Introduction

This project aims to analyze and predict student performance using machine learning techniques. The goal is to identify key factors influencing student outcomes and provide actionable insights to improve academic success. By combining data cleaning, visualization, predictive modeling, and interpretability methods, the project offers a **complete data-driven workflow**.

Objectives:

- Clean and preprocess student data for modeling.
- Explore relationships between features and outcomes using visualization.
- Train and evaluate predictive models (Decision Tree and Logistic Regression).
- Interpret model predictions using feature importance and SHAP analysis.
- Provide actionable insights for educators and administrators.

2. Data Collection

- Dataset was collected locally (habits.csv) and included attributes such as:
  - Demographics: age, gender, parent_education
  - Academic metrics: gpa, exam_score, study_hours, attendance
  - Extra-curricular activities and resources: extracurricular, internet_access
  - Target variable: Pass/Fail (0 = Fail, 1 = Pass)
- **Initial dataset shape:** 500+ records with 10+ features.

3. Data Preprocessing

Steps Taken:

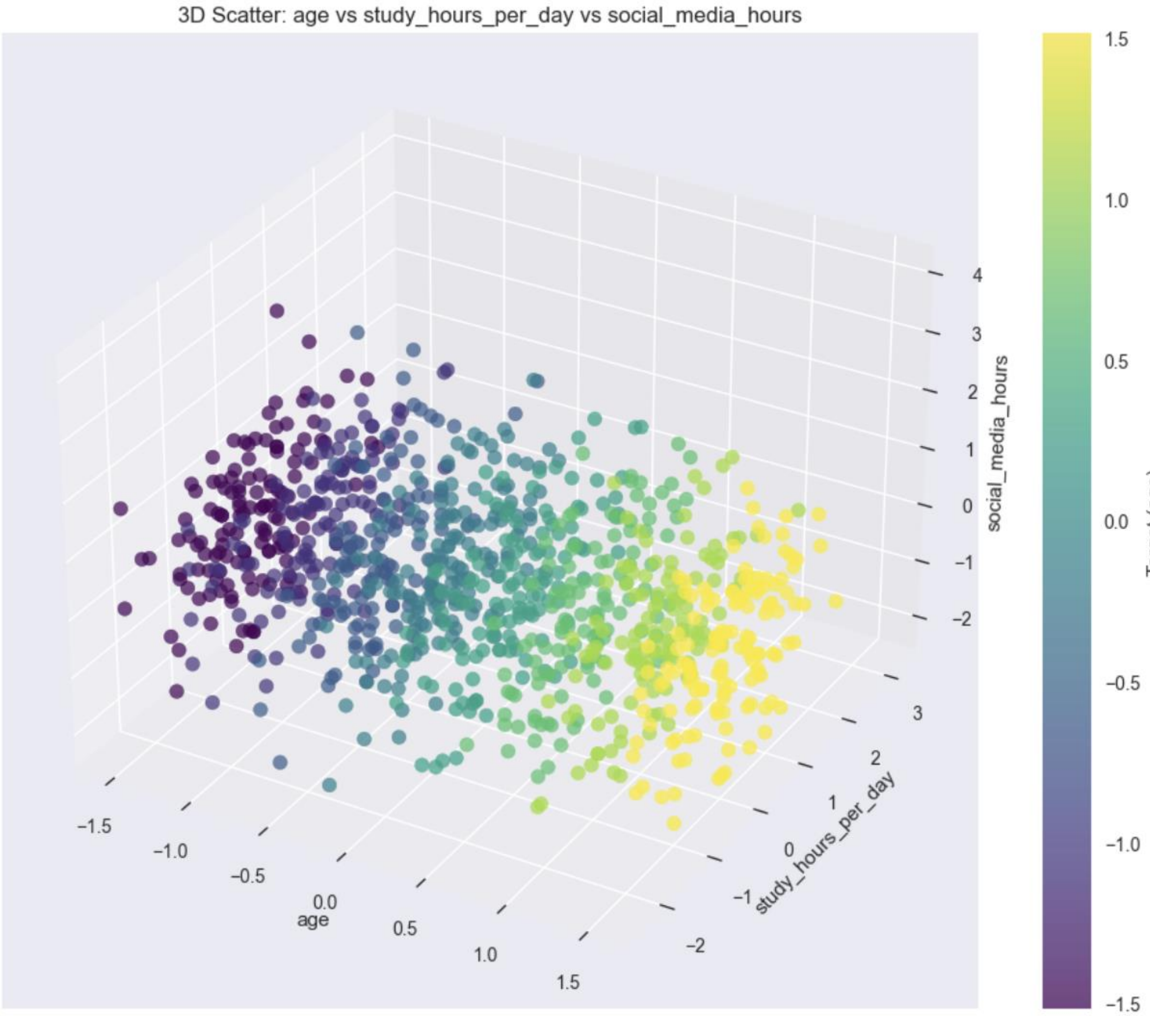
1. **Missing Values Handling**
   a. Numerical missing values filled using **mean** (for low skew) or **median** (for high skew).
   b. Categorical missing values filled using **mode**.

