# Huji IML Hackathon

## Challenge #2: Detecting Attributes of Breast Cancer

*Conclusions by Matan Hirschhorn, Ofri Baruch and Tzlil Ovadia*

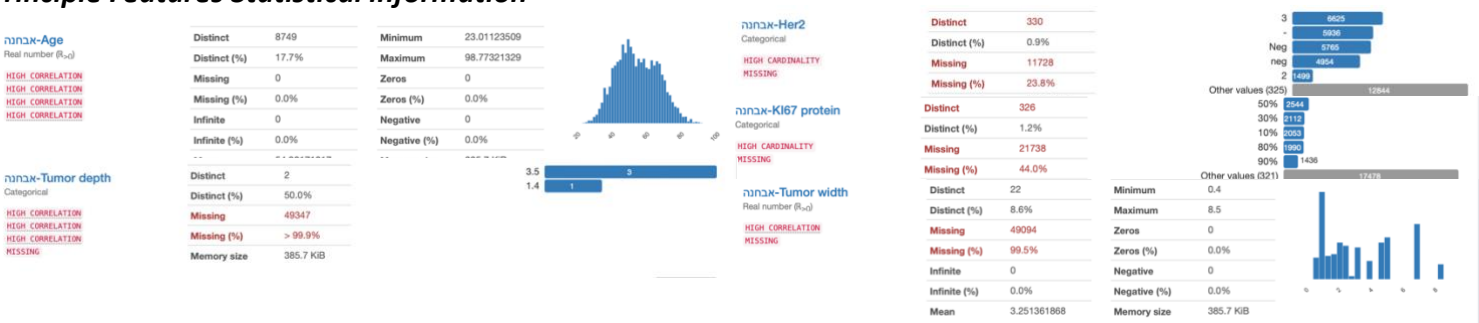
**Data Preprocessing and Data Analysis**

On this step, we tried various ways of processing the raw data, trying to deal with missing values, and deciding whether to drop the missing samples, or fill it with an alternative representative value. In addition, we tried to find values with the Highest correlations with the labeled data and getting rid of the features which are highly correlated with other features in order to reduce data redundancies.

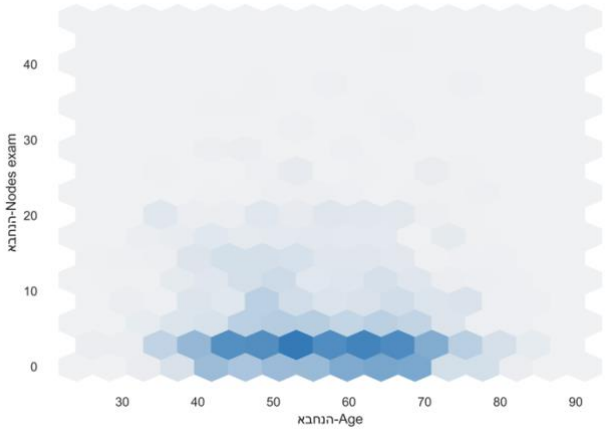
Another **major** challenge we found along the way, was the ultra-high variance found in relevant features, i.e., human typed-based feature fields such as ***HER2, KI67 Protein*** etc. We chose to use Regex based solution in order to parse the data and extract out of it the principal data components (e.g., positive diagnosis, percentage rate etc.). This task was a major time consumer.

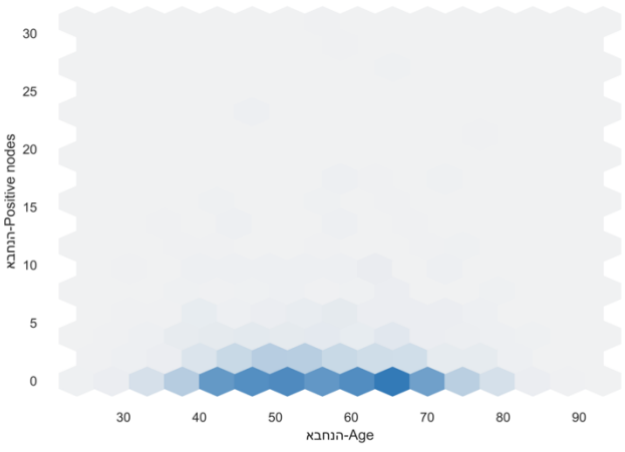
Our greatest mistake was to start with relatively big number of features, we thought could be relevant to model's predictions, and the fact we were already time-invested due to the work done on the preprocessing part, didn't make it any easier for us.
This gave us relatively long running time, and results that just couldn't justify it. We tried to use **PCA** and **Regularization** to improve the results of the test set. Unfortunately, the efforts were in vain.

