

Student Performance Prediction Project

We will approach this problem using complete MLDLC (Machine Learning Development Life Cycle).

Step-1: Framing the Problem

Q1) What is the actual problem?

A1) Our problem is to predict a student's performance on the basis of his/her efforts, attitude towards education. So, on the basis of the student's previous scores, questions practiced, sleep hours and more we will predict his/her performance index.

Q2) Who is our target audience?

A2) Students who are having sufficient knowledge of their scores and are willing to quantify their performance by predicting performance index.

Q3) What are the basic characteristics of ML Model for this problem?

A3) To solve this problem I will be using supervised machine learning and training will be done in offline mode(without the use of any API). This will be a regression problem.

Q4) What do I know about data?

A4) I will use the data from Kaggle('Student_Performance.csv'). It has data of 10,000 students. As input I have 5 columns and as output I have performance index. All the columns have numerical data except 'Extracurricular Activities' as it is categorical data.

Q5) What will be my approach to solve this issue?

A5) I will follow all the steps of ML Development Life Cycle till 'Model Deployment'.

Step-2: Data Gathering

Data is used from **Kaggle**. Provided By **Nikhil Narayan**.

Source: <https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression>

Data Sample:

	Hours Studied	Previous Scores	Extracurricular Activities	Sleep Hours	Sample Question Papers Practiced	Performance Index
8920	1	97	No	5	4	72.0
8561	7	85	Yes	9	7	78.0
3039	9	77	No	7	3	74.0
7889	3	62	Yes	6	6	44.0
6661	2	58	No	4	9	35.0

About Data:

Inputs:

- 1) Hours Studied: The total number of hours spent studying by each student.
- 2) Previous Scores: The scores obtained by students in previous tests.
- 3) Extracurricular Activities: Whether the student participates in extracurricular activities (Yes or No).
- 4) Sleep Hours: The average number of hours of sleep the student had per day.
- 5) Sample Question Papers Practiced: The number of sample question papers the student practiced.

Output:

Performance Index: A measure of the overall performance of each student.

Step-3: Data Preprocessing

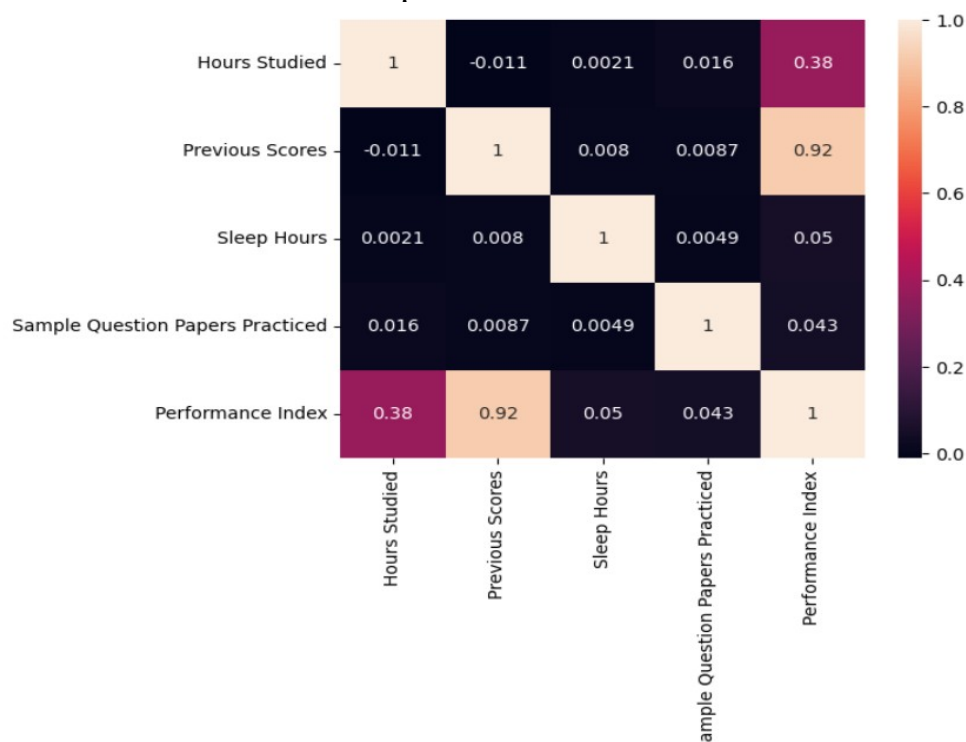
- Since, None column has max value greater than 128 we can reduce the datatype to int8 and float16 to save memory. Using this step I reduced the memory usage from 468.9+ KB to 136.8+ KB (Almost 70.82% Decrease).
- I also dropped 127 duplicate values.

Step-4: Exploratory Data Analysis (EDA)

Since, This is my streamlit project (project with proper GUI) so I have not explored such steps here to see complete understanding, data analysis, visualizations and experimentation on this dataset you can see my notebook on Kaggle.

