האקטון 2022, משימה שנייה

מגישים:

בר נמש: 312162464. אביתר בן: 312245087. יובל עומר: 316514314.

Preprocessing:

* The first challenge we encountered when working on the challenge were the medical terms

And annotations which are prevalent in the data.

To face this challenge, we did extensive research on the internet.

Furthermore, at first glance it was noticeable that a lot of features were Contain very basic errors in the way they were inserted, but it was possible to understand most of these errors and confidently restore the original value.

In the same way, we have put a lot of effort into thinking for each feature - in which way to treat its values, Is it in creating categories? Is the right way to create categories with order? How to deal with features whose values are dates? Are there new features that we can create from existing features? Which features seem to be of great importance and which are identified as noise? What is the connection between the different features?

Some examples:

As for the KI67 protein we decided to categorize the values into a group of five categorical values- from low value to high value- when in the values of this feature it was necessary to carefully and creatively extract the value in order to catalog it correctly (the values in this feature were given in various format, and it was noticeable that, that feature of great importance and we did not want to lose this important information).

Regarding dates features we conclude that the right way to deal with them is to calculate the difference between them – days amount between diagnosis and surgery for example.

* The data is made of the following features:
* Dates, patient's age, tumor size,
* We tried various methods that emits different results:

Different hyperparameters in different classifiers , such as Random Forest, Decision Tree,

KNN. In addition to that each classifier was observed in different multiclass classifier, such as, Power Set, Chain Model, and even Binary model.

* We carefully learned our data, and take extra care to learn it without overfit the data.

In order to do that we split our data to train and test parts, and we learned our data without peak at the test part until the very end.

In order to understand about the variance and bias that exists in our model we resampled our train data using the cross-validation method.

* We ended up using the Decision Tree classifier as it was the one that emits the better results.