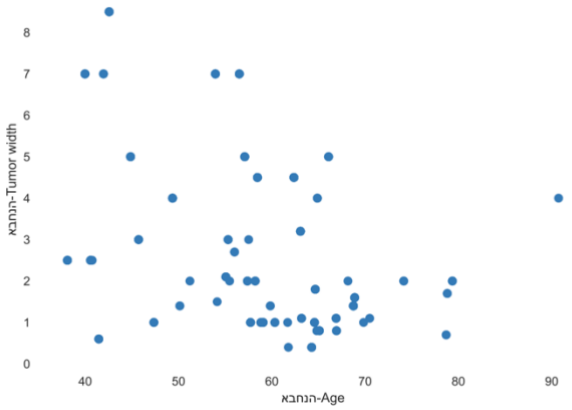
***Principle Features Statistical information***

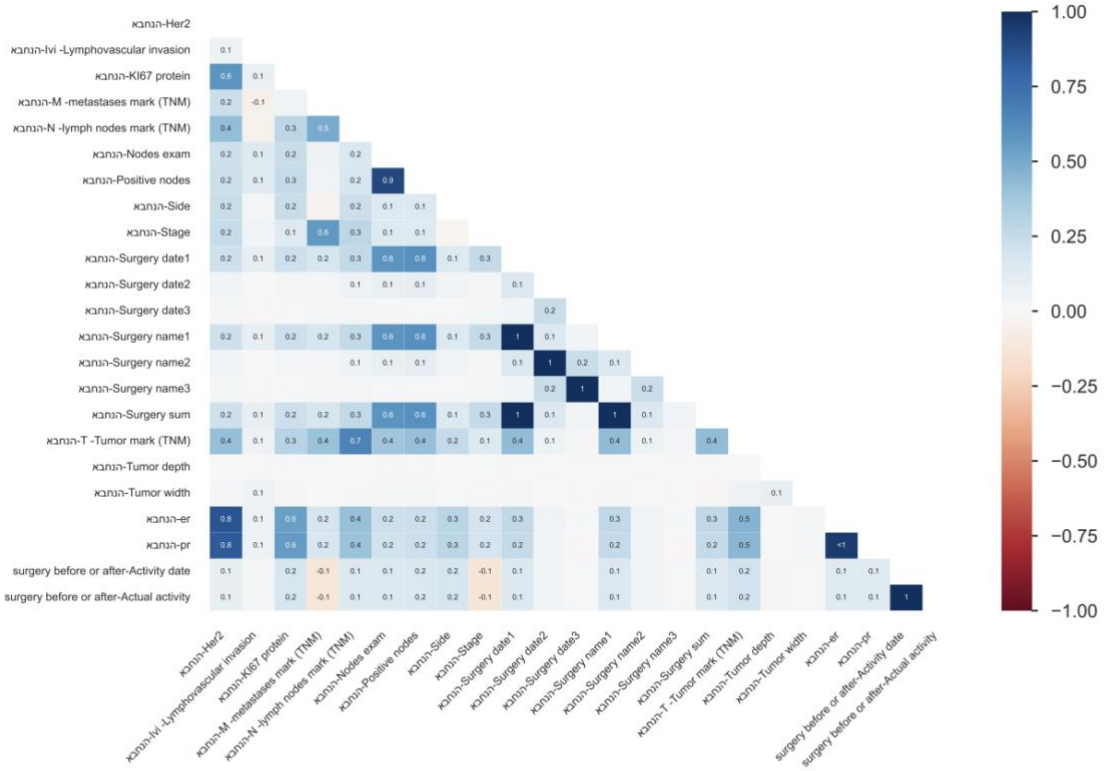



Examples for some of the visuals we generated in order to analyze the interactions between the different features given in the dataset, and below we can also find correlation heatmap between all features

