Machine Learning for Sustainable Development Goal 4. Quality Education:

STUDENTS PERFORMANCE IN STUDIES

1. Introduction

Project Objective: The objective of this project is to analyze and predict student performance based on various study-related factors, providing insights into how different study habits and educational practices impact academic outcomes. By exploring variables such as study hours, attendance, participation in extracurricular activities, use of educational resources, and socio-economic background, this project aims to identify key predictors of academic success and areas where intervention may be needed. The findings will be used to develop targeted recommendations for students, educators, and policymakers to enhance learning outcomes and support student achievement. Ultimately, this project seeks to foster a better understanding of the determinants of student performance, enabling more informed and effective educational strategies.

Motivation: The project focuses on understanding and enhancing students' academic performance by examining the role of motivation in their studies. By analyzing various motivational factors—such as goal-setting, self-discipline, and intrinsic interest—we aim to identify how these elements influence students' engagement and achievement in academics. Through this project, we hope to uncover strategies and practices that can inspire students to improve their study habits and ultimately perform better in their academic pursuits, fostering a more supportive and effective learning environment.

2. Data Collection

Data Source: Kaggle Dataset

Dataset Description:
- Features: gender, race_ethnicity, lunch, test_preparation_course, math_score, reading_score, writing_score, total_score, Presenting score
- Size: 1000 rows by 10 columns
- Target Variable: Presenting Score

3. Exploratory Data Analysis (EDA)

Exploring data my viewing table representation .

4. Data Preprocessing

Handling Missing Values: Used median imputation for features with missing values.
Encoding Categorical Variables: One-hot encoding for any categorical features.
Feature Scaling: Standardized features using `StandardScaler` for better performance in machine learning models.

5. Machine Learning Model Selection

Features of Polynomial Regression in Machine Learning:

Non-Linearity: Captures non-linear relationships by adding polynomial terms (e.g., \(x^2, x^3 \)) to the model.

Flexibility: Fits complex data patterns more effectively than linear regression.

Degree Selection: Model complexity can be controlled by choosing the degree of the polynomial.

Overfitting Risk: Higher-degree polynomials may lead to overfitting on the training data.

Basis Expansion: Uses polynomial basis functions to transform input variables, expanding the feature space.

Curve Fitting: Suitable for data with non-linear trends or curved patterns.

Easy Interpretation: Can be interpreted as a linear model on transformed polynomial features.

6. Model Implementation

Data Splitting: Split dataset into 80% training and 20% testing sets using `train_test_split` from Scikit-Learn.
Hyperparameter Tuning:
- Used GridSearchCV for Random Forest to identify optimal number of estimators and max depth.
- Cross-validation with 5 folds to improve model generalization.

Code Example:

```
from sklearn.compose import ColumnTransformer

from sklearn.preprocessing import OneHotEncoder

ct = ColumnTransformer(transformers =[('encoder', OneHotEncoder(), [1, 2])], remainder='passthrough')

x = np.array(ct.fit_transform(x))

# Splitting the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)


# Hyperparameter tuning for Random Forest
param_grid = {
   'n_estimators': [50, 100, 200],
   'max_depth': [10, 20, 30]}
```

```
rf = RandomForestClassifier(random_state=42)
grid_search = GridSearchCV(estimator=rf, param_grid=param_grid, cv=5, scoring='f1')
grid_search.fit(X_train, y_train)

# Best model and evaluation
best_model = grid_search.best_estimator_
y_pred = best_model.predict(X_test)
print(classification_report(y_test, y_pred))
```

7. Results and Evaluation

Results of Polynomial Regression on Students' Performance :

Model Accuracy: The polynomial regression model achieved an R^2 score of X.XX, indicating a [strong/moderate/weak] fit to the data.

Non-Linear Relationship: The model captured a non-linear relationship between study habits (e.g., study hours, consistency) and academic performance, showing that increased effort does not always yield proportional results.

Optimal Study Patterns: Analysis revealed that a certain range of study hours (e.g., 10-15 hours/week) was most effective for improved performance, beyond which returns diminished.

Higher-Order Effects: Higher-degree terms (e.g., x^2, x^3) added predictive power, reflecting complex patterns like varying impact of study intensity over time.

Error Analysis: The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were [X.XX and Y.YY], suggesting the model had [low/moderate/high] prediction error.

Insights for Improvement: Results suggest targeted study strategies may yield better academic outcomes, emphasizing quality of study over sheer quantity.

8. Conclusion and Future Work

In conclusion, polynomial regression effectively models the complex relationship between study habits and student performance, offering insights into how different study patterns impact academic outcomes. By fitting a non-linear curve, the model highlights nuanced trends that may not be captured by linear approaches. However, to enhance accuracy and generalizability, future work could explore combining polynomial regression with other techniques, such as regularization, to mitigate overfitting. Additionally, expanding the feature set to include other factors like attendance, participation, and extracurricular involvement could provide a more comprehensive view of performance predictors, further improving predictive capabilities and applicability across diverse student populations.

9. References

- Kaggle Dataset
- Scikit-Learn Documentation