Ethical and Socially-Aware Data Labels

Juan Carlos De Martin

Nexa Center for Internet & Society
Politecnico di Torino - Italy
http://nexa.polito.it

Publication Authors:
Elena Beretta, Antonio Vetrò, Bruno Lepri,
J.C. De Martin

Publication available at:
https://www.springer.com/us/book/9783030116798
https://nexa.polito.it/node/1515

crucial role of data for the
design and development of
machine learning algorithms
and, more generally,
of many digital systems

“if the admission models to American universities had been trained on the basis of data from the 1960s, we would probably now have very few women enrolled, because the models would have been trained to recognize successful white males.”

Cathy O’Neal, “Weapons of Math Destruction,” Crown Books, 2016

to avoid discrimination and
other unintended negative
effects, care is needed
at all stages of the design
and development process

1. data collection
2. data usage
3. …

key idea:

to support computer scientists using datasets by means of easy-to-understand labels certain data characteristics may lead to discriminatory decisions and therefore it is important to identify them and show the potential risks.

1. data collection
2. data usage
3. …

labeling datasets using measures of certain input data characteristics (e.g., uneven distribution in gender balance, co-linearity of attributes, etc.) that represent a risk of discrimination if used in decision making (or decision support) systems useful to software engineers to be more aware of the risks of discriminations and to use the dataset in an more ethically and socially-aware manner.

In addition, it could be used by third parties to more easily identify risks on a given dataset.

Other initiatives in the same direction:

“The Dataset Nutrition Label Project”

https://datanutrition.media.mit.edu

“Datasheets for Datasets”

by Gebru *et al.*

Three building blocks for EASAL

1. Disproportionate Datasets

2. Correlations and collinearity

3. Data Quality

We propose the **ISO/IEC 25012** and **25024** standards models as a reference for quantita- tively assessing the quality of data input and the consequential confidence of the decision made out of that data. In particular, we refer to the inherent quality dimensions: accuracy, completeness, consistency, credibility, currentness.

Testing the EASAL approach on a real case

Credit Card Default dataset

information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

The dataset does not contain the protected attribute “race”, but contains other personal information that can be used in a discriminatory way if applied to assess creditworthiness, such as gender and level of education.

1. Disproportionatess

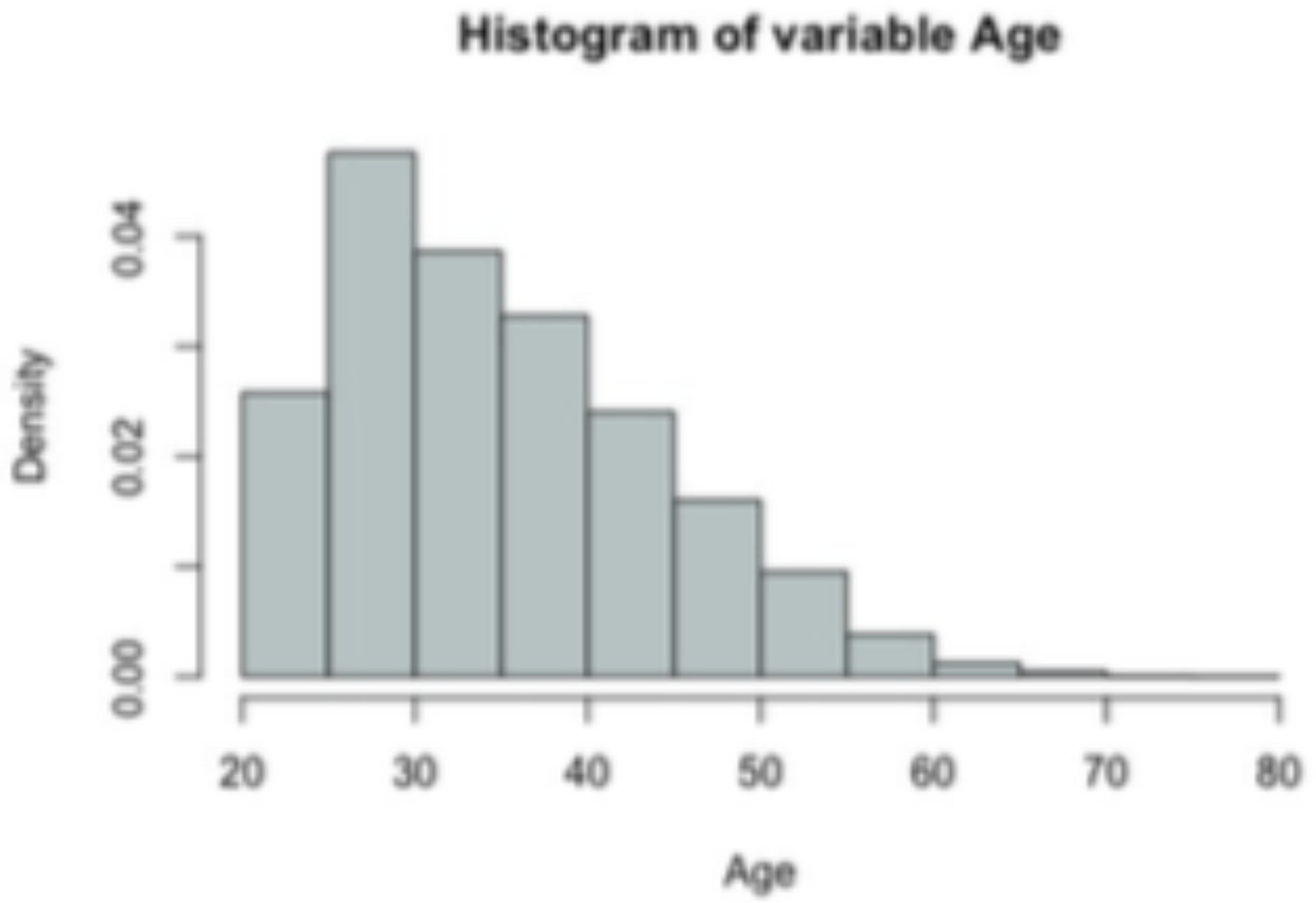

Fig. 1. Frequency of variable *age*

60% women

46.7% have attended college

50-50% single vs married

2. Correlations and collinearity

CORRELATION between “payment default” condition and:

- education level
- gender
- marital status

3. Data Quality

accuracy, **completeness**, consistency, credibility, currentness

CONCLUSIONS

The EASAL approach could help datasets users to be more aware of the potential biases and problems of the dataset before using them to develop systems, therefore reducing the risk of downstream unintended problems.

thank you