DAT565/DIT407 Assignment 8

Jonatan Markusson
jomarkusson@gmail.com

William Norland
Williamnorland@gmail.com

2024-05-20

Problem 1: Create a Datasheet

Questions

1. For what purpose was the data- set created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
Answer: As stated on Kaggle this dataset was created to help optimize HR functions.

2. Who created the dataset (for ex- ample, which team, research group) and on behalf of which entity (for ex- ample, company, institution, organization)?
Answer: This dataset was created by a Pakistani student and "Kaggle expert" called Fahad Rehman. Gauging from the profile of the creator it seems like this dataset was created on behalf of the Kaggle community and for the purpose of self promotion in the data science community (by contributing good data).

5. What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edg- es)? Please provide a description.
Answer: The instances, or rows, in this dataset is company employees. The columns represent different attributes of the employees.

6. How many instances are there in total (of each type, if appropriate)?
Answer: There are 14999 instances in this dataset.

8. What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.
Answer: Each instance is a csv row with 10 columns. The data is processed and ready to use out-of-the-box.

9. Is there a label or target associ- ated with each instance? If so, please provide a description.
Answer: No each instance is not labeled but each instance has the following attributes: satisfaction level, last evaluation, number project, average montly hours, time spend company, Work accident, left, promotion last 5years, sales

and salary.

15. Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
Answer: Yes, the dataset includes data that can be considered confidential. Satisfaction level, Number of projects, Average Monthly hours and work accidents are all containing personal information that the employee can be damaged if leaked, but in this dataset the employees are anonymous, so you cant connect a person to the specific data.

16. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. If the dataset does not relate to people, you may skip the remaining questions in this section.
Answer: Yes, if you take a look on an employee who has caused a lot of accidents, our haven't got a promotion in the last five years or have a lower salary than the rest of their department. The employee in person might feel insulted since the data basically tells them they are not doing their job correctly.

17. Does the dataset identify any sub- populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
Answer: Yes, by the column department, for example IT, Sales or Management. Here is a table of the ratio: 1

Table 1: Subpopulation Distribution by Department

Department	**Distribution**
Sales	27.60%
Technical	18.13%
Support	14.86%
IT	8.18%
Product Management	6.01%
Marketing	5.72%
Research and Development	5.25%
Accounting	5.11%
Human Resources	4.93%
Management	4.20%

18. Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.
Answer: No, not in this dataset. The data consists over 14000 employees, so identifying one indidivual would be impossible.

19. Does the dataset contain data that might be considered sensitive in any way (for example, data that re- veals race or ethnic origins, sexual orientations,

religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
Answer: Yes, the data can be mapped to provide financial data, since salary a part of this, although it's only in the categories low, medium and high one can still together with other data connect departments with salary to map peoples financial situation.

26. Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. If the dataset does not relate to people, you may skip the remaining questions in this section.
Answer: According to the website where the data was retrieved, no ethical review processes were conducted.

27. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?
Answer: The data was obtained via a website: [2]

28. Were the individuals in ques- tion notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
Answer: This is not clear on the website the data was retrieved.

29. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
Answer: This is not clear on the website the data was retrieved.

40. Is there anything about the composition of the dataset or the way it was collected and preprocessed/ cleaned/labeled that might impact future uses? For example, is there any- thing that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individu- als or groups (for example, stereotyp- ing, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there any- thing a dataset consumer could do to mitigate these risks or harms?
Answer: This dataset is very straightforward with a csv file with 10 columns which can easily be made into a dataframe.

41. Are there tasks for which the da- taset should not be used? If so, please provide a description.
Answer: This data should mainly be used for HR analyzing, so let's say you want to analyze pet food prices(our anything other not related to HR-questions), this data should not be used.

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

Answer 1.1: The dataset contains performance data of employees, which could be used to make decisions about the employees. This could be very anxiety inducing for the employees since what might be a very understandable diversion from work might not show up in a very non emotional dataset.

Answer 1.2: The dataset does not identify any subpopulations which could lead to decision about vurnerable subpopulations that might be very damaging to the community as a whole. For example the classic issue with women being significantly more likely to have to take maternity leave. Since we cant see the gender of the employees its much harder if not impossible to account for this in our analysis.

2. or, if you think the dataset is completely without such ethical issues, write 200 words motivating why.

Answer 2.1: No, we stated our issues above

Problem 3: Data Privacy and the law

For these we used EU GDPR article 6 as source [1]

(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.
Answer: This violates GDPR article 6 as the university does not seem to have consent from the students to do this for this specific purpose (6a). The university also violates 6d since it does not protect to interests of the students, who may not want offers on study help.

(b) The university suspects some students for plagiarism, and passes their assignments on to the university legal team.
Answer: This should be fine if this is handled in the right way, according to 6f. Since the university has a legitimate reason to forward this data to their legal team, on suspicion of cheating. Also their legal team can be seen as the same party as the university themselves.

(c) The university submits statistics to the national board of education about the number of students passing and failing the course.
Answer: This is fine since the university does not need to include personal information about the students to supply this data. This is allowed according

to GDPR Article 6, since the university has a legal obligation to report this data to the national board of education.

(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!).
Answer: According to article 33 the university must instantly notify this leakage to the supervisory authority, no more than 72 hours after becoming aware of the leak. If the data is sensitive the university must also contact the students affected by this so they are aware of the extend of their leaked information.

References

[1] EU. *GDPR in EU.* Retrieved 2024-05-20. Unknown. URL: `https://gdpr.eu`.

[2] Fahad Rehman. *HR Dataset.csv.* Retrieved 2024-05-20. Unknown. URL: `https://www.kaggle.com/datasets/fahadrehman07/hr-comma-sep-csv/data`.

A Code

```python
    import pandas as pd
import matplotlib as plt
raw_df = pd.read_csv('HR_comma_sep.csv')

department_ratio = raw_df['Department'].value_counts(normalize=True)

management_df = raw_df[raw_df['Department'] == 'management']
print(management_df)
```