Link: <https://www.kaggle.com/code/uditkishoregagnani/complete-ml-model-development>

Example: correlation matrix of input variables



Step-5: Feature Engineering and Selection

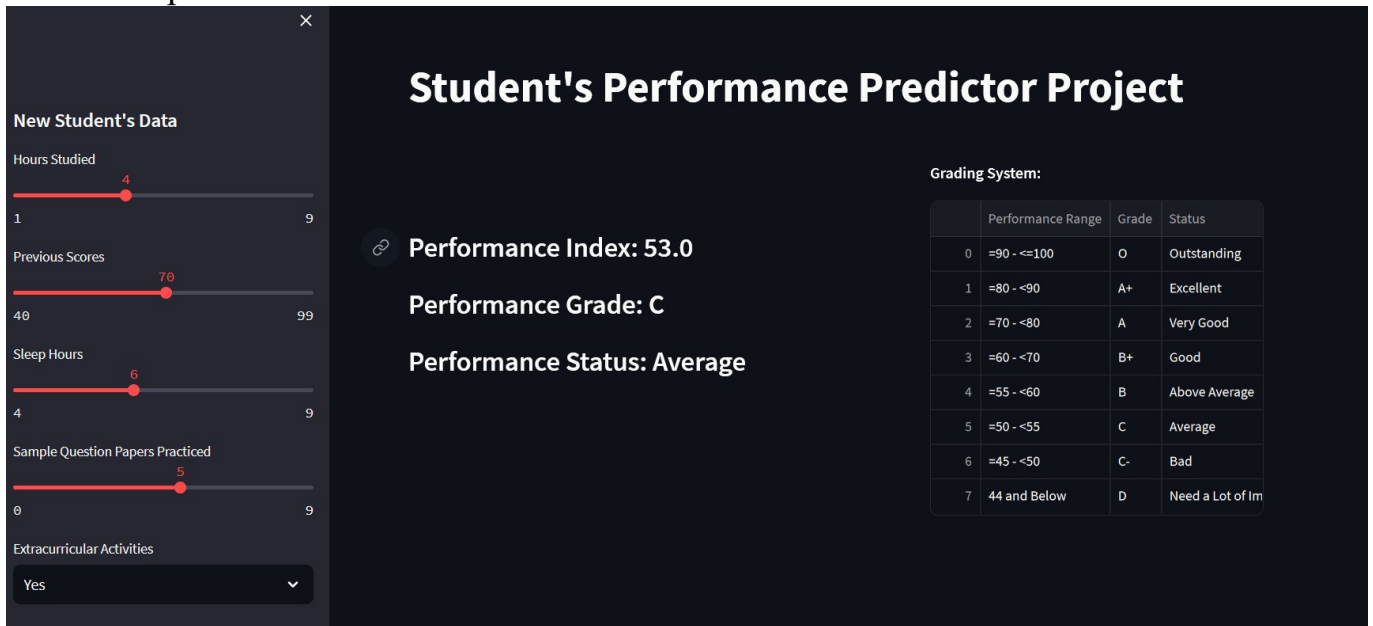
- I used Pipeline to do feature engineering tasks (in this case OneHotEncoding and MinMaxScaling).
- Using this pipeline I fit_transform my training data and transform my test data.

Step-6: Model Training, Evaluation and Selection

- To solve this particular problem I decided to go with Stacking.
- Here I used 5 different algorithms (SGDRegressor, Ridge Regressor, SVR, GradientBoosting Regressor as 4 estimators and Linear Regression as final estimator).
- I did try various other models and RandomSearchCV to look for best hyper-parameter for each model I selected. (Everything is in my Kaggle Notebook).
- This model gave me minimum 98.75% accuracy everytime I used it and that too with 10 cross validations.

Step-7: GUI for This Model

Sample:



	Performance Range	Grade	Status
0	=90 - <=100	O	Outstanding
1	=80 - <90	A+	Excellent
2	=70 - <80	A	Very Good
3	=60 - <70	B+	Good
4	=55 - <60	B	Above Average
5	=50 - <55	C	Average
6	=45 - <50	C-	Bad
7	44 and Below	D	Need a Lot of Im

Since, Extracurricular Activities is a categorical variable so it is a select box and rest all input variables are sliders.

I have improved this project a bit as the model is predicting only Performance Index value and so I created various if, elif and else conditions to see the student lies in which category of grading system. (I also cheated a bit by replacing 'Fail' with 'Need a lot of Improvement' as I feel 'Fail' is a very negative word and may demotivate a student).

Step-8: Scope of Improvement / Future Enhancements

In future a bigger project can be created for not just one student but for whole school where teachers can monitor each students past and current progress and can also forecast future scores. Also this model can be deployed on Cloud.