ZENTHIC AI INTERNSHIP

-DEETHIKSHA R

TASK-3

Dataset: Student Academic Performance Dataset

Objective

The aim of this analysis was to study student performance data, understand patterns, and identify factors that affect test scores. The goal was to find useful insights that can help improve student academic outcomes.

Data Cleaning

The dataset was cleaned by:

- Replacing missing and placeholder values
- Standardizing text values (like gender formats)
- Converting numeric columns to proper numeric type
- Handling missing data carefully

This made the data ready for proper analysis.

Univariate Analysis

Histograms were used to study numeric variables like study hours, attendance rate, GPA, and test scores.

Bar charts were used to analyze categorical variables like gender, internet access, and scholarship status.

This helped understand how each variable is distributed.

Bivariate Analysis

Relationships between variables were examined using scatterplots, boxplots, and correlation analysis.

We studied:

- Study hours vs test score
- Attendance vs test score

- Internet access vs performance
- Prior GPA vs test score

Key Findings

1. Students who study more hours generally score higher.
2. Higher attendance is linked to better performance.
3. Students with internet access tend to perform better.
4. Prior GPA is a strong indicator of current test scores.
5. Scholarship students show competitive academic results.

Conclusion

The analysis shows that study habits, attendance, previous academic performance, and access to internet resources play an important role in student success. These findings can help institutions design better academic support programs.

TASK-4 Feature Engineering and Transformation

1. Creation of New Features

To improve model performance and capture meaningful patterns, new derived features were created from existing variables.

a) Attendance Category

The continuous variable attendance_rate was converted into categorical groups:

- **Low** → Attendance < 50%
- **Medium** → 50% – 75%
- **High** → > 75%

This transformation helps the model understand attendance behavior in grouped form rather than raw percentages.

b) GPA Band

The variable prior_gpa_10pt was categorized into performance bands:

- **Low** → GPA < 5

- **Medium** → GPA between 5 and 8
- **High** → GPA > 8

This allows the model to capture academic performance levels more effectively.

c) Study Hours Bin

The variable study_hours_per_week was discretized into:

- **Low** → 0–10 hours
- **Medium** → 10–20 hours
- **High** → Above 20 hours

Binning reduces noise and helps identify study pattern groups.

2. Handling Skewed Numeric Columns

Skewness was checked for numeric variables such as:

- fee_paid_inr
- study_hours_per_week

If a variable showed high positive skewness, a **log transformation (log1p)** was applied to reduce distribution imbalance and improve model stability.

Example:

- fee_paid_log = log(1 + fee_paid_inr)

Log transformation helps in normalizing highly skewed financial data.

3. Encoding Categorical Features

Machine learning models require numeric input. Therefore:

- Categorical variables were encoded using:
    - **Label Encoding** for ordinal categories
    - **One-Hot Encoding** for nominal categories

This ensures that the model can interpret categorical information correctly.

4. Scaling Numeric Features

To ensure all numeric features are on the same scale:

- **StandardScaler** was applied.

Scaling prevents features with large ranges (e.g., fee_paid_inr) from dominating smaller-scale variables (e.g., GPA).

After scaling:

- Mean ≈ 0
- Standard deviation ≈ 1

This improves model convergence and performance.

5. Final Model-Ready Dataset

After completing:

- Feature creation
- Skewness handling
- Encoding
- Scaling

The processed dataset was saved as:

dataset_model_ready.csv

This dataset is fully prepared for machine learning model training and evaluation.

TASK-5 Predictive Modeling

1. Define target variable: test_score

The dependent variable selected for modelling is:

test_score

It represents the academic performance of students. As it is a continuous numerical variable, the problem is defined as a **supervised regression task**.

The dataset was divided into:

- **Feature Matrix (X)** – All independent variables
- **Target Variable (y)** – test_score

2. **Train-Test Split**

To evaluate model performance fairly, the dataset was split into:

- **80% Training Data**
- **20% Testing Data**

The training set was used to train the models, while the testing set was used to evaluate predictive performance on unseen data. A fixed random state was used to ensure reproducibility.

3. **Baseline Models Implemented**

Three regression models were trained and compared:

3.1 Linear Regression

Linear Regression was used as the baseline model. It assumes a linear relationship between input features and the target variable.

- Simple and interpretable
- Serves as benchmark model

3.2 Decision Tree Regressor

Decision Tree Regressor captures nonlinear relationships and feature interactions.

- Handles complex patterns
- May overfit if not controlled

3.3 Random Forest Regressor

Random Forest is an ensemble learning method that combines multiple decision trees.

- Reduces overfitting
- Improves prediction stability
- Generally provides better accuracy than a single tree

4. Model Evaluation Metrics

Two evaluation metrics were used:

4.1 Root Mean Squared Error (RMSE)

RMSE measures the average magnitude of prediction error.

- Lower RMSE indicates better accuracy.
- Penalizes large errors more heavily.

4.2 R^2 Score (Coefficient of Determination)

R^2 measures the proportion of variance explained by the model.

- Range: ($-\infty$ to 1)
- Closer to 1 indicates better model performance.