Introduction to Machine Learning_Midterm Project Report

This project tries to predict the exam score of the student based on their studied hours, attendance, tutoring sessions and previous exam score. As a student, this problem has been an interesting and familiar topic to do. By doing the topic that is familiar to me, I think it will help me to understand machine learning models more effectively without the worry of not understanding the content and context of the data.

The dataset is the synthetic dataset generated for educational and analytical purposes that was found on Kaggle. The data do not specify the sector of education and subject of the exam score. The data has 20 columns, numerical and categorical. Among the 20 columns, there are 7 numerical columns that could be used for machine model: 'Hours_Studied', 'Attendance', 'Sleep_Hours', 'Previous_Score','Tutoring Sessions', 'Physical_Activity' and 'Exam_Score'. As we will predict the final exam score, the target of this problem will be 'Exam_Score'. For features, I choose 'Hours_Studied', 'Attendance', 'Tutoring_Sessions' and 'Previous_Score' for this problem. There are other features but I think they are not that relevant to the target.

As a beginner, I choose linear regression to use for this problem because of its simplicity. I use the train_split_test method from *sklearn* python library to split the data into train data and test data for the preprocessing. Train data will be used to train the model while test data will be used to test whether the model is accurate or not. Then I put the four chosen features, 'Hours_Studied', 'Attendance', 'Tutoring_Sessions' and 'Previous_Score', into variable X and target, 'Exam_Score' into y. Then I use the *StandartScaler* method from *sklearn.preprocessing* to scale the data. To build the linear regression model, I called the *Linear_Regression* method from *sklearn.linear_model* and fit the scaled train data into the model. Then I generate the coefficient of each feature to the target and the intercept from the linear regression.

For the evaluation, I use mean absolute error, mean square error and R2 score for both the train data and test data. For the train data, I got Mean absolute error: 1.34, Mean squared error: 6.61, R2 score: 0.58. The test data also get Mean absolute error: 1.28, Mean squared error: 4.93 , R2 score: 0.64. The predicted data will be a little different from the actual data and the fitting is more than 50%, which indicates a good model.

I did not face much challenge for this problem because I chose linear regression which is easy to understand. But I learnt that even the easiest model requires hand-ons experience to

understand the model and the method thoroughly. As this is my first hand-ons project for machine learning, I think I will try to do more projects for the other lessons. Next time, I think I will challenge myself and do another model that is not familiar to me. Another lesson is to choose a real life dataset from a different sector which might pose more challenges and give more lessons to learn.

Dataset source - https://www.kaggle.com/datasets/lainguyn123/student-performance-factors