

STUDENT PERFORMANCE ANALYSIS

Prepared and Presented By: Uyen Mai



MOTIVATION

High college dropout rate



Predict whether an student pass the course or not



Provide early insights into student progress to support academic achievement and course completion



Data Overview

Data Source: Student Performance Prediction (Kaggle)

```
df = pd.read_csv("student_performance_prediction.csv")
df.head(5)
```

	Student ID	Passed	Study Hours per Week	Attendance Rate	Previous Grades	Participation in Extracurricular Activities	Parent Education Level
0	S00001	Yes	12.5	NaN	75.0	Yes	Master
1	S00002	No	9.3	95.3	60.6	No	High School
2	S00003	No	13.2	NaN	64.0	No	Associate
3	S00004	No	17.6	76.8	62.4	Yes	Bachelor
4	S00005	No	8.8	89.3	72.7	No	Master

```
rows, variables = df.shape
print(f"Number of rows: {rows}")
print(f"Number of variables: {variables}")
```

Number of rows: 40000
Number of variables: 7

Data Overview

Data Source: Student Performance Prediction (Kaggle)

Before Data
Processing

Item	Stats
Total variables	7
Multi-categorical variables	2
Numerical variables	3
Binary variables	2
Total records (rows)	40,000
Class distribution	balanced
Missing values	yes



After Data Processing

Item	Stats
Total variables	10
Multi-categorical variables	0
Numerical variables	3
Binary variables	7
Total records (rows)	31,774
Class distribution	balanced
Missing values	no

DATA PREPROCESSING

Step 1: Check missing values and unique values of all variables

Step 2: Handle outliers, missing and categorical data

- All columns: impute missing data for numerical features using mean and for categorical features using mode
- Study Hours per Week: replace negative value with absolute value
- Attendance Rate: remove values greater than 100 or less than 0
- Previous Grades: remove values greater than 100
- Passed, Participation in Extracurricular Activities: replace binary values with 0's and 1's.
- Educational Parent Level: one hot encoding for nominal categorical variable

Step 3: Check correlation among input features to remove one that have strong correlation to others

Step 3:

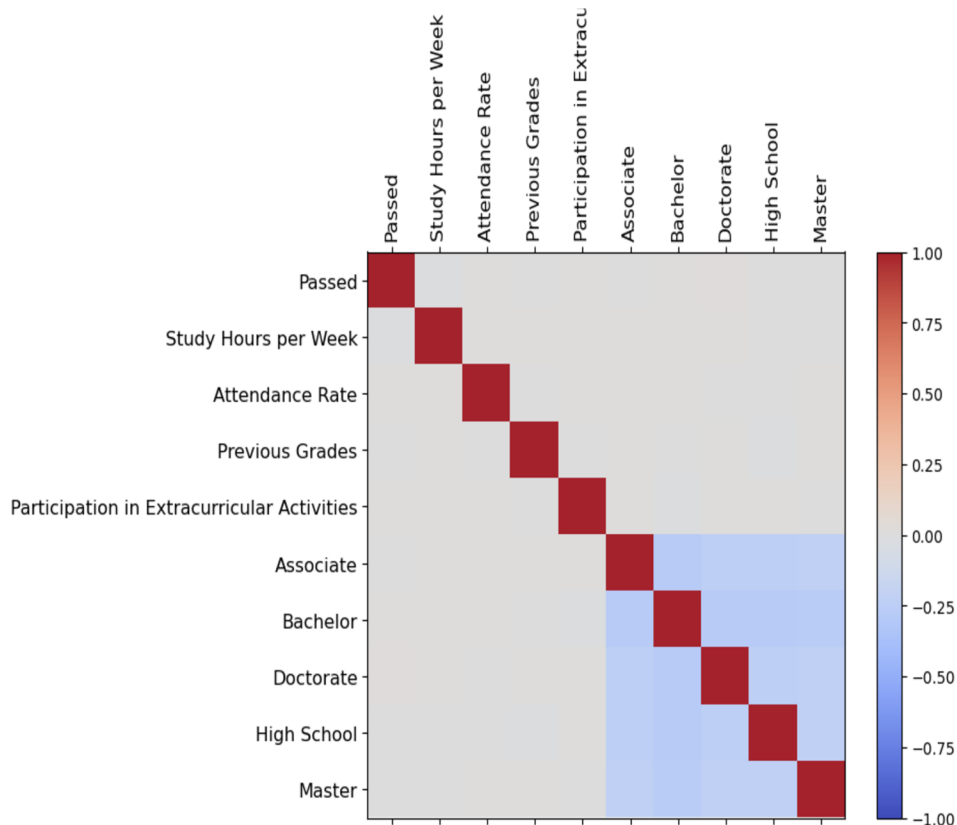
Step 4: Split dataset into: 80% train, 10% validation, 10% test

