Technical test

Machine learning

General questions :

1) What's the interest behind a **ROC curve** ? What is it used for ? Can you briefly explain its principle ?

2) Why would you normalize (**z-score**) your numerical features before using it to train your model ?

3) Can you cite and explain briefly one (or more) **recommendation system** ?

Coding questions (provide lines of codes and results) :

1) Say that you want to compare 2 versions of a website, the original one (**A**) and the new featured one (**B**). The aim of the test is to see if the version B has a better conversion rate than the version A.

After 1 month of testing, here are the results of your test :

→ **version A :** **4000** sessions, **200** conversions (which gives a conversion rate of **5%**).

→ **version B :** **4000** sessions, **260** conversions (which gives a conversion rate of **6.5%**).

Is the difference **statistically significant** ? Meaning : is the version B performing better than the version A essentially thanks to the new feature ? We will say that the difference is significant if the probability that version B is better (under the null hypothesis) is inferior to **5 %**. Prove your result by simulating the experience.

2) We gathered data from an e-commerce website (data from https://datafiniti.co/). You'll find a list of 71,044 reviews of 600 different products. Let's say that we have a user, named John who is interested in Food, Movies, Personal Care, Music, Book and Sport. Your mission is to determine, among the 600 products, which are the **best 3 products** that John could buy knowing that he cares a lot about product ratings.

The base code, and the database is available to download here : https://github.com/JeanSavary/mileON-technical-test

There might be different interesting results depending on your approach. Be creative 😉

NB : Results of the analyses could be provided either on **Jupyter Notebook(s)** or **scripts + written report**.