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 many features contain basic errors such as typos, but most errors were recoverable and we managed to turn them into useful data.

Similarly, we spent a lot of time thinking about each feature - in which way to process its values, Is it categorical? Is it ordered? 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 dispensable? What is the correlation[[1]](#footnote-2) between the different features?

Some examples:   
One of the features we encountered is the KI67 protein which is a protein that effects cell growth and proliferation, and many studies show it has a direct connection to cancer growth.  
Which led us to decided to categorize the values into a group of five categorical values- from low value to high value, this process required us to extract the correct value from a given input in spite of the many formats the input can be given.

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.

Figure 1: Correlation between features

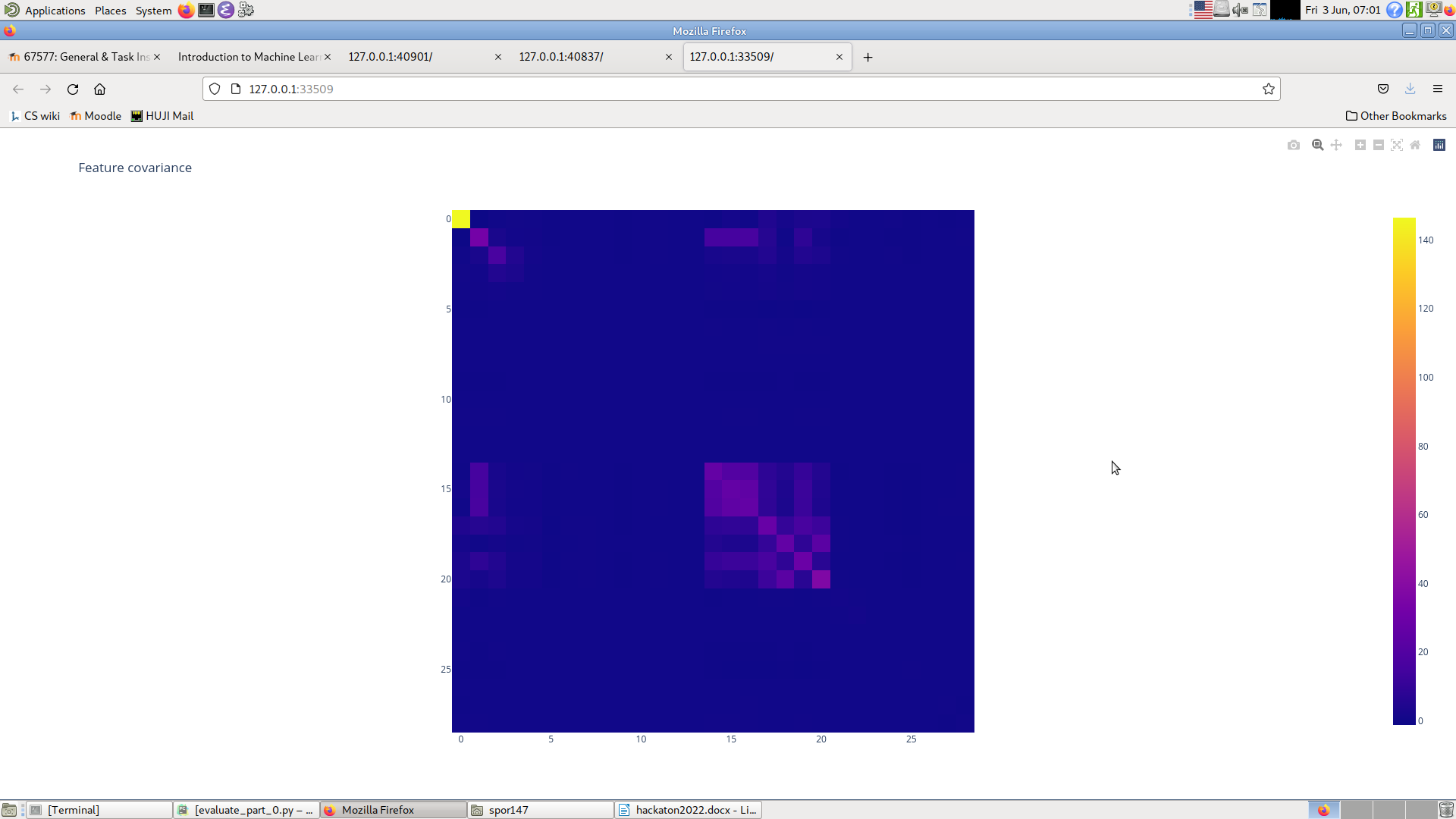


Figure 2: Kmeans clustering with age as color

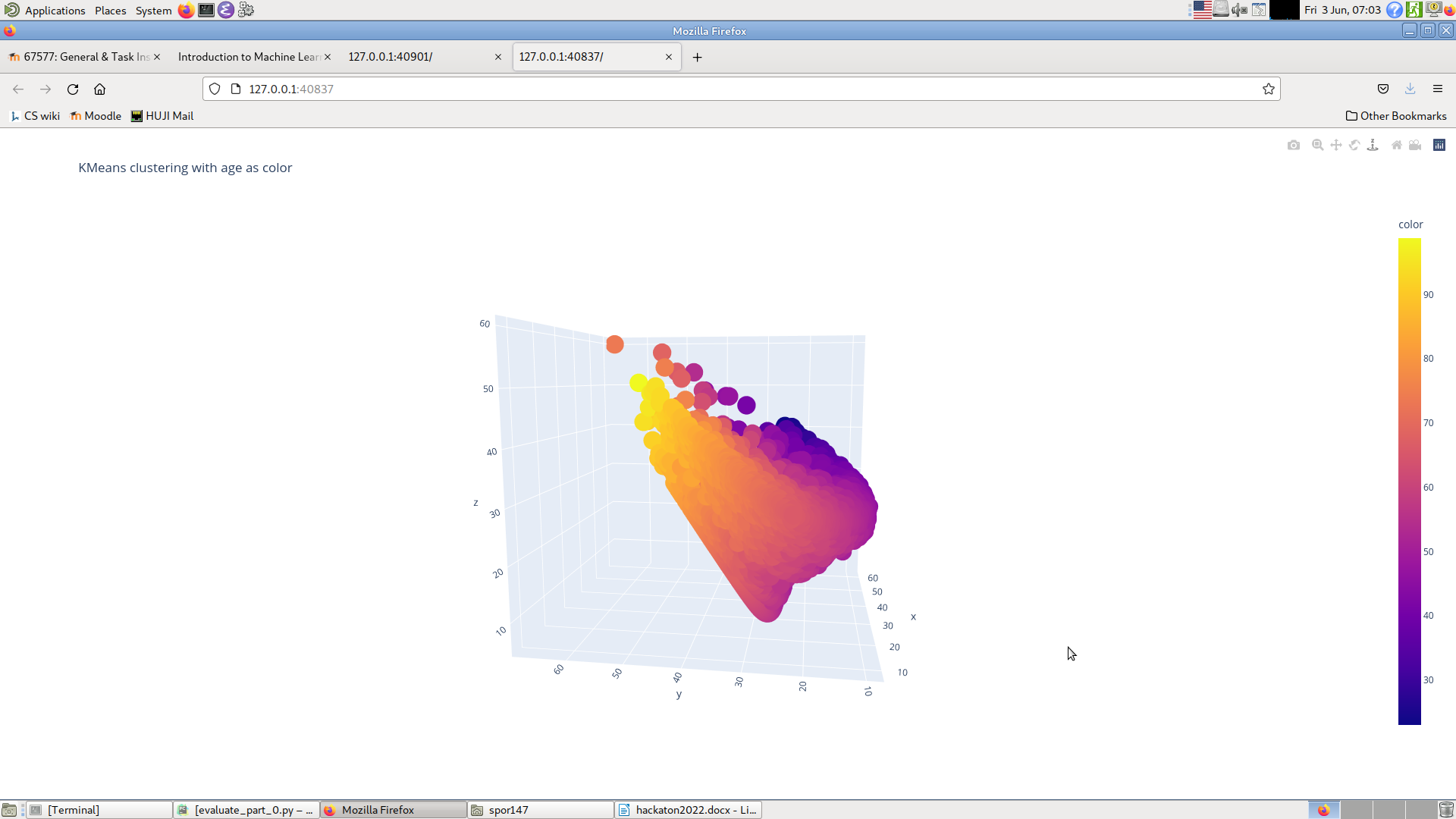
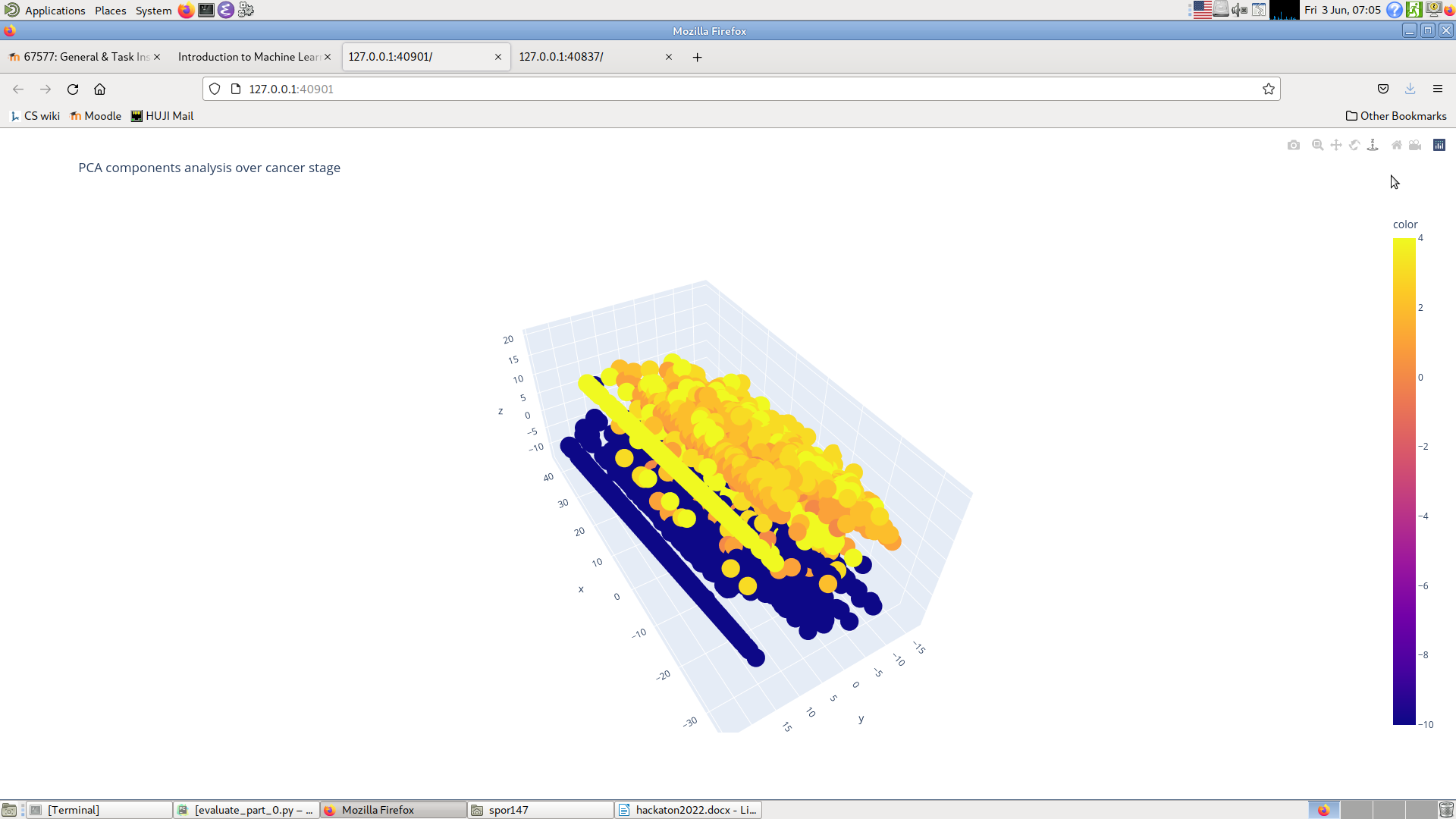


Figure 3: PCA components analysis over cancer stage



1. See figure 1 [↑](#footnote-ref-2)