Machine learning phase 1

preprocessing techniques :

1-first we apply on the string dataset the function feature encoder to make it measurable

2-we replace the missing values with median value

3-and we apply the scaling with the function feature scaling

4-then we apply the feature selection with the correlation method

5- and finally we split the data to training data (80%) and testing data (20%)

analysis on the dataset:

1-these features (overall rating, potential, wage, reactions, club rating, release clause euro, international reputation) have effect on each other and also on the value

regression techniques:

we used the polynomial and multiple linear regression

differences between the two models:

1-the MSE for polynomial model is 426351315339.7569 and accuracy is 98.5% and training time is 0.15s

2-the MSE for multiple linear regression is 1106620339532.842 and accuracy is 96.1% and training time is 0.32s

So the differences between them are the accuracy and MSE and training time, polynomial model is more accurate and has the least MSE and the least training time so it is more efficient

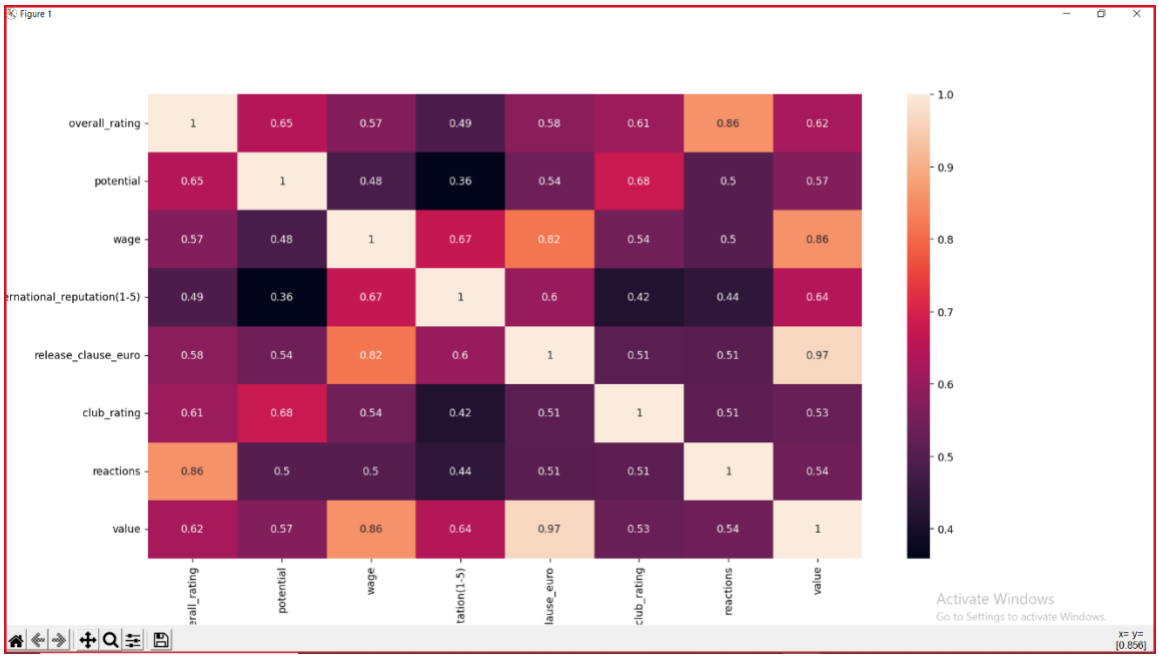
features we discarded :

we discarded the following features : (id, name, full name, birth date, nationality)

the sizes of our training, testing set :

20% for the testing set and 80% for the training set

screenshot of the feature selection using correlation :



conclusion about this phase of the project and what intuition we had about our problem and how it was proved/disproved :

1- we faced a problem in position column because it has one or more values, so we split this column into other columns (maximum 4 columns) then we applied one hot encoding

2- the columns that have “+” sign we split them with the function (str.split) and discard what comes after the “+” sign

3- and in the date if we found any missing data, we replace it with 1/1/2020 and we have to split the date with “/” sign to make it a numeric value