Exp AI Lab Assignment-2 Report

Assignment: Feature Importance Analysis using SHAP
Student Name: Veekshitha Adharasani
Roll Number: 2303A52175
Date: 22-08-2025

Introduction

In the education domain, analyzing factors that affect student academic performance is essential for better planning, targeted interventions, and improving overall learning outcomes. Machine learning models can predict student grades by analyzing academic, and behavioral features. To enhance interpretability, SHAP is applied to identify the most influential factors.

This assignment focuses on applying SHAP to the Students Performance dataset to train, identify features and interpret results using SHAP.

Dataset Description

The dataset used is the Students Performance dataset from education domain.

- **Source**: Publicly available on Kaggle.
- **Size**: 145 student records, each representing a unique student's profile.
- **Features (Independent Variables)**: Student ID, Student Age, Sex, High School Type, Scholarship, Additional work, Sports activity, Transportation, Weekly study hours, Attendance, Reading, Notes, Listening In class, Project Work
- **Target Variable**: Grade

Preprocessing Steps

The dataset underwent the following preprocessing steps before model training:

1. **Data Cleaning**:
    - Checked and handled missing values (none found).
    - Removed duplicate rows.
2. **Feature Transformation**:
    - Encoded categorical variables such as *gender*, *parental education level*, and *test preparation course* using Label Encoding.

- Converted target variable *Grade* into numeric labels.

3. **Normalization**:
    - Standardized numerical features (study hours, attendance) for consistent scale.
4. **Data Splitting**:
    - Dataset split into 80% training and 20% testing.

Model & Performance:

A **Random Forest Regressor** was selected as the predictive model due to its ability to capture nonlinear feature interactions and its compatibility with SHAP's TreeExplainer.

- **Model Parameters**:
    - n_estimators = 100
    - max_depth = None
    - random_state = 42
- **Accuracy :** 85%
- **Classification Report:** High precision/recall for major classes, lower for minority classes.

SHAP Analysis :

Explainer Used: TreeExplainer

Plots Generated:

Summary Plot → Shows the global feature importance.

Force Plot → Highlights individual student predictions.

Waterfall Plot → Step-by-step contribution of features for one prediction.

Comparison with Model's Built-in Feature Importance:

Both SHAP and Random Forest importance agree that academic scores are the strongest predictors. SHAP adds interpretability by showing direction of influence (e.g., higher math score increases probability of good grade).

6. Conclusion:

Academic scores (math, reading, writing) are the most decisive factors in predicting student performance. SHAP enhances interpretability compared to traditional feature importance