TEECE 2 Capstone Project
Lifestyle and Learning – Predicting Student Performance

I. Introduction

This project utilizes a simulated dataset of 1,000 student records sourced from Kaggle. Each record captures key lifestyle habits—such as study hours, sleep patterns, screen time, diet, and mental health—and relates them to academic performance, specifically the final exam score. The dataset is ideal for educational machine learning applications, enabling learners to perform data preprocessing, visualization, clustering, regression, and classification.

You are tasked with analyzing how these lifestyle factors affect student performance, and building predictive models based on insights you derive.

Dataset: student_habits_performance.csv

II. Project Goals

- Determine relationships between lifestyle habits and final exam scores
- Discover meaningful student groupings based on lifestyle through clustering
- Build and evaluate models that predict academic performance
- Summarize and communicate findings through data storytelling

III. Project Components

1. Problem Definition
    - Formulate a clear research question.

2. Data Understanding and Preprocessing
    - Load and inspect the dataset
    - Handle:
        - Missing values
        - Categorical variables (apply label/one-hot encoding)
        - Scaling (for models like K-Means and regression)
    - Engineer new features if helpful (e.g., combine screen time metrics)

3. Exploratory Data Analysis (EDA)
    - Visualizations:
        - Histograms for feature distributions
        - Scatter plots and box plots for comparing habits vs. scores
        - Correlation heatmap

4. Clustering (Unsupervised Learning)
    - Apply K-Means clustering using lifestyle features (excluding exam score)
    - Determine optimal K using:
        - Elbow method (inertia plot)
        - Silhouette score
    - Label and describe each cluster

5. Regression Analysis (Supervised Learning)
    - Use the following models to predict Final Exam Score:
        - Linear Regression

    - Decision Tree Regressor
    - Random Forest Regressor
- Evaluate models using:
    - MAE, RMSE, and R^2 score
    - Train/test split and cross-validation

6. Optional Classification Task
- Convert scores into performance levels:
    - Low (bottom 33%), Average (middle 34%), High (top 33%)
- Train classification models (e.g., Logistic Regression, Decision Tree)
- Evaluate with confusion matrix, accuracy, and F1-score

IV. Interpretation and Insights

Prepare a dedicated section in your notebook summarizing your analysis and conclusions. Your insights should include:

A. Feature Importance
- For tree-based models, plot and analyze feature importance
- For linear models, interpret coefficients
- Identify the top 3–5 features that most affect performance

B. Cluster Profiling
- For each cluster:
    - Describe common behaviors (e.g., "Cluster 1 sleeps less, studies more")
    - Associate average exam scores with each group
    - Comment on trends you observe (e.g., does screen time correlate with lower scores?)

C. Model Performance
- Which model performed best? Why?
- Are there trade-offs between interpretability and accuracy?

D. Real-World Implications
- What advice could you give students based on your findings?
- Are there surprising or counterintuitive results?

V. Final Deliverables

A. GitHub Repository
- Upload your complete, well-commented Jupyter Notebook
- Include:
    - README.md summarizing the project, methods, and findings
    - All supporting files (dataset, plots, outputs)
- Use meaningful commit messages and organize code into clear sections

B. YouTube Presentation
- Record a 10 minute (max) video discussing:
    - The problem and your approach
    - Key findings and model results
    - Visualizations and interpretations
    - Recommendations or insights
- Speak clearly, show your notebook, and use visuals to support your discussion