DAT565/DIT407 Assignment 8

Giacomo Guidotto
gusguigi@student.gu.se

Leong Jia Yi, Janna
gusleoji@student.gu.se

2024-10-31

1 Create a Datasheet

Question 1: For what purpose was the dataset created?
This dataset was created to analyze and help optimize key HR functions such as workforce planning and diversity monitoring.

Question 2: Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
The author is Fahad Rehman, a student at BSCS at the Abasyn University of Peshawar. He is an individual author, not for any organization.

Question 5: What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)?
The instances of the dataset represent employees with several attributes to describe them as an employee.

Question 6: How many instances are there in total (of each type, if appropriate)?
There are a total of 14999 instances of employees

Question 8: What data does each instance consist of?
The instances consist of 10 features, namely:

- $satisfaction\_level$: Employee satisfaction score (1-5 scale)
- $last\_evaluation$: Score on last evaluation (1-5 scale)
- $number\_project$: Number of projects employee worked on
- $average\_monthly\_hours$: Average hours worked in a month
- $time\_spend\_company$: Number of years spent with the company
- $work\_accident$: If an employee had a workplace accident (yes/no)
- $left$: If an employee has left the company (yes/no)
- $promotion\_last\_5years$: Number of promotions in last 5 years
- $Department$: Department of the employee
- $salary$: Annual salary of employee

Question 9: Is there a label or target associated with each instance?
There are some labels and this depends on how this dataset is used. For example, suppose the analysis intends to determine the relationship between employee characteristics and whether a person has left the company. The "left" feature may be used as a label. A similar case stands for other features such as "work_accident".

Question 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)?
No, all the data was not confidential. All the data could be publically accessible as it did not involve the employees' personal information like their social security number or phone number

Question 16: Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
Even though this data contains scores that rate the employees which could be sensitive, the score cannot be linked back to the employee as there is no id attached to each row. Thus, it cannot cause the ill effects mentioned in the question.

Question 17: Does the dataset identify any sub-populations (for example, by age, or gender)?
It does not identify any sub-populations.

Question 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?
No, the data represent employees from an unknown company with no id attached to each instance.

Question 19: 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)?
No, this data does not contain any sensitive information

Question 26: Were any ethical review processes conducted (for example, by an institutional review board)?
This information was not explicitly stated. However, seeing as this dataset was created by a single author and is not backed by a reputable organisation, it is likely that there were no ethical review processes.

Question 27: 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 obtained through web scraping from job posting sites.

Question 28: Were the individuals in question notified about the data collection?)?
This is not explicity stated in the source but because it was obtained through web scrapping, it is unlikely that the author contacted the individuals about data collection.

Question 29: Did the individuals in question consent to the collection and use of their data?
This is not explicity stated in the source but because it was obtained through web scrapping, it is unlikely that the author contacted the individuals to consent.

Question 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?
There was minimal description of how the data was collected and cleaned. However, we know this dataset was produced by a single individual who was not backed by a reputable organization. Thus, results can be biased if there were mostly collected from a particular subgroup(eg. majority race). Consequently, future models trained on this dataset may yield results that are unrepresentative of minority groups

Question 41: Are there tasks for which the dataset should not be used?
Yes, the dataset should not be used for tasks with severe consequences. For example, it should not be used to determine whether to fire an employee. This is because this is a limited dataset with only a few features from an unknown company. Thus, the accuracy of models trained from this dataset would not be high enough to justify affecting someone's life because of its result.

2 Ethics

This dataset presents a series of ethical issues that can become problematic and affect the application for which it can be used.

The first issue concerns the purpose of this dataset which may not be ethical. Following the 12 ethical guidelines in the Data Values and Principles manifesto, we need to "Consider carefully the ethical implications of choices we make when using data, and the impacts of our work on individuals and society."[1]. Thus, in this case where the author states that one use of the data is to determine the quality of employees - "segmenting high-risk employees" and "identify high potential ones", we need to consider the ethical implications. We believe that is not ethical to rely on tools based on this dataset to rate employees. From our analysis in question 1, the features describing the instances in the dataset are not extensive and its accuracy is not vouched for. Thus, it could potentially give inaccurate results and it is unethical to allow a potentially inaccurate tool to alter a persons career trajectory.

The second issue concerns the author's non-transparency regarding the data collection process. Another ethical guideline mentioned was that one should "Recognize and mitigate bias in ourselves and in the data we use"[1]. However, from our analysis in question 1, it does not clearly state if there were ethical reviews to process the data, which populations were the dataset sampled from or even the author's background. Thus, we do not see deliberate attempts to mitigate bias and ensure fairness in this dataset. Utilizing a biased dataset would not accurately reflect the patterns of all populations which is problematic.

3 Data Privacy and the law

a) Not legal. According to Article 6 paragraph 1, at least one of the conditions must be fulfilled for the processing of data to be legal[2]. However, none of the conditions are fulfilled because the university doesn't need to sell the data to a tuition company and the students have not given consent for this transaction.

b) Legal. According to Article 6 paragraph 1, at least one of the conditions must be fulfilled for the processing of data to be legal[2]. In the case of plagiarism, it fulfills part 1e as processing this information is necessary to determine whether an actual misdemeanor has been carried out and it is being conducted by an official authority vested in the controller, the university administrators.

c) This would depend on the purpose of why the national board of education is receiving the grade results. Under Article 6 paragraph 1 part, processing data is allowed if it is necessary for the performance of a task to be carried out in the public interest[2]. Thus, if the national board was processing the data to improve the school system and ultimately benefit the students, it would be legal.

d) Firstly, according to article 55[2], the university must notify the supervisory authority within 72 hours without delay. Furthermore, the university must also notify affected students of this breach. According to article 33[2], the university must document any personal data breaches, the effects, and their response to the situation.

References

[1] Daniel Kaplan Benjamin Baumer and Nicholas Horton. *Modern Data Science with R*. Retrieved 2023-10-17. 2023. URL: https://mdsr-book.github.io/mdsr2e/ch-ethics.html#some-principles-to-guide-ethical-action/.

[2] General Data Protection Regulation. *Complete guide to GDPR compliance*. Retrieved 2023-10-17. nd. URL: https://gdpr.eu/.