Datasheet for Infrastructure Intelligence Dashboard datasets

Dayse Rivera, Larissa Roberta, Lidiane Parisi, Lucas Romao, Maísa Mendonça, Vinícius Guimarães

January 23, 2024

1 Datasheets for datasets

Datasheets for Datasets "document [the dataset] motivation, composition, collection process, recommended uses, and so on. [They] have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted biases in machine learning systems, facilitate greater reproducibility of machine learning results, and help researchers and practitioners select more appropriate datasets for their chosen tasks."

The motivation behind the proposal was the electronics industry, where every component has a datasheet that describes its operating characteristics and recommended uses. In machine learning, data is the input for model training. Using the wrong dataset, or using a dataset outside of its original intent, or even not understanding well enough the limitations of a dataset, has dire consequences for the model. However, "[d]espite the importance of data to machine learning, there is no standardized process for documenting machine learning datasets. To address this gap, we propose datasheets for datasets."

2 Template

Purpose and Origin

What motivated the creation of the dataset, who created it, and who funded its creation?

All the datasets were provided by The London Datastore. The London Datastore, created by the Greater London Authority, aims to make London's data accessible for public use. The GLA is committed to influencing other public sector organizations to release their data on the platform, supported by Mayor Sadiq Khan.

What specific gaps or tasks does it aim to address?

To allow transparency into account various socioeconomic indicators, such as education, drug use, criminality,

health wellbeing, and financial weelbeing.

Composition and Representation

What types of data are included (e.g., documents, photos, people), and how many instances are there?
Only excel files which contains 224 instances each.

Is it a complete set or a sample, and how representative is the sample of the larger set?
Complete set of entire London between 2016 and 2022

Data Characteristics and Labeling

What are the characteristics of each data instance (raw data or features)?
Raw data for each London borough between 2016 and 2022

Are there labels or targets associated with the instances, and are relationships between instances explicitly defined?

- Well-being health (%)
  - Life satisfaction
  - Obesity
  - Anxiety
  - Exercise practices
  - Chronic deceases
- Financial well-being (%)
  - Rough sleepers
  - Up-to-date with household bills
  - Unemployment
  - Inequality
- Criminality
  - Total of crimes
  - Individual safety crimes
  - Societal order crimes
  - Property crimes
  - Violent Crimes
- Drug
  - Total Drug users
  - Cannabis
  - Cocaine
  - Ecstasy
  - Other Drugs
- Education (%)
  - NVQ0
  - NVQ1
  - NVQ2
  - NVQ3
  - NVQ4+

Ethical and Legal Compliance

Does the dataset adhere to legal and ethical guidelines, such as GDPR?
Yes, it does.

Are there concerns about confidentiality, offensive content, or identification of individuals or sensitive sub-populations?
There is no problem with confidentiality, offensive content and identification of individuals or sensitive subpopulations

Collection Process and Validation

How was the data collected and validated?

All the data was provided by The London Datastore.

What mechanisms or procedures were used, and who was involved in the data collection process?

The datasets where only downloaded from the database

Preprocessing and Data Integrity

What preprocessing, cleaning, or labeling methods were applied?

All process of data treatment was made in excel, using formulas and creation and removal of columns and lines.

Was the raw data saved alongside processed data, and is the preprocessing software available?

Yes

Usage and Application

How has the dataset been used, and what other potential applications does it have?

The datasets were used to build a dashboard using Tableau Software. The dashboard will be used as an MVP to provide insights in public investments.

Are there any limitations or considerations for its use in specific tasks?

It is only supposed to be used to provide insights, not as a decision tool

Distribution and Accessibility

How will the dataset be distributed?

Open in the London Datastore https://data.london.gov.uk/

What are the copyright, licensing, and regulatory considerations?

It is open to use

Maintenance and Updates

Who is responsible for dataset maintenance and updates, and how can they be contacted?

The dataset maintenance and updates are response of London Datastore

Are there plans for regular updates, and how will these be communicated?

The dataset maintenance and updates are response of London Datastore

Preprocessing and Data Integrity

Is there a process for receiving external feedback and contributions to the dataset?

Feedbacks are made in the London Datastore e-mail datastore@london.gov.uk

How are these contributions validated and integrated?

By the London Database support team