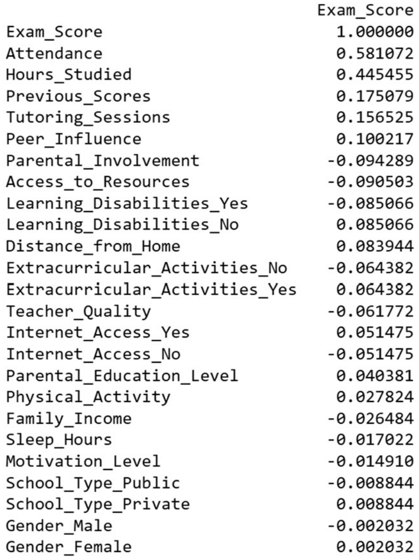
Problem Description

The problem is about prediction of the student scores from the activities such as studies hours, sleep hour, Sample question preparation and so on. Students nowadays are stuck with different kind of studies hours and the effective use of their time is important in order to have maximize the efficiency. Through the prediction, the student could estimate their final score and enhance their targeted score by leveraging their capability. With the experience with this sample dataset, it could be applied from the start of the survey and till the prediction of the score in our country.

Dataset

The data from the Kaggle, StudentPerformanceFactors data file. In the dataset, there are 20 features with one Target variable, the Exam Score with a total of 6000 data. Among those 20 features, apart from the tutoring session and some outliers, the data is nearly normally distributed. In order to select the features, correlation between the exam score and the rest is calculated and resulted as below.



```
                              Exam_Score
Exam_Score                      1.000000
Attendance                      0.581072
Hours_Studied                   0.445455
Previous_Scores                 0.175079
Tutoring_Sessions               0.156525
Peer_Influence                  0.100217
Parental_Involvement           -0.094289
Access_to_Resources            -0.090503
Learning_Disabilities_Yes      -0.085066
Learning_Disabilities_No        0.085066
Distance_from_Home              0.083944
Extracurricular_Activities_No  -0.064382
Extracurricular_Activities_Yes  0.064382
Teacher_Quality                -0.061772
Internet_Access_Yes             0.051475
Internet_Access_No             -0.051475
Parental_Education_Level        0.040381
Physical_Activity               0.027824
Family_Income                  -0.026484
Sleep_Hours                    -0.017022
Motivation_Level               -0.014910
School_Type_Public             -0.008844
School_Type_Private             0.008844
Gender_Male                    -0.002032
Gender_Female                   0.002032
```

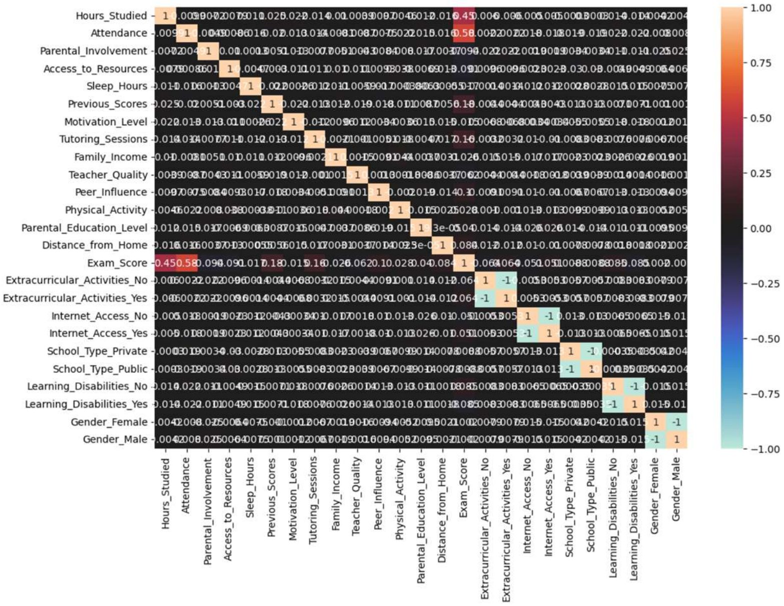


Table1 **Figure1**

Based on the correlation Table, Attendance, Hour_studied, Previous_Scores, Tutoring_sessions, Peer_Influence, Paretal_Involvement, Access_to_Resources, Learning_Disabilities_Yes, Learning_Disabilities_No, Distance_from_Home features are selected.

Modeling Approach

The data does not have missing value but there are Nominal and Ordinal data in the feature and the data is preprocess as below two type of encoding;

One hot encoding : is applied to nominal categorical features (e.g., school_type = [Public, Private], Gender = [Male, Female])

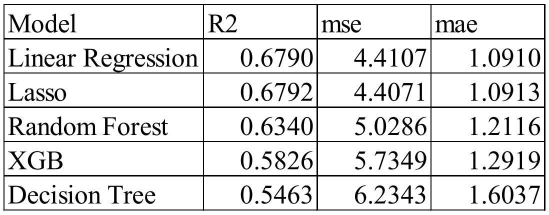
Label encoding : is applied to Ordinal Data (e.g., parental_income = [Low, Medium, High], Distance_from_Home = [Near, Moderate, Far])

After performing train test split, standard scaler is applied to X in order to normalize the feature data (X) by centering to zero mean and scaling to unit variance.

In order to find the best performance, Linear_Regression(Lasso) , Random Forest, Decision Tree and XGBoost models are used in the prediction.

Evaluation

Upon checking the four model, the different r2, mse and mae is resulted.



Model	R2	mse	mae
Linear Regression	0.6790	4.4107	1.0910
Lasso	0.6792	4.4071	1.0913
Random Forest	0.6340	5.0286	1.2116
XGB	0.5826	5.7349	1.2919
Decision Tree	0.5463	6.2343	1.6037

Table 2

Base on the resulted, Lasso Linear is the best performance intern of r2 score and MSE. So, the Lasso linear can be use for the further prediction.

Reflection

The main challenges of the prediction is looking for the data and the coding. As the familiarity of the Kaggle is low, looking for the relevant dataset take longer times and I don’t remember codes. As a learning, the correlation method is learned in the selection of the features and the new model XGBoost is also learned to used in model selection although deep understanding is still needed. By analyzing the dataset, I have found out that different dataset can be relevant to different models and they cannot be constant. I am looking forward to leverage my knowledge and try for the different fields such as environmental, medical fields with some missing data in order to expend my knowledge and understanding while at the same time improve my coding skill.