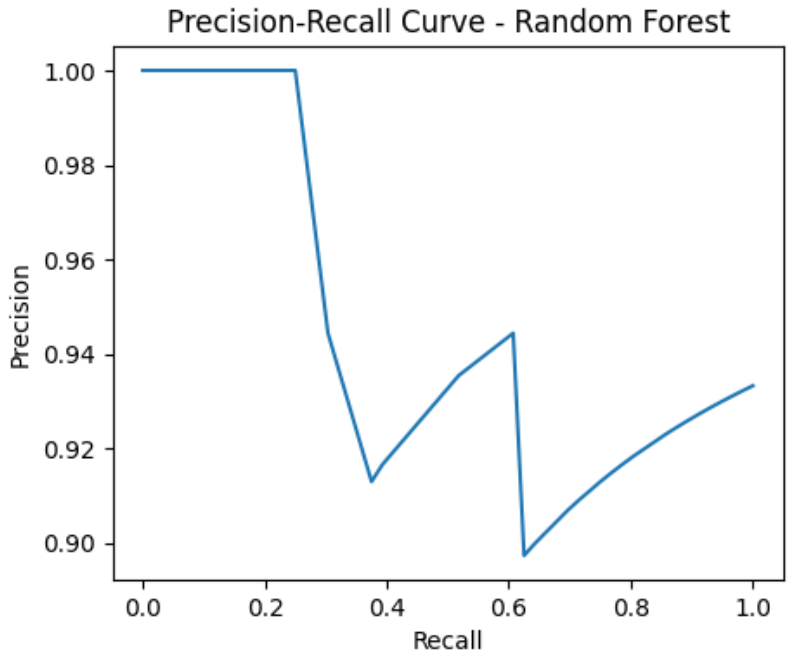
## 6. Results (Test Data)

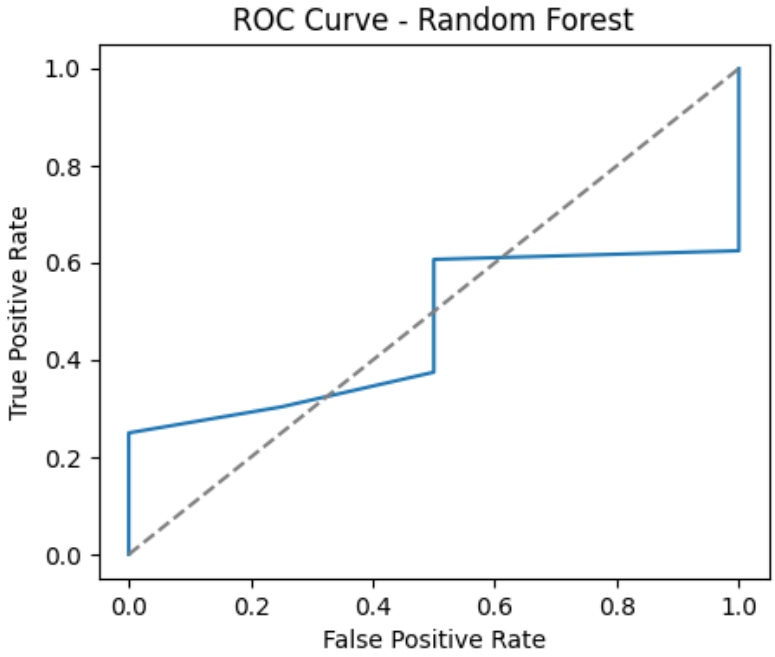
| Model | Accuracy | Precision | Recall | F1-score | ROC AUC |
|---|---|---|---|---|---|
| Logistic Regression | 93.3% | 93.3% | 100% | 96.6% | ~0.60 |
| Random Forest | 93.3% | 93.3% | 100% | 96.6% | ~0.42 |

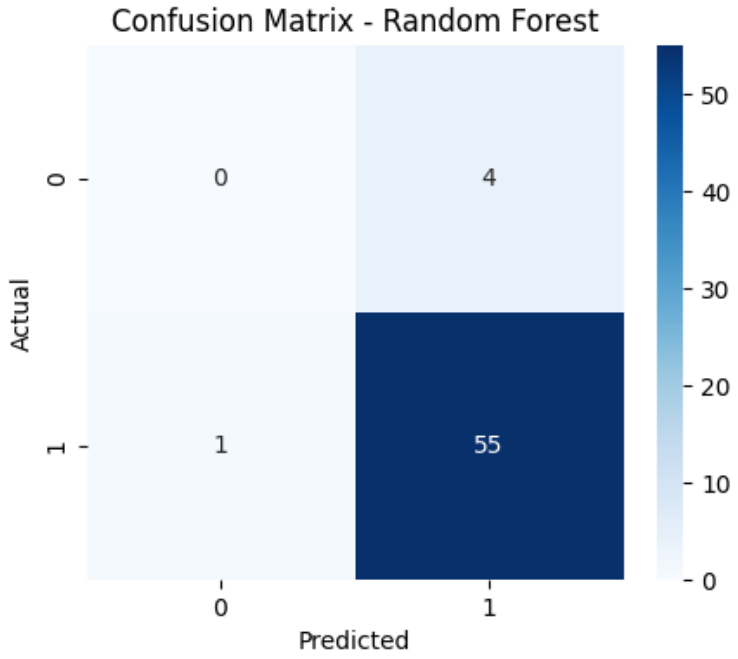
📊 *Insert Chart 3: Confusion Matrix Heatmap*

📊 *Insert Chart 4: ROC Curve (Random Forest)*

📊 *Insert Chart 5: Precision-Recall Curve*






## 7. Key Insights

- **Study Hours** and **Attendance** strongly impact passing probability.
- **Parental Education** positively influences student performance.

- **Extracurricular Activities** show a small but positive effect.
- **Internet Access** helps moderately, but not the strongest predictor.

---

## 8. Conclusion & Interview Notes

- Both Logistic Regression and Random Forest perform well.
- Logistic Regression → highlight **simplicity & interpretability**.
- Random Forest → highlight **accuracy & robustness**.
- Preprocessing ensures **clean and consistent input** for models.
- In interviews, follow this flow:

  **Problem → Data → Features → Preprocessing → Models → Metrics → Insights → Real-world Value**.

📌 **Business Value:** This model can help schools identify students likely to fail and provide tutoring or support before exams.

---

## 9. Next Steps (if asked in Interview)

- Handle imbalance (e.g., SMOTE, class weights).
- Add more features (e.g., family income, study environment).
- Deploy as a **Flask/Django app** for real use.
- Use **SHAP/Feature Importance** for deeper interpretability.

---

## 10. Appendix (Visuals to Add)

- **Chart 1:** Target distribution (Pass vs Fail).
- **Chart 2:** Preprocessing pipeline (flowchart).
- **Chart 3:** Confusion Matrix heatmap.
- **Chart 4:** ROC Curve.
- **Chart 5:** Precision-Recall Curve.