Step 5: Compute class weight to handle class imbalance in training set

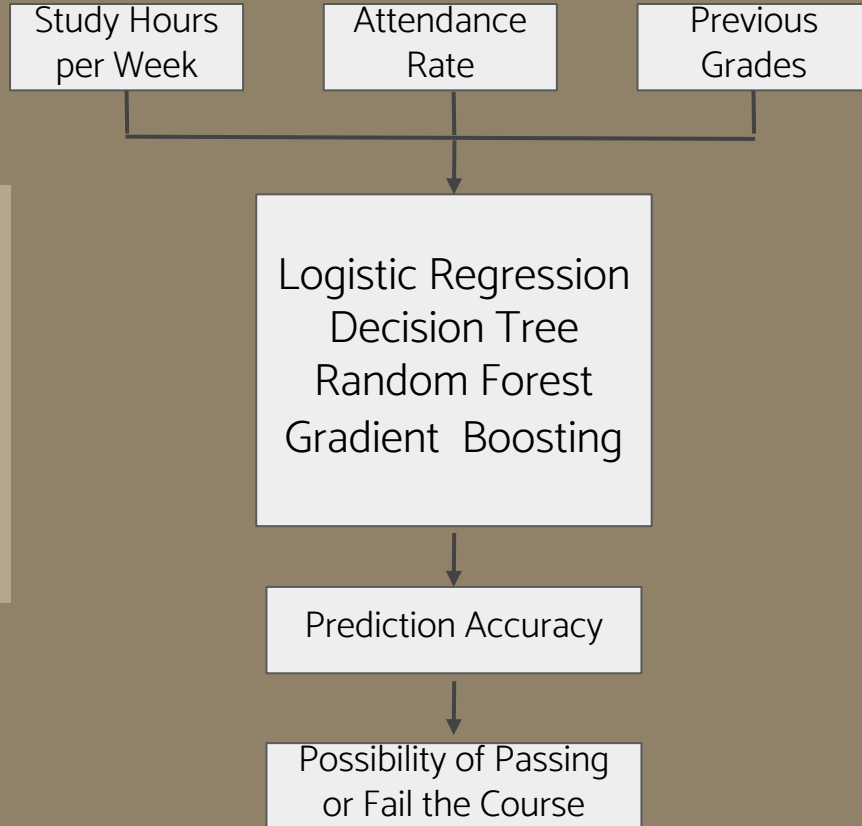
Step 6: Scale numerical input features (Study Hours per Week, Attendance Rate, Previous Grades)

Step 7: Check relationships among input features and target feature

Weak or negligible relationships with the target variable detected



MODELING PROCESS



1. Example Features

2. Prediction Models
(Using GridSearchCV to find the best parameters)

3. Model Evaluation

4. Impact Prediction

MODEL SELECTION

Rank	Models	Accuracy Score (Validation Set)
1	Logistics Regression	0.514
2	Decision Tree	0.500
3	Random Forest	0.520
4	Gradient Boosting	0.528

OUTCOME

- **Data Accuracy:** 0.51 ~ 0.53
- **Areas for Improvement:**
 - Add more features with significant predictive power, such as participation in class through discussion, accumulative grade, teaching quality and etc.
 - Deploy more advanced machine learning models to achieve better results

THANK YOU

