Purpose

The purpose of this project is to apply Exploratory Data Analysis (EDA) techniques to a dataset to select the relevant features to build and compare Machine Learning models to predict a student's performance on exams based on their score. The techniques, libraries and models used during the project are in line with the content covered by the course "Introduction to Machine Learning: Supervised Learning" of the University of Colorado Boulder and **are for educational purposes only as the dataset is fictional**.

About the Data

Source

Dataset URL: https://www.kaggle.com/datasets/desalegngeb/students-exam-scores Dataset single sourced from: Expanded_data_with_more_features.csv Size: 15 columns, 30k rows , 2.9MB

Part A: Exploratory Data Analysis (EDA)

As part of the EDA, 2 steps were performed before to first clean the data and then visualize it to build the ML model:

1. Data Cleaning

As part of the data cleaning we'll evaluate the dataset and adjust the data to fit our Machine Learning Model, as follows:

1. Target Value selection
2. Feature Cleaning

2. Data Visualization

1. Checking the distribution of categorical features
2. Checking the distribution of numerical features

Learnings Target Selection:

When comparing the distribution curves for the 3 calculated options (average score, best score and worst score) it is possible to see that the distribution for best and worst are slightlty dislocated from the center of the data. Best score provides a higher median score while the worst score provides a lower median score. **Avg score** provides a smoother distribuition with a distribution in line with the student scores fairly representing the data and the best of the options to use as a target for our model.

Learnings Feature cleaning:

Data Cleaning: After cleaning the data we are left with 19k observations, 10 features and 1 target. Out of the 10 features 8 were are numeric and 2 are categorical.

We can see that some columns have larger number of null values (e.g UsesSchoolBus - 3k null values). For the future, depending on the relevance of this feature and others to the regression model, we could it feature from the feature list before dropping NAs and this would ensure we keep a higher number of the observations.

By doing the right data modifications our model has now only 2 categorical columns instead of 11, making it easier in the future to build regression models without a large increase in the number of features (curse of dimensionality)

Data Visualization: When looking at the categorical features, one can note that the average and the distribution for the Ethnic Group E is a little above the the other groups but not noticiable difference between groups in different ParentMaritalStatuses.

By evaluating the pairplot and correlation matrix for the numerical features, there is no clear correlation between them and overall the correlation of the individual features with the target are low, with StandardLunch as the highest feature with a low 0.32 correlation score.

In order to find the right features to build a regression model, we'll need to do further investigation to find the best feature combination for the model.

Part B: Building Machine Learning Models

1. Create a training and testing dataset
2. Model A: Linear Regression Models
3. Model B: Support Vector Machines (Regressor)
4. Evaluate Model performance

4.Evaluating Model Performance

The table below describes the model performance against test data for each of the models build in this exercise and it's relevant metrics:

- R2: R Squared
- MSE: Mean Squared Error
- MAE: Mean Absolute Error

It is possible to notice that the performance of the models were very close, with SVR M2 having the best R squared, followed by Linear Regression M1. When looking at the erro values, It is possible to see that the Support Vector Machines were able to deliver lower values than the linear regressions, very likely due to the optimization of the parameters coming from applying RandomSearchCV.

Part C: Conclusion & Discussion

Model Performance:

- None of the models built in this exercise displayed a relevant R squared, meaning that they won't do a good job predicting the target values. In this case, this is a high indicator that these features are not relevant enough to predict a student's score.
- The linear regression models had a very similar performance than the SVM models, if we had to chose a model, we should pick the Linear Regression M1 as it is always better to use simpler models if they had similar performance tha more complex ones.

Key Learnings:

- Data Cleaning: By doing the right data modifications to the dataset, the models ended up with only 2 categorical columns instead of 11. Making it easier to build and understand the regression models. Without this previous data cleaning, our models would end-up with a large increase in the number of features, increasing complexity.
- Null Values: When cleaning the data, almost one third of the observations were dropped due to null values. In a real scenario, I would try and evaluate what features are relevant to the problem and drop them before removing their null values. Additionally, we should consider feature interpolation in scenarions where it would be too expensive to drop the data.
- Target Selection: Aiming for simplification of the model and interpretability of the data, I decided to use the average score of the students' scores in different disciplines. This was only possible as the 3 different scores ['MathScore', 'ReadingScore', 'WritingScore'] were highly correlated and their distributions were very similar.

Future Exploration:

- Target Change: Predicting scores prooved to challenging with the current dataset.Dependending on the goal of a future exploration, I would convert the scores to letter grades (A, B, C, etc). Enabling the use of classification models.
- Feature Interpolation: In order to improve the usable data size, we could use interpolation method to try and estimate observations based on similar observations.