ML Lab Assignment 1: "Student Habits vs Academic Performance"

25MCMI20, Sachin A

23/01/2026

1 Introduction

In this experiment, I analyzed how different student habits influence academic performance using a supervised machine learning approach. A Decision Tree classifier was used to predict grade categories derived from exam scores. This experiment was carried out with a focus on understanding the full pipeline including data preprocessing, exploratory data analysis, model training, and evaluation.

2 Dataset Description

The dataset used in this experiment is the *Student Habits vs Academic Performance* dataset obtained from Kaggle. Each row in the dataset corresponds to a student, and the columns represent different demographic, behavioral, and academic attributes.

The dataset contains the following attributes:

- **age**: Age of the student
- **gender**: Gender of the student
- **study_hours_per_day**: Average daily study hours
- **social_media_hours**: Time spent on social media per day
- **netflix_hours**: Time spent on streaming platforms
- **part_time_job**: Whether the student has a part-time job
- **attendance_percentage**: Percentage of classes attended

- **sleep_hours**: Average hours of sleep per day
- **diet_quality**: Self-reported diet quality
- **exercise_frequency**: Frequency of physical exercise
- **parental_education_level**: Education level of parents
- **internet_quality**: Quality of internet access
- **mental_health_rating**: Self-reported mental health rating
- **extracurricular_participation**: Participation in extracurricular activities
- **exam_score**: Numerical exam score

The column `student_id` was removed during preprocessing since it is only an identifier and does not contribute to prediction.

3 Libraries Used

The following Python libraries were used in this experiment:

- **NumPy**: Used for numerical operations and basic array handling.
- **Pandas**: Used for loading the dataset and performing data cleaning and preprocessing.
- **Matplotlib**: Used for plotting graphs and visualizing results.
- **Seaborn**: Used to generate statistical visualizations such as pair plots.
- **Scikit-learn**: Used for splitting the dataset, training the machine learning model, and computing evaluation metrics.

4 Data Preprocessing

Before applying machine learning models, I inspected the dataset using functions such as `info()`, `describe()`, and `isnull()` to understand data types, distributions, and missing values.

4.1 Handling Missing and Duplicate Values

Missing values were handled using simple strategies (these were strategies that I googled and found were appropriate):

- Numerical features were filled using the median value.
- Categorical features were filled using the mode.

Duplicate rows, if any, were removed to avoid bias during model training.

4.2 Grade Label Creation

The numerical exam score was converted into grade labels based on the SCIS grading scale in this experiment:

Score Range	Grade
≥ 90	O
85–89	A+
75–84	A
65–74	B+
60–64	B
55–59	C
50–54	P
< 50	F

After assigning grade labels, the original exam score column was removed, since labels were already assigned.

4.3 Encoding Categorical Variables

Category based input features were converted into numerical form using something called one-hot encoding. This step is necessary because machine learning models require numerical inputs. The target variable (Grades) was not encoded.

5 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to understand how selected student habits relate to academic performance. Instead of visualizing all feature combinations, I have

presented some focused pair plots for features that seem intuitively relevant to academic outcomes.

