DAT565/DIT407 Assignment 8

Saif Sayed
gussayedfa@student.gu.se

Gona Ibrahim Abdulrahman
gusibrgo@student.gu.se

2024-05-17

This is a report for assignment 8 of the course *Introduction to Data Science & AI* from Chalmers and Gothenburg University.

Problem 1: Create a Datasheet

Firstly, we utilized Python and the Pandas library to read and analyze the dataset. We import the Pandas library and read the **HR_comma_sep.csv** dataset into a dataframe called df. Appendix E contains our source code.

Datasheet

A Motivation

For what purpose was the dataset created?

It's designed for HR analytics to uncover workforce patterns, gauge employee engagement, forecast turnover, and inform strategic HR decisions.

Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?

The dataset's origin is unspecified, but it's hosted on Kaggle, indicating it was likely shared by a user, possibly "fahadrehman07".

B Composition

What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)?

The "HR Dataset" catalogs individual employees' profiles.

How many instances are there in total (of each type, if appropriate)?

There are 1000 instances of employees from a company.

What data does each instance consist of?

The “HR Dataset” profiles employees with metrics like satisfaction level and performance score (0-1), project count, average monthly hours, years at the company, work accident presence (0 or 1), company departure status (0 or 1), recent promotions (0 or 1), department, and salary range (low, medium, high).

Is there a label or target associated with each instance?

Regarding the question about labels or targets associated with each instance, the information provided in the dataset does not explicitly mention the presence of a specific label or target variable. It seems that the dataset primarily focuses on capturing various attributes and characteristics of employees rather than predicting a specific outcome or target variable.

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)?

HR datasets generally include information related to employment, such as employee performance, demographics, and job-related attributes. The dataset, in question, does not include sensitive or confidential information such as private communications or personal medical data.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

These datasets aim to enhance workforce management insights. While not inherently offensive, data analysis must be conducted ethically, respecting privacy and well-being.

Does the dataset identify any subpopulations (for example, by age, gender)?

The CSV dataset features “Department” and “Salary” columns, identifying employee subgroups by work area and pay grade. “Department” sorts employees into areas like Sales or IT, while “Salary” classifies pay as low, medium, or high. Analyzing these can reveal the workforce’s departmental and financial composition.

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?

When handling datasets with personal data, privacy should be prioritized. Direct identification risks arise from unique identifiers like employee IDs. Indirect risks stem from merging data sources. But in our dataset there is not such information to direct identifying the person.

Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals 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)?

The dataset is not inherently sensitive.

C Collection process

Were any ethical review processes conducted (for example, by an institutional review board)?

No ethical review processes were conducted according to the information on the Kaggle website.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?

The data was collected from Job posting sites for a company containing 1000 employee records.

Were the individuals in question notified about the data collection?

The data was scraped using selenium package in python environment. According to the information on the Kaggle website, we don't think the individuals were notified.

Did the individuals in question consent to the collection and use of their data?

Since the data was likely open and publicly available on job posting sites, we believe that the individuals might have given their consent for their data to be collected and used. However, it depends on the licensing of the data and whether it is protected by copyright.

D Uses

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

The dataset's labeling may influence its application, but it doesn't inherently cause unfairness. Users must evaluate datasets for biases and privacy issues to prevent stereotypes and ensure findings are equitable and representative.

Are there tasks for which the dataset should not be used?

There is not directly any tasks warning to don't be used but I think possibly could be salary but in our dataset salary is not counted by amounts it is only showing by average.

Problem 2: Ethics

Ethical issues in HR datasets may include privacy breaches, data security lapses, workplace biases, lack of consent, transparency issues, data inaccuracies, misuse of information, and excessive employee monitoring. However, in our dataset, we don't directly observe any factors that could potentially lead to ethical issues in the future. The dataset is anonymous and does not directly exploit individuals' privacy. Although it contains private data such as salary, the dataset does not state the exact amount but rather categorizes it from low to high. Even if the dataset were cross-checked with other sources to identify employees, it would still be impossible to determine their exact salary, which is the only private and sensitive data in the dataset. This also reduces the chances of data security breaches since sensitive data are not explicitly stated.

Moreover, the dataset does not seem to include information about employees' races or genders, thus avoiding workplace biases. However, there might be a lack of consent when collecting the data. We cannot be certain if this is always the case since the data is scraped from job posting sites. It depends on whether the data is publicly available for use on the sites under copyright laws and whether the employees have given their consent.

Problem 3: Data Privacy and the law

(A) Selling the results data and student contact details to a private company for the purpose of offering discounted tutoring:

This would likely not be considered a lawful processing activity under the GDPR. The purpose of selling the data to a third party for commercial gain is not compatible with the original purpose of collecting the data for the university course. Additionally, there is no indication that the students have given consent for their personal data to be shared in this way. This would violate the principles of purpose limitation and consent under Article 6.

(B) Passing student assignments to the university legal team for suspected plagiarism:

This scenario would likely be considered a lawful processing activity under Article 6(1)(c), which allows processing necessary data for compliance with a legal obligation. If the university has policies or procedures in place that require investigating suspected academic dishonesty, then passing the assignments to the legal team would be necessary to fulfill this legal obligation.

(C) Submitting course statistics to the national board of education:

This would be a lawful processing activity under Article 6(1)(e), which allows processing necessary for the performance of a task carried out in the public interest. Providing educational statistics to a government authority would likely be considered a task in the public interest, as it helps inform policy and oversight of the education system.

(D) Data leak resulting in publication of student names, contact details, and assignment results:

This scenario goes beyond just the lawful bases for processing covered in Article 6. The GDPR also has strict requirements around data security and breach notification under Articles 32-34. In the event of a data breach where personal data is exposed, the university would have a legal obligation to notify the relevant supervisory authority and the affected data subjects. They would also need to take appropriate measures to secure the data and mitigate the risks to the individuals. Failure to do so could result in significant penalties under the GDPR.

Appendix

E Python code

```
import pandas as pd

# Read the dataset into a Pandas dataframe
df = pd.read_csv("HR comma sep.csv")

# Display the first few rows of the dataset
print(df.head())

# Calculate basic statistics of the dataset
statistics = df.describe()
print(statistics)

# Obtain information about the dataset columns
info = df.info()
print(info)

# Write the datasheet to a file
datasheet = pd.DataFrame()
datasheet['Feature'] = df.columns
datasheet['Description'] = ['Employee satisfaction 
    level',
                                    'Employee evaluation score',
                                    'Number of projects',
                                    'Average monthly hours',
                                    'Time spent with the company',
                                    'Workplace accidents',
                                    'Left the company',
                                    'Promotion in the last 5 years',
                                    'Department',
                                    'Salary']
datasheet['Data type'] = df.dtypes
datasheet['Missing values'] = df.isnull().sum()
datasheet['Unique values'] = df.nunique()

datasheet.to_csv('datasheet.csv', index=False)
```