# 01URZSM DATA ETHICS AND DATA PROTECTION

# COMPUTER SCIENCE PART (DATA ETHICS)

# Exam 16 September 2022

## A. CASE (7,5points).

**Predicting absenteeism at work**

The managers of a large chain of supermarkets in Italy want to contrast the absenteeism at work of the company's cashiers. They want to understand which factors are related to the most frequent absences. Given such a goal, they built a dataset merging the available information on the absences with the demographic data of the cashiers. The obtained dataset contains absences from 1st September 2015 to 1st September 2022. For each absence, the information in the list below is available: variables A, B and C are related to when the absence took place; variables from D to K are related to the person who made the absence; variable L is the target variable. Notice that holidays and sick days do not count as absences, but only work permits.

A) Day of the week (five possible values, from Monday to Friday)
B) Month (12 possible values, from January to December)
C) Season (4 possible values: winter, springtime, summer, autumn)
D) Distance from residence to work (number of kilometers)
E) Current salary (€)
F) Service time (number of years the person is at the company)
G) Education (3 possible values: bachelor degree, master degree, phd)
H) Son (number of children)
I) Smoker (yes/no)
J) Weight (kgs)
K) Height (centimeters)
L) Target variable: number of hours of absence

The company signed a consultancy contract with a data scientist, who is charged with the tasks of building a prediction model and giving advises on its strengths and limitations.

You are that data scientist who signed the consultancy contract.

**Answer concisely to the following questions (give a separate answer for each point). Clearly state your own hypotheses, and any other information that you suppose in addition to the provided information, in order to coherently support your line of reasoning.**

**1) Which measurements issues, model limitations or data collection issues do you report to the managers? Identify two issues (in total), and explain them. You don't have to suggest remedies. (4 p.)**

**2) Which risk of systematic discrimination towards a protected attribute do you identify and report to the managers? Identify only one and explain why. (3,5 p.)**

***

**B (7,5 points). Explain what the Campbell's law and the Goodhart's law assert (also known as reflexivity problem) (4p.), then make an example of real or hypothetical automated decision making system where they could apply (3,5p.)**

# Indications for a possible solution

## A. CASE (7,5points). Predicting absenteeism at work

The following comments focus on the interpretation of the case. It should be used as a guide and not as the only possible solution, which is not unique, and it might be partially dependent on the reasoning presented and the hypotheses made (if valid). The exam, given the available time, requires synthetic (but precise and logically coherent) answers, therefore the length of the analysis reported herein should not be taken as a reference.

1) Examples of

- measurement issues
  - A) Day of the week: large supermarkets can be open 7/7 or 6/7, this variable does not track weekends;
  - G) Education: low levels are excluded, as a consequence it is probable to have majority of missing values due to the common low level of education required for this job;

- model limitation
  - considering the organization of the dataset (each row is an absence), and the target variable "number of hours of absence", the model can predict the longest absences, but not the most frequent absences, which is the goal of the managers;
  - data of years 2020 and 2021 might be considered for a separate model because of probable highest absenteeism due to the difficulties of the hardest pandemic times;

- data collection limitations
  - the current organization of the dataset is by absence (each row is an absence), hence it might probably include very few data about people hired recently in the company;
  - the year is not tracked: absences in the same day and month of different years are not distinguished.

2) Risk of systematic discriminations:

First of all, it has to be considered that the work of cashiers is a typical example (also illustrated during the course) of highly skewed job by gender: thus, the database will be probably filled with more records related to women, ***under the hypothesis that the distribution of absences will reflect such imbalance***. In addition, women's child-care is usually predominant, that would imply ***more and longest absences for women***. These are the necessary premises to make a robust hypothesis on the ***correlation between the target variable and the protected attribute gender***.

With the premise of the hypothesis explained above, proxy variables for gender need to be identified and explained (remember: a **discrimination will occur only if the proxy variables are significant predictors of the target variable)**:

- the variable "current salary" is expected to be a proxy for gender, because of the salary gap between women and men;
- feature K (height) is expected to be highly correlated with gender: globally, the mean height of women is shorter than that of men.
- feature J (weight) is expected to be slightly correlated with gender: on average it is higher for men.

A specific note concerns variable H ("Son", number of children): there is no rational to hypothesize that men cashiers (globally and in that supermarket) have significantly more/less children than women cashiers, hence this variable should not be considered as a proxy for gender. However, if we assume that on average the more the children the longest and numerous the absences are, such correlation with target variable might also reinforce the systematic discrimination towards gender explained above.

The exam has been designed by adopting the variables of an existing dataset on absenteeism at work: https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

**Most common errors**

Question 1:

- answering with discrimination issues: the question was not about that;
- as stated during the course and in several past exams' examples, just writing that a variable is useless is not an acceptable answer;
- no explanations given.

Question 2:

- it is not explained why the discrimination would occur (e.g. correlation with target variable);
- not referring to a protected attribute;
- answering with measurement issues: the question was not about that;
- no explanations given.

## B. Question on theory(7,5points)

Check slides and video-lectures.