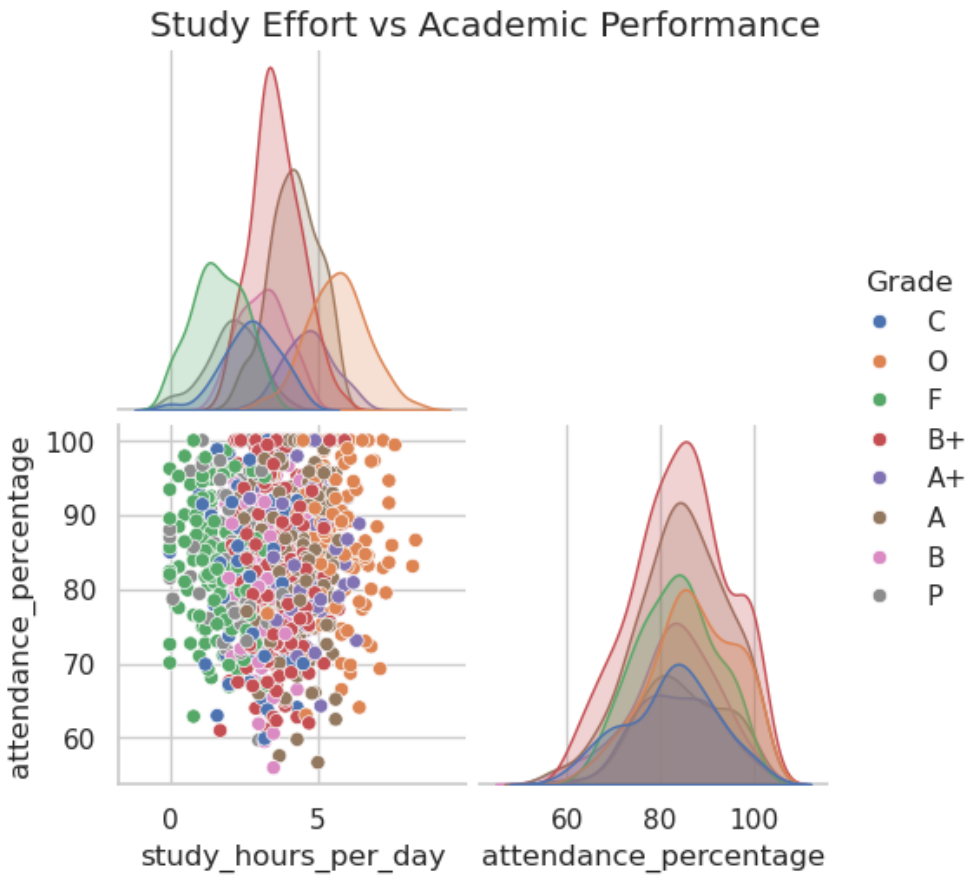

Figure 1: Study Hours and Attendance Percentage vs Academic Performance

From this plot, I observed that higher grades are more frequently associated with higher study hours and better attendance.

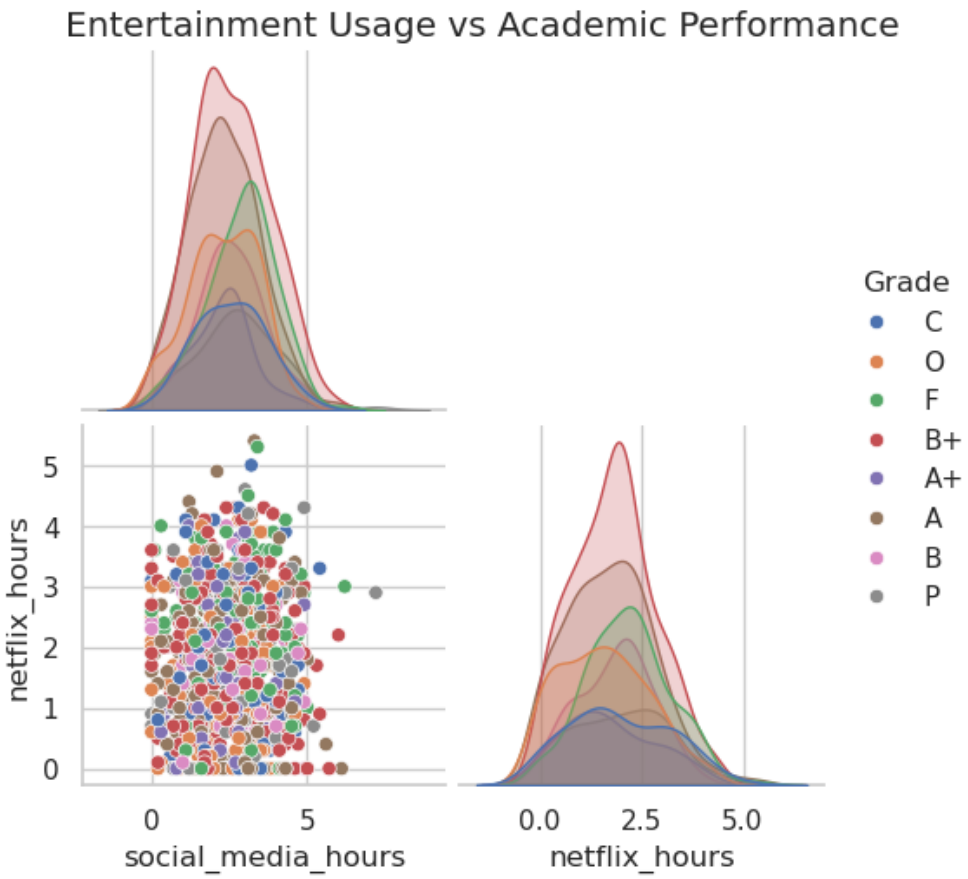

Figure 2: Social Media and Entertainment Usage vs Academic Performance

From this plot, I noticed that students with higher social media and entertainment usage tend to appear more often in lower grade categories, suggesting a possible negative association with academic performance.

6 Model Training

A Decision Tree classifier was selected for this experiment. The dataset was split into training and testing sets using various splits from 90–10 to 50–50.

7 Model Evaluation

7.1 Confusion Matrices for Different Train–Test Splits

To better understand how the classification performance varies with different train–test split ratios, confusion matrices were generated for each split. The confusion matrix is used to

understand how well the classification model predicts each grade category. Each row of the matrix represents the actual grade of a student, while each column represents the grade predicted by the model. The values along the main diagonal correspond to correct predictions, where the predicted grade matches the true grade. The off-diagonal values indicate misclassifications, showing cases where the model predicts a different grade than the actual one. By examining the confusion matrix, I was able to identify which grades are predicted accurately and which grades are more frequently confused with others. This provides a detailed view of the models' performance than overall accuracy alone.

