Student Pass/Fail Classification – Model Evaluation Report

1. Introduction

This project focuses on predicting whether a student will pass or fail based on academic-related features such as study hours, attendance, previous performance, and assignment scores. The problem is formulated as a **binary classification task**, where:

- 1 = Pass
- 0 = Fail

Three supervised learning models were implemented and compared:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Random Forest

The models were evaluated using Accuracy, Confusion Matrix, and F1-Score.

---

2. Data Preprocessing

The dataset includes the following features:

- Hours (study hours)
- Attendance (percentage)
- Previous_Score
- Assignments
- Result (Pass/Fail)
- To prepare the data:
    1. A Pass/Fail target variable was created using a score threshold.
    2. The dataset was split into training (80%) and testing (20%) sets.
    3. Feature scaling (StandardScaler) was applied for Logistic Regression and KNN, as these models are sensitive to feature magnitude.
    4. Random Forest was trained on unscaled data, as tree-based models are scale-invariant.

---

3. Model Evaluation

3.1 Logistic Regression

Logistic Regression is a linear classification model that estimates probabilities using the logistic function. It is simple, efficient, and highly interpretable.

- Strength: Easy to interpret coefficients.
- Weakness: May struggle with complex nonlinear relationships.

Performance was strong and stable, with balanced precision and recall.

3.2 K-Nearest Neighbors (KNN)

KNN is a distance-based model that classifies data points based on the majority class of nearest neighbors.

- Strength: Simple concept and effective for smaller datasets.
- Weakness: Computationally expensive for large datasets and sensitive to scaling.

KNN performed well but may vary depending on the chosen value of K. It captures local patterns in the data.

- ### 3.3 Random Forest

Random Forest is an ensemble model that builds multiple decision trees and combines their predictions.

- Strength: High accuracy and ability to handle nonlinear relationships.
- Weakness: Less interpretable compared to linear models.

Random Forest generally achieved the highest accuracy due to its ensemble nature and ability to reduce overfitting.

---

4. Comparison of Models

Model	Accuracy	Interpretability	Complexity
Logistic Regression	High	Very High	Low
KNN	Moderate-High	Low	Medium
Random Forest	Very High	Low	High

Key Observations:

- Logistic Regression provides a clear understanding of how each feature impacts the prediction.
- KNN depends on the local structure of data and requires careful tuning.
- Random Forest achieves strong predictive performance but behaves like a “black box.”

- # 5. Model Complexity vs Interpretability

  An important objective of this assignment was to understand how increasing model complexity affects interpretability.

  - As complexity increases, model accuracy often improves.
  - However, interpretability decreases.
  - Simpler models (Logistic Regression) allow direct explanation of feature influence.
  - Complex models (Random Forest) provide better predictive power but make decision

reasoning difficult.

In real-world educational systems, if transparency and fairness are important, Logistic Regression may be preferred. If maximizing prediction accuracy is the goal, Random Forest becomes more suitable.

---

6. Conclusion

This project demonstrates how different classification models perform on a student performance dataset.

All three models successfully predicted pass/fail outcomes, but their strengths differ:

- Logistic Regression is best for interpretability.
- KNN is useful for local pattern recognition.
- Random Forest provides the highest predictive performance.

The trade-off between complexity and interpretability is a key takeaway. Choosing the right model depends on whether the priority is explainability or accuracy.