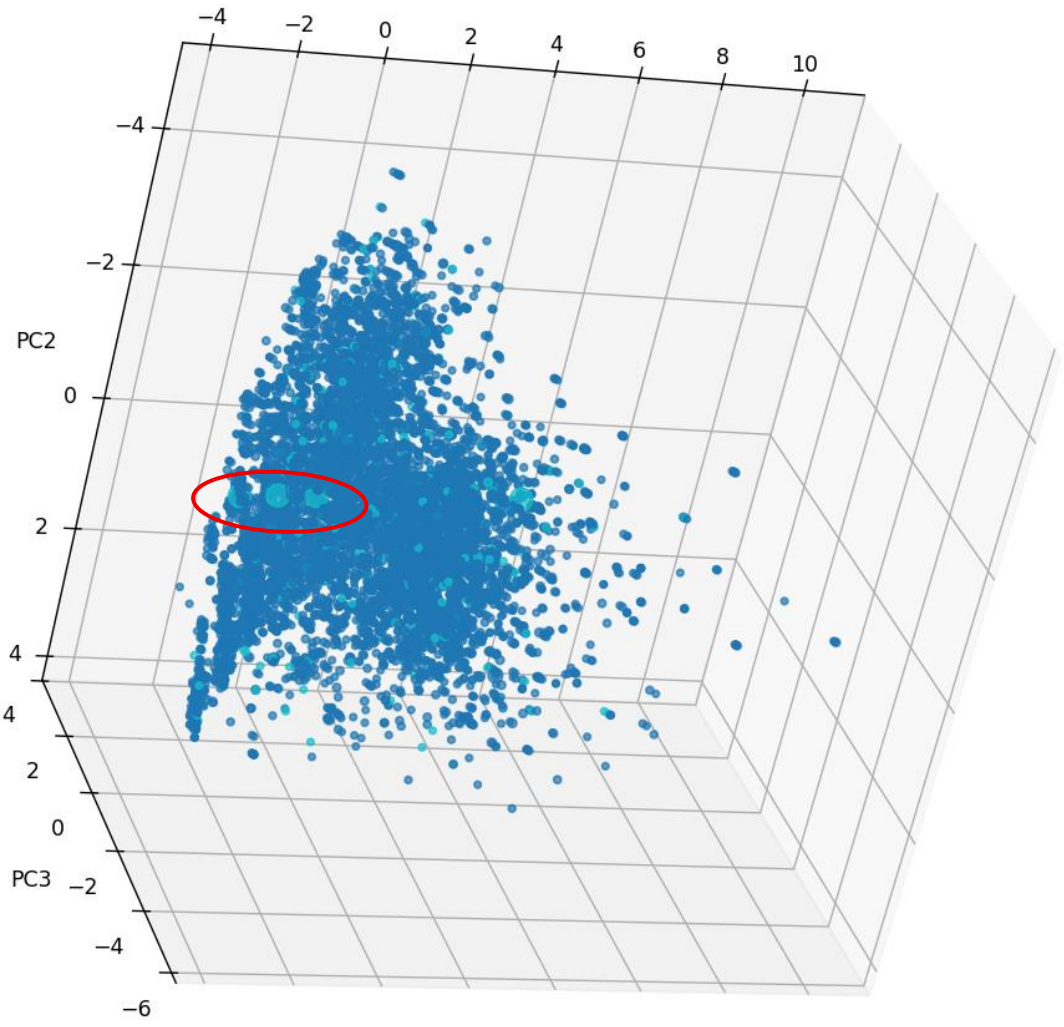
IML – Hackathon

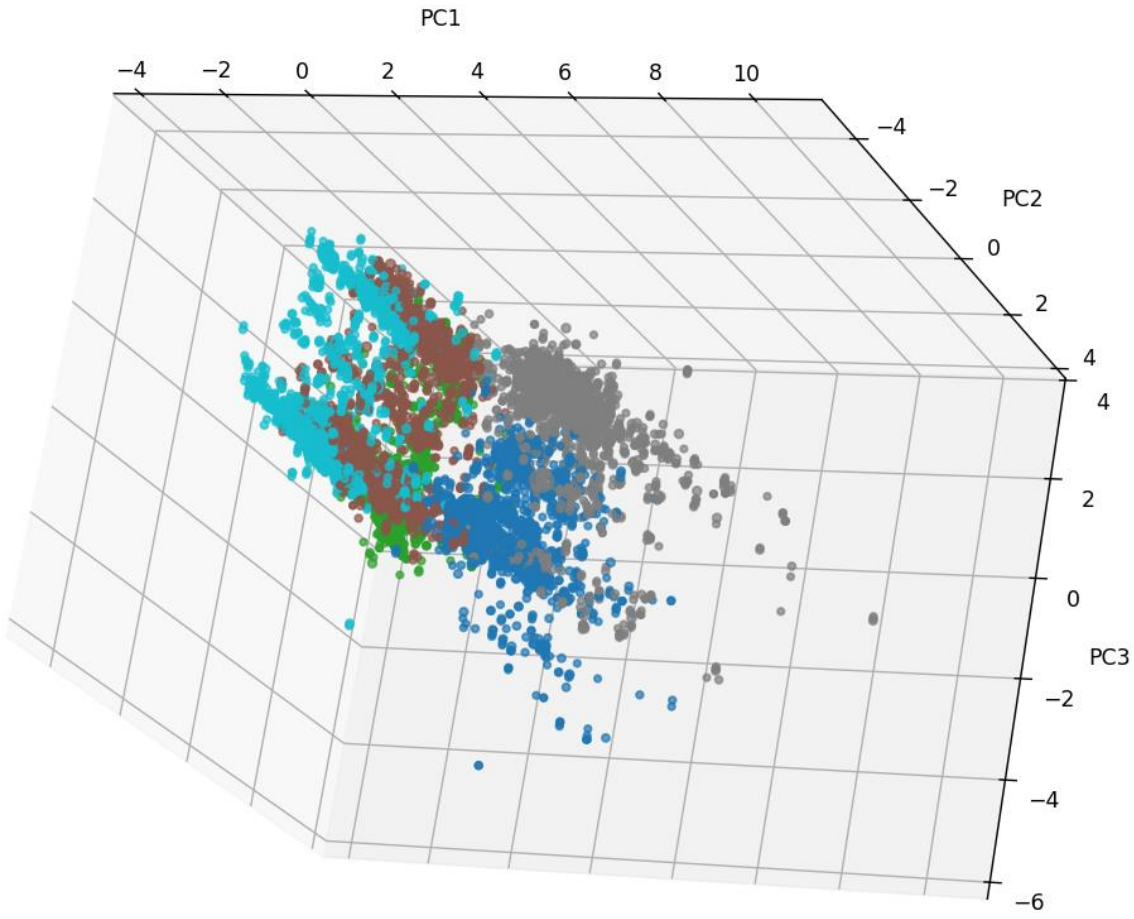
Part 3

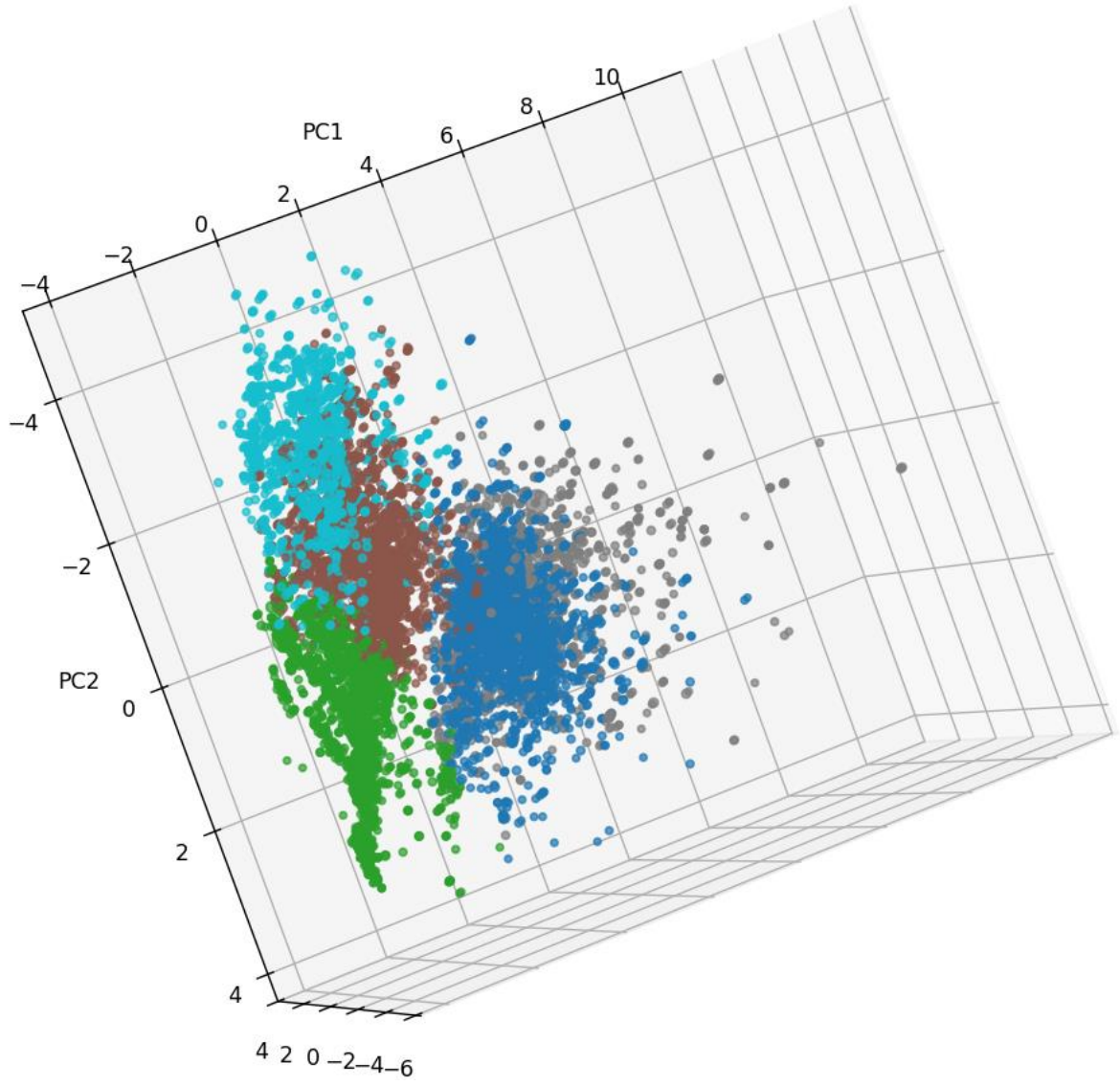
- We created a 3D PCA graph, where each dot has a size proportional to the size of the tumor, and colored according to the existence of metastases (a base size was added so points representing samples without tumors are still visible).
-