2. **Encoding Categorical Variables**
    1. **Low cardinality:** One-Hot Encoding.
    2. **High cardinality:** Label Encoding.
3. **Feature Scaling**
    1. Standardized numerical columns using StandardScaler to ensure mean = 0 and SD = 1.
4. **Final Cleaned Dataset**
    1. Saved as students_clean.csv
    2. No missing values; ready for modeling.

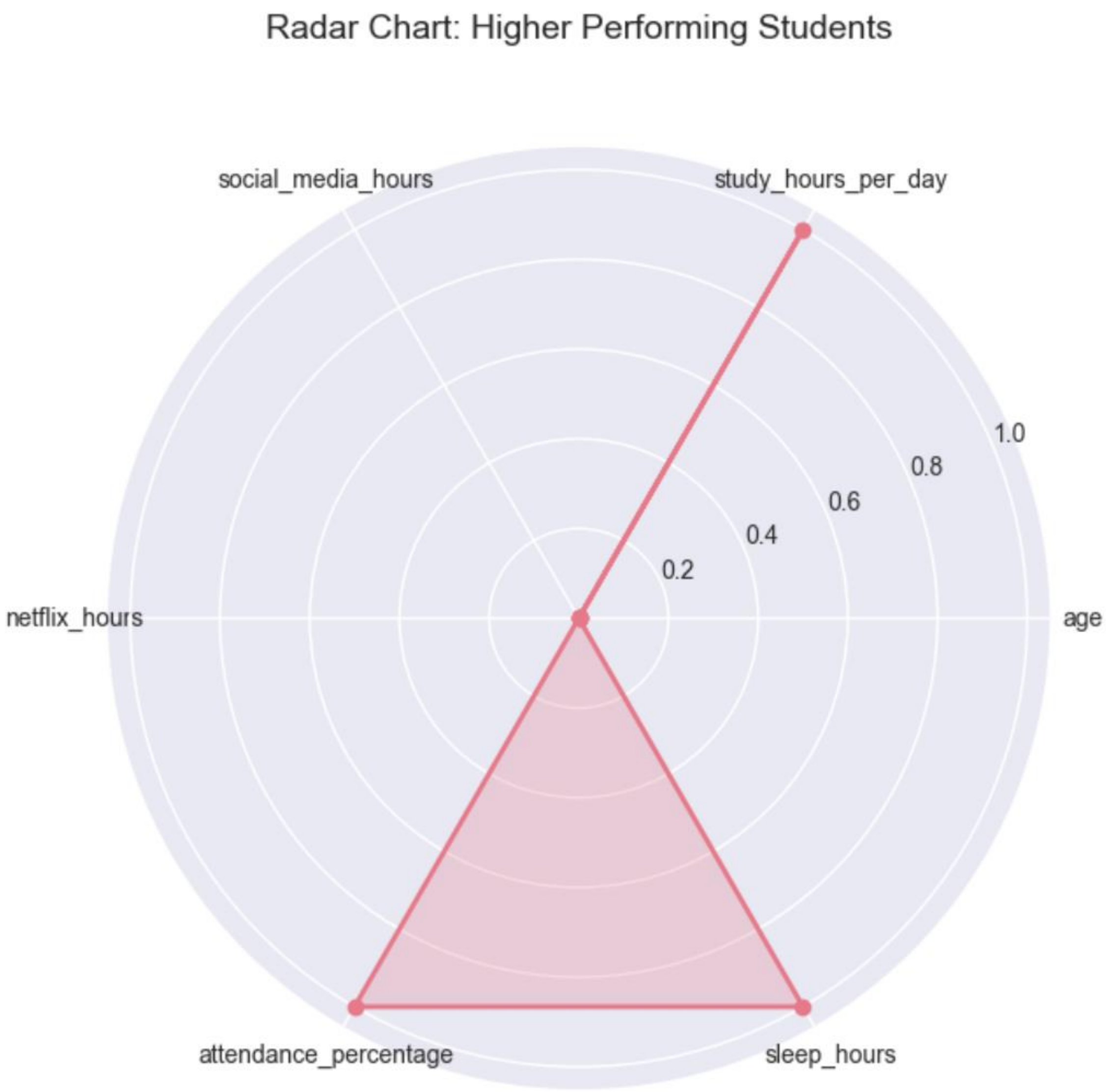
4. Exploratory Data Analysis (EDA)

Visualization Techniques:

- **Scatter Plots (2D & 3D):** Examined relationships between study hours, attendance, and exam scores.



