

Project Summary

Project Title: What Should Students Do To Ace Their Final Exam?

Group Members:

Theingi Aung (uda7cf)

Megan Chhu (tvg8yz)

Isidore LaRocco (zdn5uf)

Experimental Set-Up:

We will predict students' final exam performance using two different machine learning methods, k-Nearest Neighbors classification and multiple linear regression, and compare the requirements and outcomes of each model. Our goal is to use a small number of features (no more than five, but preferably two) to make accurate predictions before using all predictive features in our data set. Our winning model will be the one with more accurate predictions.

Methods:

Data Set: We used the “Student Performance Factors” data set with 6,607 students’ habits, environment, demographics, and final exam results from Kaggle.

Data: We wrote code to check for nulls and outliers, and to encode categorical data. We wrote a Train-Validate-Test split function. We used the `pd.corr()` function to make a correlation matrix.

kNN: We wrote a kNN regression algorithm by adapting kNN classification algorithms from the homeworks. We wrote code to find the best “k” for the predictive features used.

Linear Regression: We used scikitlearn’s `LinearRegression()` function for our linear regression model.

Model Analysis: We wrote code for a root mean square error (RMSE) function and applied this RMSE function to our predicted test labels to compare our two models.

Results

We found that kNN classification was an entirely incorrect approach to our prediction problem.

After applying kNN regression, we found that the two models performed similarly when using only two predictive features with linear regression slightly outperforming.

We found that linear regression outperformed kNN regression when using all of our 19 predictive features.

Learning Objectives

Correlation matrices are useful for pre-analyzing data to avoid overfitting certain features (in addition to looking really cool).

kNN regression is distinct from kNN classification and can be considered when predicting continuous labels. But, kNN regression does not perform as well as linear regression with many predictive features.