Score prediction based on hours of study.

José Antonio Bobadilla García A01734433, Tecnológico de Monterrey Campus Puebla

I. Introduction

It has always been said that the more hours of study dedicated to an exam, the better grade will be obtained, and to a certain extent this is true, but some prediction can be made depending on how many hours of study are dedicated to a specific subject. ?.

Performing a linear regression model could help us complete this prediction to find out if it is true or false that depending on the hours you study, there are more chances of getting a high grade.

II. Data set

The data set is structured as a CSV that contains the information of the students' study hours along with the grades obtained. The dataset has approximately 110 values which will help us to predict the grade that the student will get depending on the hours of study that they invest.

III. Model Proposal

The structure of the model that is planned to be used is a linear regression with 2 variables using the Python library Scikit Learn to make said prediction. The independent variable would be the number of hours that the students invest and the dependent variable would be the grade obtained.

IV. Test and Validation

A reshape was applied to the dependent and independent variables of the numpy library which resizes the arrays in arrays so that the model can work correctly.

For the test, 50% of the data was chosen for training and the scikit function learn train_test_split was used, which divides our dataset into the following variables taking into account the entered configuration:

X_train
X_test
y_train
y_test

Once this was done, a linear regression object from the scikit learn library was instantiated to be able to perform the training with the established data.

V. Plots

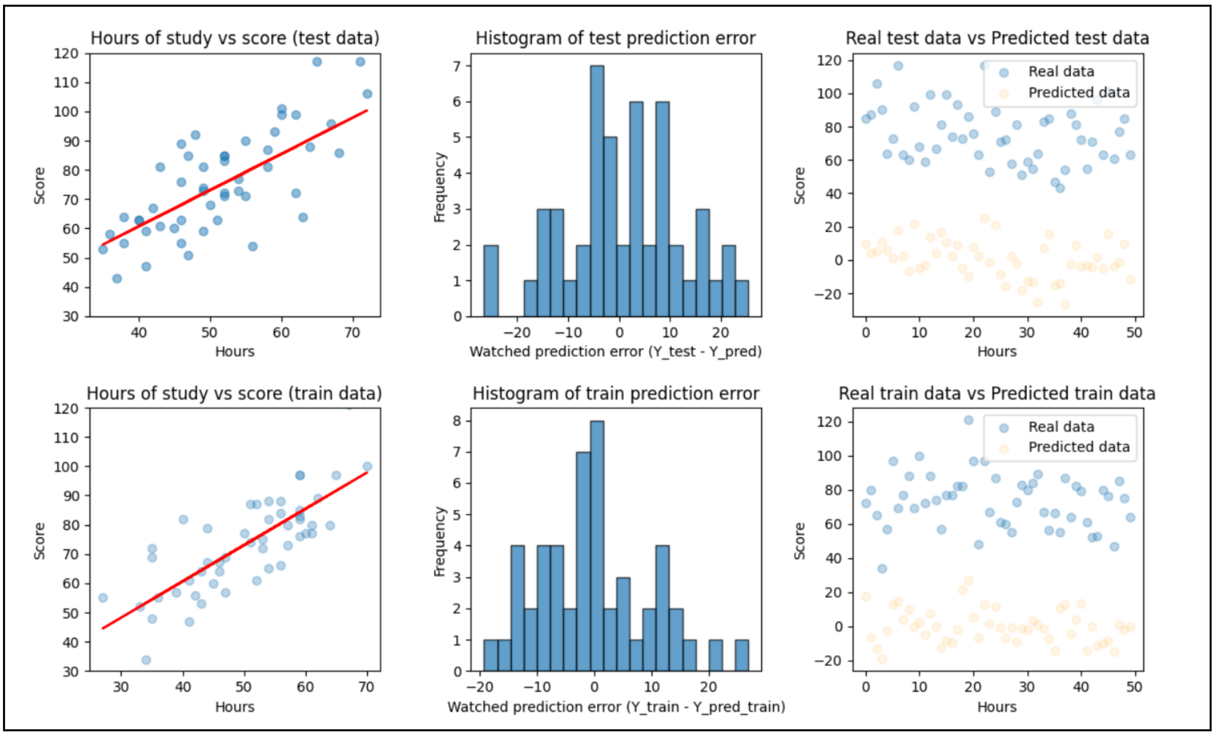
Once the predictions were obtained, the following graphs as follows:

The calculation of the prediction error was carried out to identify the bias of the model and we can observe that for testing histogram there is a more disorganized data structure compared to that of train in which we can see the typical "bell" structure. This means that the model has a high bias.

for the variance we can observe the histogram data are more grouped compared to the test data, which are more dispersed in the map. Also the width of the histogram tells us the variance that this has.

Also we can see that the distance between the true value is slightly to the right on the test prediction histogram, so a large distance means a larger bias.

Although the shapes are similar, we can see that the data of the train is slightly placed upwards.


Hours of study vs score (test data)

Histogram of test prediction error

Real test data vs Predicted test data

Hours of study vs score (train data)

Histogram of train prediction error

Real train data vs Predicted train data