Exercise Sheet 7 – Probability and Statistics

INF161 Autumn Semester 2021

- Deadline: 5/10/2021 , 23:59
- Format: Your answers are to be returned in a single pdf report. You can also return scanned pages for your calculations. For results, your answers must include any values that are requested in the Notebook.

Exercise 1 *We throw two fair dice and a coin at the same time. What is the probability that the sum of the numbers on the dice are* ***higher than 4 or*** *the coin lands on* ***Heads****?*

Exercise 2 *Table 1. shows the information for a movie recommendation system that includes the rating of users and Genre for each movie in the databse. Fill the missing values in the table using both methods of item-based and collaborative filtering.*

	Drama	Comedy
Money Heist	1	0
Mr. Robot	1	0
Rick and Morty	0	1
Brooklyn nine-nine	0	1
Peaky Blinders	1	0
After Life	0	1
BoJack Horseman	0	1

	Money Heist	Mr. Robot	Rick and Morty	Brooklyn nine-nine	Peaky Blinders	After Life	BoJack Horseman
Teodor Haraldson	6	9	8	8	5	6	8
Kamilla Hansen	8	8	**??**	9	5	4	3
Marthe Berge	7	5	1	8	**??**	5	3
Lene Beck	5	9	**??**	8	3	7	1
Ludvig Sæter	5	7	5	8	6	7	5
Emillie Rasmussen	3	2	4	9	7	2	**??**

Table 1: Ratings for each movie by each user

Exercise 3 *Write a python procedure that reads the content of the file traffic_rules.xlsx and generates, for each sheet in the excel file, a dataframe encoding the data. The sheet "rules ordinal" represents the rules in the sheet "rules" with only ordinal data, while the sheet "rules categorical" represents the rules in the sheet "rules" with only categorical data. Fit a linear regression model for each dataset (ordinal and categorical) that predicts if a vehicle should pass the crossing. With what data does the model perform better? Why?*

Exercise 4 *Create a Jupyter notebook to visualize the Coronavirus, with data from:*

`https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset`

For each visualization you create you should write down

- *Why you think this visualization is interesting*
- *What conclusions and insights can be drawn from the visualization*

You should at least add (use the `COVID19_line_list_data.csv` *file for these):*

- *A histogram over the age of the patients*
- *A plot showing the fraction of male and female patients*

Beyond that it is up to you how far you want to go. The sky is the limit here; there is no end to the number of ways to visualize this data. See the following page for some examples:

`https://informationisbeautiful.net/visualizations/covid-19-coronavirus-infographic-datap`

Exercise 5 ***(The (infamous) chocolate study)***

This question is inspired by a favorite study of mine, where the researchers intentionally made mistakes to highlight problems with how science is published. Your goal is to list the mistakes made by the researchers in conducting the study and drawing conclusions from the data. Information about the study follows.

Study goal: *Measure the health impact of eating chocolate.*

Study overview:

- *The study consists of* 20 *volunteer participants (*10 *men and* 10 *women). Each participant is paid a total of* 100*.*
- *At the start of the study, researchers take blood tests of all participants, measuring blood sugar, liver function, kidney function, cholesterol, and glucose.*
- *Next, all participants are instructed to eat* 200 *grams of chocolate per day for* 1 *month. The participants are all given the same kind of chocolate.*

- *At the end of the study, the researchers again take blood tests of all participants. Blood sugar, liver function, kidney function, and glucose are very similar to at the beginning of the study, so only cholesterol is considered for the study.*

Outcome:*The cholesterol level of the participants had, on average, im- proved at the end of the month compared to the test at the start of the study. The researchers determine that eating chocolate improved cholesterol with a certainty of* $> 95\%$.