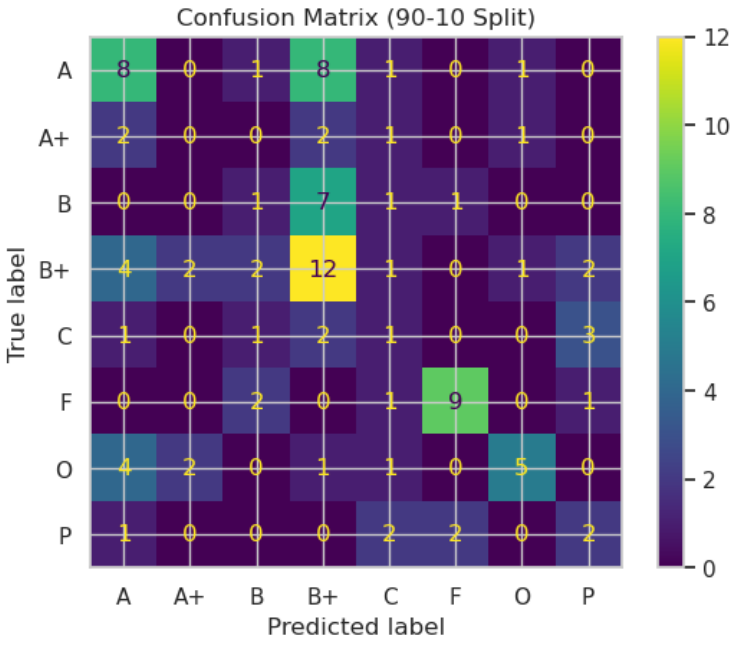

Figure 3: Confusion Matrix for 90–10 Train–Test Split

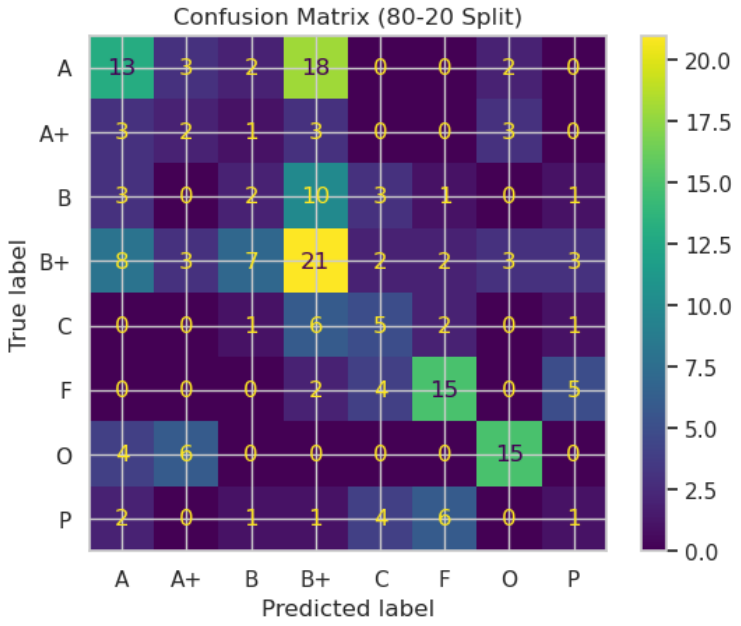

Figure 4: Confusion Matrix for 80–20 Train–Test Split

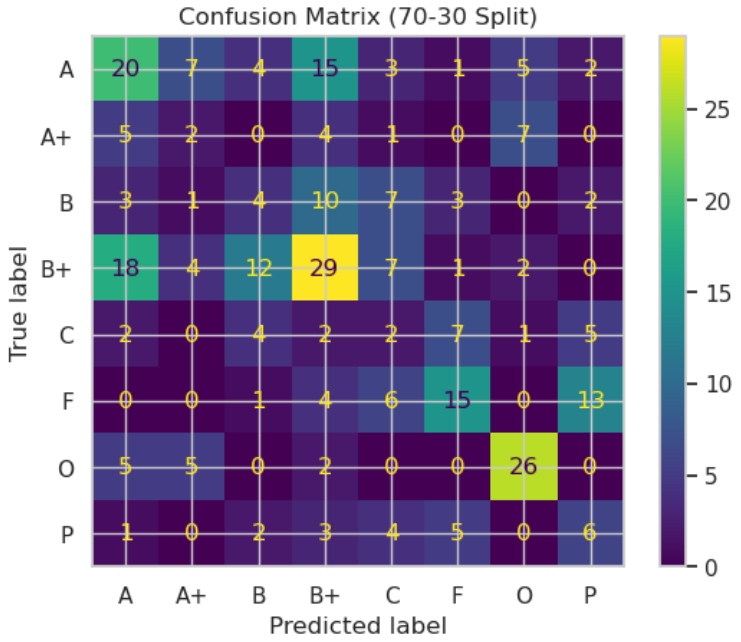

Figure 5: Confusion Matrix for 70–30 Train–Test Split

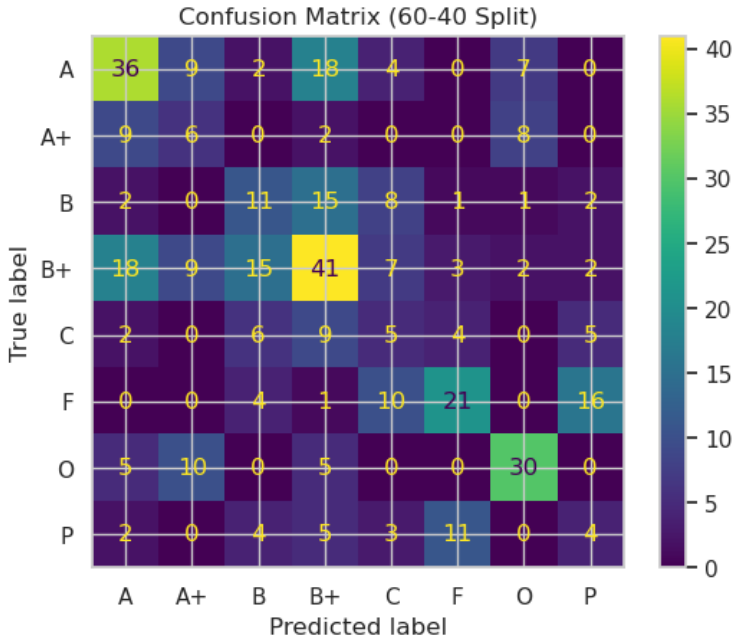

Figure 6: Confusion Matrix for 60–40 Train–Test Split

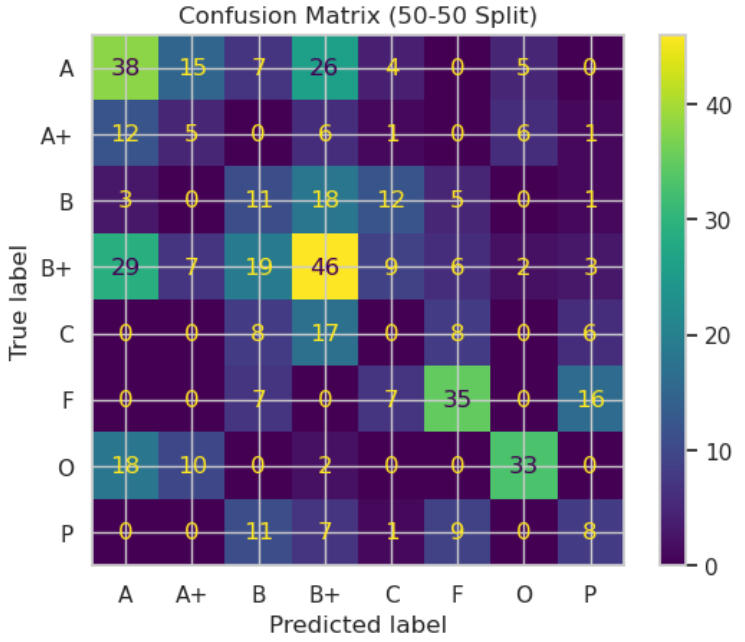

Figure 7: Confusion Matrix for 50–50 Train–Test Split

Overall, the confusion matrices show that prediction stability improves for moderate split ratios such as 70–30 and 60–40, while very small or very large test sets lead to more variation in class-wise predictions.

7.2 Evaluation Metrics

The model was evaluated using the following metrics:

- Accuracy
- Precision (weighted)
- Recall (weighted)
- F1-score (weighted)

Weighted averaging was used to account for class imbalance. I observed that the weighted recall values were numerically equal to accuracy, which is expected in a multi-class classification setting.

8 Effect of Train–Test Split Ratios

The experiment was repeated using different train–test split ratios: 90–10, 80–20, 70–30, 60–40, and 50–50.

Figure 8: Comparison of Evaluation Metrics Across Different Train–Test Splits

From the results, I noticed that model performance was more stable for moderate train–test splits such as 70–30 and 60–40, while very small or very large test sets showed more variation in performance.

9 Conclusion

In this experiment, I applied basic machine learning techniques to analyze how student habits influence academic performance. Through data preprocessing, exploratory analysis, and classification using a Decision Tree model, several meaningful patterns were observed.