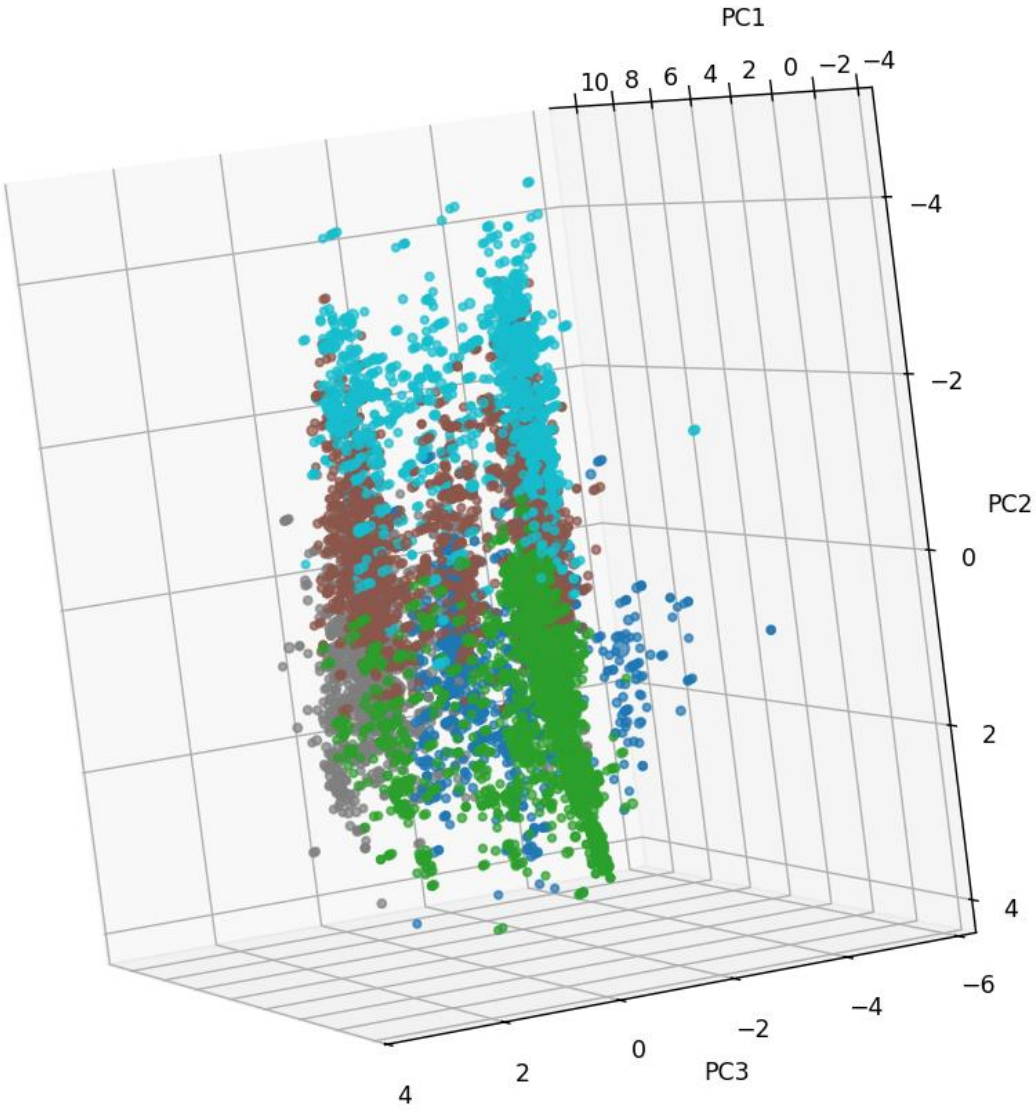


As we can notice in the graph, there are a number of samples that are relatively large and light blue – They represent samples with large tumor size and metastases. They are in close proximity, and located in the negative side relative to the PC2 component, and relatively centered around the other axis's.
From Analyzing the entries of the PC2 vector, we noticed the largest values are matching to features "Surgery Sum" and "PR sensitivity". A low sensitivity to PR doesn't allow the treatment with progesterone hormone. Additionally, low "Surgery Sum" suggests the patient had low number of surgeries. Both factors might contribute to the increase of tumor size, and of metastases.

- We created the same PCA graph, while coloring dots according to a k=6 KMeans clustering:


PC1
−4
−2
0
2
4
6
8
10
PC2
−4
−2
0
2
4
PC3
4
2
0
−2
−4
−6


PC1
10
8
6
4
2
0
−2
−4
PC2
−4
−2
0
2
4
4 2 0 −2 −4 −6



As we can see in the graph, the data has a clear partitioning into different group in the 3d space which matches the resulting clustering.

The partitioning into clear separate groups might help to classify groups of patients with similar features, in order to deduce some properties of the tumor on all of the group, after performing some observation on some of it.

We believe some of the clustering derives from the way we performed preprocessing – many verbal features were converted into numerical values, creating discrete gaps along the matching axis's. Therefore, we could expect the dots to be spaced in constant gaps along the matching axis, and also pass similar property to the lower dimension representation.

- The first principal components largest entries are:

אבחנה-Surgery sum	0.412138
Surgery names score	0.393951
אבחנה-Nodes exam	0.305096
אבחנה-Histopatological degree	0.294269
אבחנה-T -Tumor mark (TNM)	0.276552

אבחנה-N -lymph nodes mark (TNM)	0.274679
אבחנה-M -metastases mark (TNM)	-0.256323
אבחנה-Positive nodes	0.233720
אבחנה-er	0.231783
אבחנה-pr	0.215734
אבחנה-Margin Type	-0.201616
aggressiveness_by_activity_timing	-0.201604
אבחנה-Side	0.130982
אבחנה-KI67 protein	0.090997
אבחנה-Basic stage	0.075793
surgery before or after-Actual activity	0.060038
אבחנה-Lymphatic penetration	0.058314
אבחנה-Histological diagnosis	-0.038631
Surgery date average wait	0.024877
אבחנה-Stage	0.022133
אבחנה-Age	0.019486
Form Name	-0.012657
אבחנה-Her2	0.011207

Where 'surgery sum' is the number of surgeries the patient had, and 'surgery names score' is a column we added in preprocessing that accounts for the different procedures the patient had (as a numerical grade of the total severity of the procedures). Therefore, the principal component patient had multiple surgeries, with a relative high score of severity (the procedures she experienced where relatively more advanced).