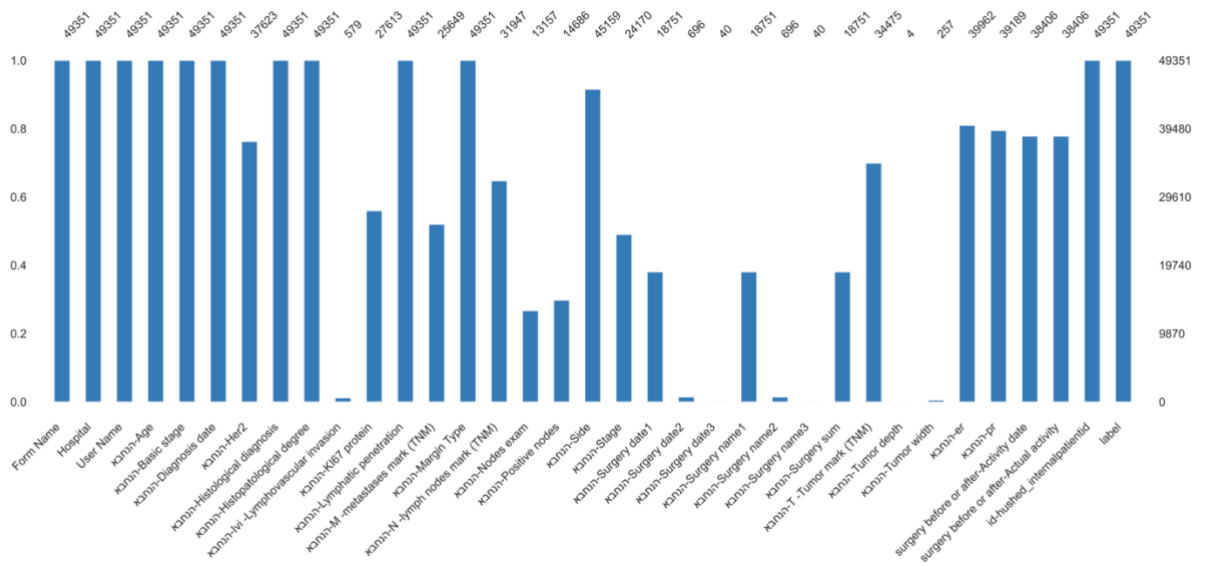





The correlation heatmap measures nullity correlation: how strongly the presence or absence of one variable affects the presence of another.




A simple visualization of nullity by column.

**Part I – Random Forest Classifier**

As for the first task, due to the nature of this problem, as complex as it be, we chose to use a Random Forest Classifier type, in order to classify each class. In addition to the data preprocessing, we realized that the labels data frame demanded some processing since we need to divide it into all of the different unique labels one can get.

Finally, after some wandering around, and some trials which lead us nowhere, we gave up and chose to get rid of many features, ending up with handful of features – ***Age, Tumor Depth, Tumor Width and HER2***.
Thanks to this brave decision, we achieved $F_{1_{MICRO}} \approx 97\%, \quad F_{2_{MACRO}} \approx 95\%$.

Another major challenge we had was the way of reconstructing the labels information in order to be able to compare it with the unknown labels.

**Part II – Random Forest Regressor**

This part was using continuous form of labels. Hence, we've changed to different type of model - Random Forest Regressor. Other than that, we haven't made further changes or any other data processing, thanks to the thorough analysis made prior.
This model scored value of $MSE \approx 0.156$ on the test set.

**Conclusions**
We learned a lot from this problem set, starting from the strategy planning for the project, through the data preprocessing stage (which was tedious), and more important is not to exaggerate with the model's complexity – start with simple, thin (but not too thin) model, and only afterwards add bits of complexity, brick by brick.