# Хакатон: математика и статистика

## I Материјали

- Документ со формулација на задолженијата (овој документ)
- Два `csv` фајлови со податоци

## II Задолженија

- Проверете дали на располагање ги имате сите материјали;
- Во целост прочитајте ја содржината на документот со формулација на задолженијата. Доколку има потреба нешто да се појасни, слободно прашајте;
- Решете ги задачите кои се формулирани подолу. Се очекува решението да содржи и објаснувања/заклучоци на некои места (некаде е експлицитно побарано, некаде не е), па не ги заборавајте. Нема *единствено* решение на задачите;
- Покрај решенијата, на/до крајот на хакатонот треба да имате соодветна презентација на истите (може и во самиот *notebook*, не мора да е *PowerPoint* или слично).

## III Распоред на активностите

Хакатонот се одвива во текот на три дена од кои два се активна работа, а третиот ден е презентацијата на сработеното.

- Ден 1: сабота, 18 јуни 2022
  - 9:00-9:30 | Отворање на хакатонот и запознавање со задолженијата
  - 9:30-12:00 | Работа во групи **без ментор**
  - 12:00-14:00 | Консултации и работа **со ментор**
- Ден 2: недела, 19 јуни 2022
  - 10:00-10:20 | Отворање на вториот ден, одговарање прашања, споделување искуства
  - 10:20-14:00 | Работа во групи **без ментор**
  - 14:00-16:00 | Консултации и работа **со ментор**
- Ден 3: понеделник, 20 јуни 2022
  - 17:00-18:00 | Подготовки за презентација
  - 18:00-20:00 | Презентации

# IV Задачи

## Задача 1.

Еден доктор собирал податоци за ефикасноста на лек за намалување на крвниот притисок кај неговите пациенти. За оваа цел докторот по случаен избор одбрал пациенти и ја следел состојбата со нивниот крвен притисок. Сите пациенти во истражувањето имаат покачен притисок (се разгледува само „горниот“, *систолички* притисок), а се претпоставува дека лекот позитивно влијае на намалување на притисокот.

Податоците за пациентите се дадени во `Prva_zadaca.csv`.

Колоната `Merenje 1` содржи податоци за измерениот притисок на првиот ден на истражувањето.

Колоната `Merenje 2` содржи податоци за измерениот притисок шест месеци после првиот ден.

Колоната `Primil lek ili ne` содржи податоци за тоа дали пациентот примил лек или не (1 ако примил, 0 ако не примил). Оваа променлива е од категориски тип.

а) Изберете соодветен начин за визуелизација на дадените податоци и напишете кратки коментари за добиените графици.

б) Категоризирајте ги податоците и направете основна статистичка анализа за добиените променливи (*descriptive statistics analysis*).

в) Со помош на соодветни статистички тестови, изведете заклучок дали:

i) постои значајна разлика помеѓу просечните нивоа на крвен притисок на пациентите на крајот на истражувањето споредено со почетокот;

ii) постои значајно намалување на просечното ниво на крвен притисок на крајот на истражувањето споредено со почетокот кај пациентите кои примиле лек;

iii) постои значајно намалување на просечното ниво на крвен притисок на крајот на истражувањето споредено со почетокот кај пациентите кои не примиле лек;

iv) постои значајна разлика во просечното ниво на крвен притисок на крајот на истражувањето помеѓу пациентите кои примиле лек и пациентите кои не примиле лек.

## Задача 2.

По завршувањето на првата фаза од истражувањето, докторот добил податоци за нова група пациенти на кои е направено истото истражување. Но, неговиот несовесен асистент заборавил да забележи кој пациент примал лек, а кој не и сега таа информација е загубена. Податоците се дадени во `Vtora_zadaca.csv`.

а) Со користење на податоците од задача 1 (дадени во `Prva_zadaca.csv`) конструирајте (истренирајте) алгоритам за класификација на податоците. Класифицирајте ги пациентите во две групи: 0 (не примил лек) и 1 (примил лек).

б) Откако ќе го конструирате алгоритмот во делот а), класифицирајте ги новите пациенти (чии податоци се дадени во `Vtora_zadaca.csv`) во според:

i) првото извршено мерење;

ii) второто извршено мерење;

iii) двете извршени мерења.

в) За класификациите што ги направивте во делот б), пресметајте ја прецизноста (*accuracy*), направете соодветни визуелизации и направете кратка споредба на основа на овие параметри.

## Бонус задача.

Дали податоците од задача 2 може да се класифицираат со помош на линеарна регресија?

Во случај на потврден одговор, како би се конструирал моделот и кои променливи би го сочинувале? Колку изнесуваат коефициентите на корелација $R$ и детерминација $R^2$ во овој случај? Колку изнесува прецизноста (*accuracy*, *model score*) на моделот? Дали овој резултат може да се подобри?

Во случај на одречен одговор, образложете зошто ова не е возможно.