DAT565/DIT407
Introduction to Data Science and AI

Assignment 8
Deadline: 2024-10-31 08:00

Before starting this assignment, you **must** read the article *Datasheets for Datasets* by Gebru et al. linked to from Module 8 on the course Canvas page. Also do the other recommended readings before you start if you haven't already.

I this assignment we will inspect and document a dataset available from kaggle.com. See description here:
`https://www.kaggle.com/datasets/fahadrehman07/hr-comma-sep-csv`.
The dataset is also available to download from our Canvas page.

Problem 1: Create a Datasheet

Download and inspect the dataset `HR_comma_sep.csv` from the course Canvas page (for instance by opening it in Python as a Pandas dataframe). Also read the information about the dataset on the above-mentioned Kaggle website.

Your task is to create (parts of) a datasheet for the above dataset. In your report clearly state both the question and your answer. Be brief but clear in your answers. None should be longer than a few sentences.

In your datasheet you should not answer all the questions from *Datasets for Datasheets*, as there are very many. You should however try to answer the following questions as best as you can:

- Motivation: 1 – 2.
- Composition: 5 – 6, 8 – 9, 15 – 19.
- Collection process: 26 – 29.
- Uses: 40 - 41

Problem 2: Ethics

Based on the potential ethical issues highlighted in the readings for this module, and your work on the datasheet in Question 1, can you identify any ethical problems related to this dataset? Either

1. clearly state each issue (1-3 issues) and write a few sentences why you think each issue is potentially problematic,
2. or, if you think the dataset is completely without such ethical issues, write 200 words motivating why.

Problem 3: Data Privacy and the law

First of all, read up on the EU GDPR regulations. The following website explains the most important things in plain and clear language: `https://gdpr.eu`.

Consider the following scenario: A university is offering a course in Data Science in which the students submit multiple written assignments during the course. Is it legal for the university to process this data in the following ways (motivate your answer with a few sentences and refer to the relevant points in GDPR Article 6 of when it is allowed to process data, as per the above website):

(a) The university sells the results data together with student's contact details to a private company offering personal tutoring, with the intention of weaker students getting a offers of discounted study help.

(b) The university suspects some students for plagiarism, and passes their assignments on to the university legal team.

(c) The university submits statistics to the national board of education about the number of students passing and failing the course.

(d) The university suffers a data leak by which names, contact details and results of assignments are published on the internet. What are the legal obligations of the university in this situation (hint: this is not covered in article 6 but read the rest of the above-mentioned website!).

Returning your report

Write a report, typeset in LaTeX, that answers *all* questions above. Clear and concise answers are preferred, don't pad answers with irrelevant text.

If you refer to outside sources, remember to add an appropriate literature reference (including websites) in references by `\cite`ing the references. It is recommended that you use the package `biblatex` to manage citations.

Place your figures in numbered `figure` environments, with descriptive captions and `\ref` to the figures in your discussion. Likewise, place your tables in numbered `table` environments with descriptive captions and `\ref` to the tables in your discussion.

After grading, you will be given another attempt to revise your report according to TA comments if it is not considered acceptable.

The deadline is *hard*. Late submissions will not be read at all and are considered failed. This means you will not get any feedback for the first round and the submission is considered a revision; there will be no third attempt, so if a late submission is failed, you will need to participate in a later iteration of the course for a re-attempt.