- **Box & Violin Plots:** Compared distributions of numerical features across pass/fail students.
- **Radar Charts:** Compared performance profiles between lower- and higher-performing students.



Key Insights from EDA:

- Students with higher study hours and attendance generally perform better.
- Features like study_hours, attendance, and exam_score strongly differentiate pass vs fail students.
- Identified outliers and distribution patterns to guide feature preprocessing.

5. Modeling

Models Trained:

1. **Decision Tree Classifier**
    a. Max depth = 5
    b. Captures non-linear relationships.
2. **Logistic Regression**
    a. Max iterations = 1000
    b. Captures linear relationships with interpretability.

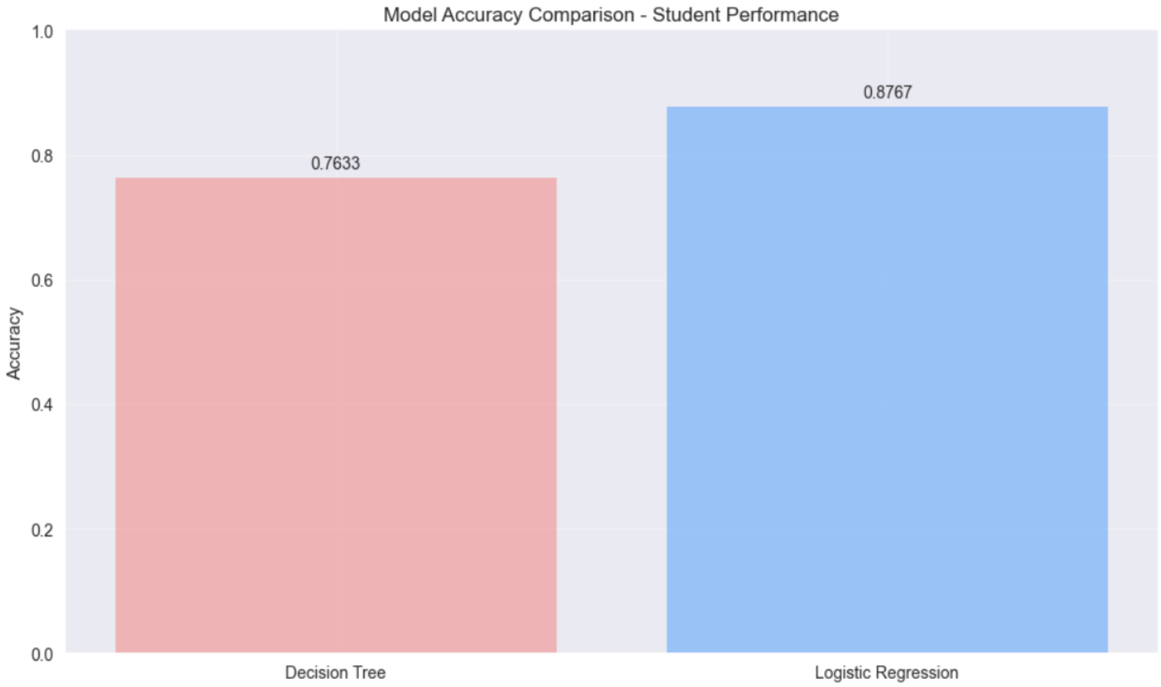
Train-Test Split:

- 70% training, 30% testing
- Stratified split to maintain class distribution.

Evaluation Metrics:

- Accuracy
- Confusion Matrix
- Classification Report (precision, recall, F1-score)

Results Snapshot:



6. Feature Analysis

1. **Decision Tree Feature Importance**

    a. Top features influencing predictions included:
        i. Study hours, attendance, exam score, extracurricular activities.
    b. Visualized with horizontal bar charts.

2. **Logistic Regression Coefficients**
    a. Positive coefficients indicate features that increase probability of passing.
    b. Negative coefficients indicate features associated with failure.
3. **SHAP Analysis**
    a. Global and instance-level explanations using SHAP:
        i. Confirmed most influential features across models.
        ii. Force plots highlighted how individual feature values contributed to predictions.

7. Conclusion

- **Key Takeaways:**
    - Study hours, attendance, and exam scores are the strongest predictors of student success.
    - Decision Tree and Logistic Regression models achieved high accuracy and complementary interpretability.
    - SHAP analysis enhanced understanding of feature impact, providing actionable insights.
- **Recommendations:**
    - Educators should monitor key features (study hours, attendance) to identify at-risk students early.
    - Targeted interventions can improve overall academic performance.
    - The workflow can be extended with additional datasets and advanced models for better prediction.