DAT565/DIT407 Assignment 8

Connell Hagen
connellh@student.chalmers.se

Måns Redin
mansre@student.chalmers.se

2024-10-24

1 Datasheet

1.1 Motivation

1. For what purpose was the dataset created?

This dataset's purpose is to provide employee data in order to optimize some key HR functions.

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

This dataset was created by Fahad Rehman on behalf of the company that this data was collected from.

1.2 Composition

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

Each instance of the dataset is a record of various numerical and categorical information.

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

14,999

8. What data does each instance consist of?

- satisfaction_level: Employee satisfaction score (0-1 scale)
- last_evaluation: Score on last evaluation (0-1 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

9. Is there a label or target associated with each instance?

- Department: sales, accounting, hr, technical, support, management, IT, product_mng, marketing, RandD
- Salary: low, medium, high

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)?

- Salary: This column divides the employees into different salary ranges. E.g. low, medium or high, which is financial data connected to a specific person.
- Workplace accidents: There may be a level of confidentiality regarding the disclosure of who was involved in a workplace accident.

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

No.

17. Does the dataset identify any sub-populations (for example, by age, gender)?

No.

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?

It may be possible to do so for people who are within categories that are uncommon. For example, someone who is making a high salary, in one particular department, who has been in a workplace accident will likely not have many other matching people.

9. 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.

1.3 Collection Process

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

We are uncertain if there was an ethical review process. Presumably there was internally within the company, since there was a survey portion to the data, implying that the company was planned the collection of this data, and it was not just taken.

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

This dataset was scraped with Selenium, so it was most likely not directly collected from the employees by the creator of the dataset.

28. Were the individuals in question notified about the data collection?

Probably not.

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

Probably not.

1.4 Uses

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?

The salary is labeled as low, medium, or high. Specific amounts are not able to be used for future uses of the data set. Also, the number of years at the company are all whole numbers, so any analysis involving exact times is not possible.

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

Determining if a particular employee should be fired. Determining if someone deserves their salary.

2 Ethics

The potential ethical issues with using this dataset that we identified are: determining if a particular employee should be fired and determining if someone deserves their salary.

We believe that these may be tempting conclusions that can be made from the data, but this dataset does not provide enough data to make a good conclusion on these matters.

For example, the number of projects that someone is working on would likely be a contributing factor to the conclusions that would be drawn here. However, this dataset does not take into account the size of the projects, or how vital someone's role is within their projects. Also, the number of hours worked is not a good representation of the value that is being provided to the company. Someone could work less hours, but actually be more productive. Someone could also have worked less hours because they have an injury or an illness, in which case expecting them to work would not be reasonable.

The metric of the evaluation scores could also have some bias embedded within it. This is because these scores are most likely determined by a human, and they will let their internal biases affect these scores. For example, perhaps they rate people different based on their gender, race, beauty, etc., and using this metric to make a decision could reinforce that bias.

3 Data Privacy and the Law

a) This is not legal. This is because in article 6, section 1, the usage of data does not fit into any of the required cases.

b) This is legal. This fits under article 6, section 1, case b: ”processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;” In this case, the contract is the contract of class. It is agreed that by enrolling in the class, plagiarism is disallowed and will be investigated.

c) This is legal. This fits under article 6, section 1, case f: "processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child." This is because although it might not be legally required for a school to disclose this data to the board of education, it may still be within their best interest to do so. This would be for the purpose of progress measurement, funding acquisition, etc. Case f applies here because a private university is not a public authority.

d) Whoever discovers the data breach needs to notify the supervisory authority within 72 hours of discovery. They need to describe the data breach, including the number of subjects concerned, the likely consequences, and measures to take to address the data breach. The controller must document the data breach.