Análisis de rendimiento regresión lineal

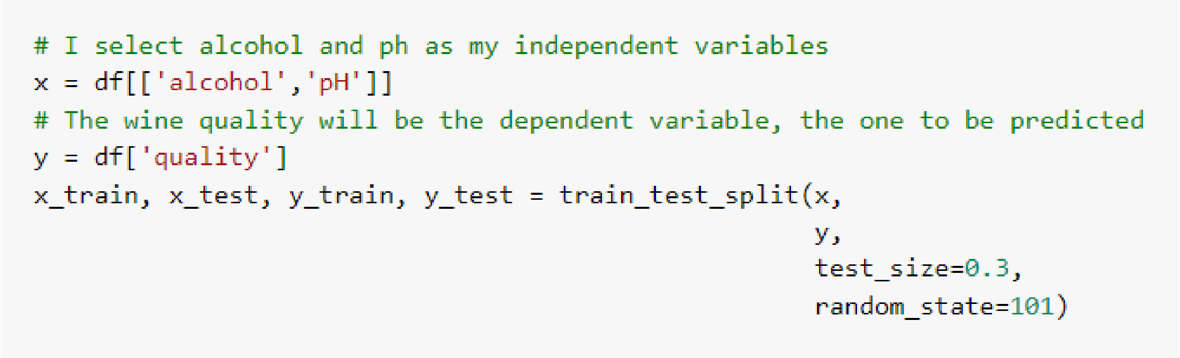
Carolina Herrera Martínez – A01411547

Campus Qro. 2022

This is the model that I create at fisrt. it predicts the value of the wine quality based on the ph and the alcohol ammounts-

To create this model, I divided the dataset into two different sets. the first one is for training the model, and the second one is for testing.

this division helps to identify different situations, such as underfitting and overfitting.



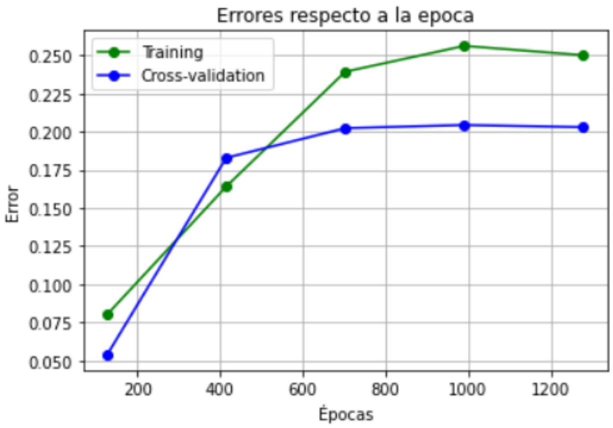
```
# I select alcohol and ph as my independent variables
x = df[['alcohol','pH']]
# The wine quality will be the dependent variable, the one to be predicted
y = df['quality']
x_train, x_test, y_train, y_test = train_test_split(x,
                                                    y,
                                                    test_size=0.3,
                                                    random_state=101)
```

Here we can observe that we have found a model that has a relatively low error.

The difference between both errors is not so big, but we can see that we have a 20% difference between the train error and the test one. This can indicate that we have a case of light overfitting, as the model could be memorizing the values from the training dataset.

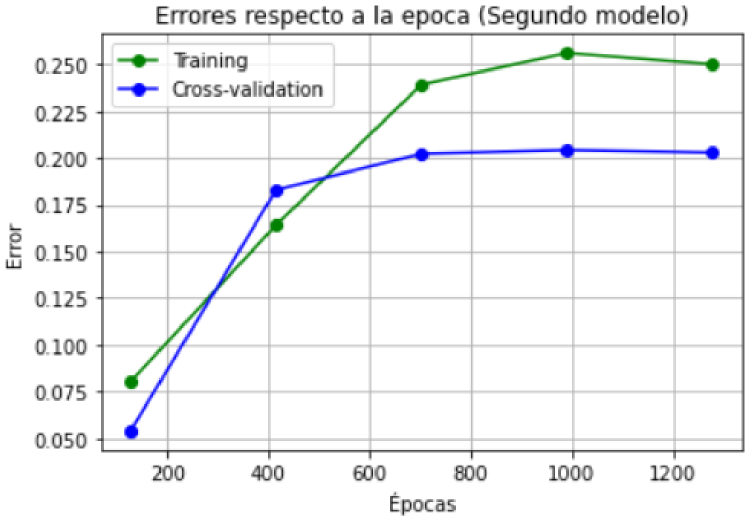
This also indicates that we have a medium strength bias as the model has adjusted itself to match the values from training, having a displacement when comaring them to the testing data

Here we can see that the distance between both errors is not that big. However, we want it to be as small as possible, so we will adjust our model to avoid this overfitting.



This second model has a better distribution of the dataset. This is considered as an adjustment to an hyper parameter. Now we will get a different modelation with slightly different coeficcients and intercept.

Now we can see that the difference between both errors has been reduced significantly. This means that we have solved our overfitting issue, and it also tells us that we have a case with a low bias and a low variance. being the low bias observed with the small change in the error when modifing the dataset, and the low variance is observed with the small magnitude of the error.



Now we can see that the difference between the training errors and the validation ones are also smaller. Something interesting to observe is that we find that the errors are almost the same in the epoch #400, and in epoch 500 they are equal.

This could also be an indicator of a slight remainance of overfitting that could be solved by delimiting the epochs to 500, so this could be a future improvement to be implemented.