Hackathon - Math & Stats

1.) From the data that Dr. Panchevski collected, he wants to conclude whether a specific medicine given to patients helps recovery. The patients include in the trial have an elevated blood pressure (BP) (data represents only the systolic BP) and it is assumed that the medicine helps patients stabilize their elevated BP.

Data is collected for 100 patients, and are given in the file named “Prva_zadaca.csv”.

The column ‘Merenje 1’ (english ‘Measure 1) represents measured BP on the first day of the trial.

The column ‘Merenje 2’ (english ‘Measure 2) represents measured BP after 6 months in the trial.

The column ‘Primil lek ili ne’ (english ‘Has medicine administered or not’) is categorical (0 or 1), i.e. it tells us whether a patient has taken the medicine, or not, for lowering the BP.

a) Decide which way you will visualize the data and write comments about the data (short text and comments for the plots).

b) Categorize the data and perform Descriptive Statistics for the variables that are gained.

c) Perform the following tests for the groups:

c-1) Conclude whether there is a significant difference between BP measured on the first day and BP measured on the last day.

c-2) Conclude whether there is a significant reduction (or increase) of the BP, in patients that have taken the medicine, between the first measure and after 6 months.

c-3) Conclude whether there is a significant reduction (or increase) of the BP, in patients that have not taken the medicine, between the first measure and after 6 months.

c-4) Is there significant difference in BP between patients that have taken the medicine and patients that have not taken the medicine?

2.) Dr. Panchevski received a new set of data after the first trial. The parameters of the trial were the same, however, the data for whether a patient has taken a medicine in the trial or not was lost.

a) Using the data in ‘Prva_zadaca.csv’, construct (train) a classification algorithm for the data. Categorize the patients in 2 groups - 0 (did not receive medicine) and 1 (received medicine). Then, for the new data in ‘Vtora_zadaca.csv’:

> a-1) Categorize patients only with respect to the first measurement.
>
> a-2) Categorize patients only with respect to the second measurement.
>
> a-3) Categorize patients with respect to the first and second measurement.
>
> a-4) For the categorizations in a-1), a-2) and a-3), compare the precision of the Models, comment (short textual description) and visualize the data (2D / 3D - in Python, write comments for the plots).

b) **Bonus:** Is it possible to categorize the data using linear regression? Would the model be a good fit? Assuming yes, which columns should be used to train the model in this case (on the data in ‘Prva_zadaca.csv’? Construct a linear regression for the categorization tasks for patients in ‘Vtora_zadaca.csv’. What is R^2? Can this score be improved? What does the score depend on in this scenario?