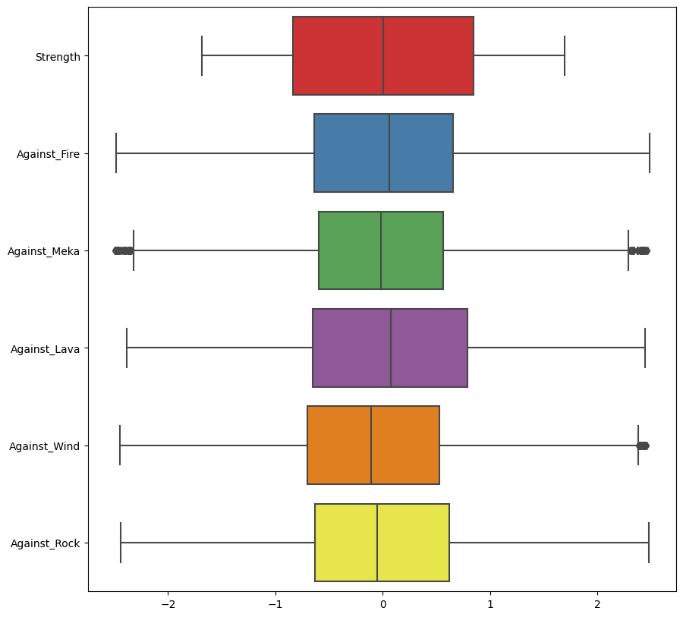
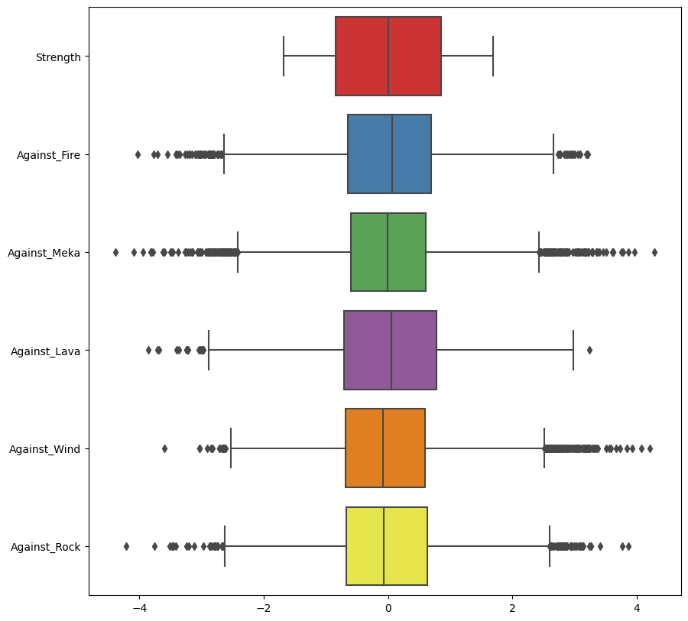
Supervised Learning Report

Final Project

Gabriele Passoni

Exploration and pre-processing

Immagine che contiene testo

Descrizione generata automaticamenteThe dataset consists of 16 columns and 10000 samples in total. There are 2 columns that are not compatible with our project. The first one is “Unnamed: 0” and is a duplicate of the index. The second one is an ID column that is useless because we want to predict the gormiti land based on features and not on an id. 6 of the 14 features remained are numerical while the other 9 are categorical. There are no duplicates or missing values in the dataset. The heatmap and the next plot show us that there is no correlation between the numerical features except in a few cases where the correlation remains low. Since the distribution of the features is Gaussian, I was able to apply a standard scaler to the entire data frame so that we have the same scaling of the features while maintaining their distribution. This procedure is not required by some of the algorithms used later but it does not compromise their results either, so I use it at this stage to avoid rewriting the same procedure for each algorithm. Next, I handled the outliers. Given the size of the dataset (high) and the number of outliers I decided to drop them. The parameter k used for the procedure was decided based on the results of the graphs. I tried to drop as few as possible while remaining effective and cleaning up unnecessary values. In order to apply the required algorithms, I had to use some encoding functions to change the feature values. The target (“Gormiti\_Land”) and two other features were encoded using a label encoder since they had to remain one-dimensional. For the other features, however, I applied one-hot encoding thanks to the “pd.get\_dummies” function. As a last thing in this chapter, I split the dataset (X and y) so that, for all algorithms, the same data would be used and the results would be unbiased.

Algorithm implementation