

Feature Engineering Report – Students Performance Dataset

• Dataset Chosen

The dataset used is the **Students Performance Dataset** from Kaggle, containing data of 1,000 students. It provides insights into student demographics and academic outcomes.

Attributes:

- **gender**: male or female
- **race/ethnicity**: group A–E categories
- **parental level of education**: from high school to master's degree
- **lunch**: standard or free/reduced
- **test preparation course**: none or completed
- **math score, reading score, writing score**: numeric test results

Goal: Analyze and identify which factors influence students' **math score**.

Problem Type: Regression.

• Feature Engineering Techniques Applied

StepTechniquePurposeReasoning
1Missing Value HandlingEnsure data completeness No missing values were found; the dataset was already consistent.
2EncodingConvert categorical variables into numeric format Label and one-hot encoding were used to transform categorical columns such as gender, lunch, and test preparation course.
3Feature ScalingNormalize numeric features StandardScaler was applied to exam scores to reduce scale bias and improve model convergence.
4Dimensionality ReductionSummarize correlated variables Principal Component Analysis (PCA) was performed to reduce exam scores into two key components.
5Feature SelectionIdentify key predictors SelectKBest with Chi-Square test was used to extract the top 5 influencing factors on math scores.

• Reasoning Behind Choices

1. Data was clean and required no imputation, preserving data integrity.
2. Encoding transformed text data into numeric form compatible with ML algorithms.
3. StandardScaler removed scale differences between exam subjects.
4. PCA simplified multi-subject dependencies while retaining the majority of data variance.
5. SelectKBest helped identify which factors (e.g., lunch type, test preparation, reading and writing scores) had the strongest impact on math performance.

• Data Understanding and Correlation Insights

Correlation analysis revealed that **reading score** and **writing score** have a strong positive correlation ($r \approx 0.95$), and both are highly correlated with **math score** ($r \approx 0.85$). This indicates that students performing well in one subject tend to perform well in others. PCA visualization further confirmed that a large proportion of variance is explained by a single academic performance

component.

Students who completed the test preparation course generally had higher scores across all subjects. Lunch type also showed influence — students receiving standard lunch tended to perform better on average.

• Observations on Transformed Features

1. The dataset became fully numeric and free of missing values.
2. Categorical variables were efficiently converted for modeling.
3. Feature scaling aligned value ranges for model fairness and interpretability.
4. PCA revealed underlying learning patterns among students, indicating consistent performance across subjects.
5. Feature selection optimized model inputs by retaining only the most relevant attributes.
6. The preprocessed dataset is now compact, standardized, and well-suited for regression modeling or machine learning experiments.

• Ethical Concerns & Bias Mitigation

Education data contains sensitive demographic features. Attributes like gender or race can lead to biased models if not handled carefully.

Ethical preprocessing ensures fairness and equal treatment by:

1. Excluding sensitive variables where unnecessary.
2. Using fairness evaluation metrics such as *demographic parity* or *equal opportunity*.
3. Applying bias mitigation libraries like **AI Fairness 360** or **Fairlearn**.
4. Ensuring transparency — documenting every transformation and explaining its purpose.

The analysis focuses on academic and behavioral factors rather than personal demographics to maintain ethical use of data.

• Future Scope

Further analysis could include:

1. Applying machine learning models such as Linear Regression, Random Forest, or XGBoost to predict math scores.
2. Exploring feature importance to understand the relative impact of socio-economic factors.
3. Conducting fairness analysis to ensure unbiased model performance across demographic groups.
4. Extending the dataset to include additional academic years or institutions for generalization.

Prepared By:

NIKHIL SHIRSATHE

202401110003