Hackathon 2022

**Data Description**

The data for the breast cancer challenge contains 47 thousand entries with information about breast cancer patients. The original data contains 31 categorical, numeric, and dated features, most of which are medical. For example, the number of surgeries a patient had, the stage of cancer, and the margin of the tumor. These are only a few of the information we had to get familiar with to best suit the features to the model. In addition, most rows in the data were duplicates of patients, while only about 8,000 patients are unique.

**Data cleaning process and Challenges**

The first challenge we had to deal with was filling Na values. For each column we had to think of the meaning of the missing value – is it a human mistake? Is it 0? Or does the lack of data indicates perhaps something we could include as a feature? We made different decisions per feature according to the data and its meaning.

The next challenge was cleaning bad entries of data by the doctors, and deciding how to use the categorical features in our model.

Finally, we dealt with the duplication of the rows. We checked for the reason for the duplication and found that multiple rows of the same patient varied only in the doctor that submitted the information, and the Clinique visited. We created dummy variables for these features and joined the information into a single row per patient.

Our learning system

Results

Suggestions

Lymphatic….

T tumor Marks – there are blanks

Er / pr are only negatives and unknowns

Cat boost

1. יצירת הסביבה הווירטואלית וייצוא קובץ requirements להגשה. אפשר לעשות את זה עם ה requirements של IML.
2. README.txt — contains a file list and a brief description of each file
3. USERS.txt — contains the logins and IDs of all the team members. Use one line per student, in the format login, ID
4. project.pdf - the description of your work.