DAT565/DIT407 Assignment 8

Sebastian Miles
miless@chalmers.se

Olle Lapidus
ollelap@chalmers.se

2024-10-24

Problem 1

We answer the selected questions from the article *Datasheets for Datasets* by Gebru et al. [**Gebru:2021**]

Motivation

1). The dataset was created to analyze HR functions, and optimize decision making related to employee management. The dataset can be used to analyze and predict attrition, employee performance. Monitoring diversity, compensation analysis and succession planning.

2). The data was created by Fahad Rehman, a student at Abasyn University of Peshawar, Pakistan. He is a Data scientist and Graphic Designer by hobby, according to his github.com biography.

Composition

5). There is only one type of instance, which is the employee record. Meaning the dataset is structured around employees only.

6). There are 15000 employee records in total in the dataset.

8). Each employee record includes the following 10 attributes:

(a) Satisfaction level
(b) last evaluation
(c) number project
(d) average monthly hours
(e) time spent with company
(f) work accidents
(g) if they are still there
(h) promotion within the last 5 years
(i) department
(j) salary

9). We could use "left" as the target label. Given all of the other parameters the model can predict whether or not the employee is going to leave or not.

15). The data contains whether or not an employee has had a work accident which could in some cases be legally confidential. Such data could be considered sensitive, especially if it is linked to an individual.

16). Salary and promotions are often private and sensitive as well.

17).

18). Even though the data does not have a direct identification of the employee, it is possible to identify if the entry stands out, but then by finding the entry, it is likely that the data was known in order to find it anyways.

19).

Collection process

26).

27).

28).

29).

Uses

40).

41).

Problem 2

Problem 3