A hands-on tutorial on explainable methods for machine learning with Python: application to gender bias

EuADS Summer School – Data Science for Explainable and Trustworthy AI

Kirchberg, Luxembourg, 8th June 2023

Aurora Ramírez
Postdoctoral researcher
University of Córdoba /
DaSCI Research Institute (Spain)

DO IT YOURSELF!

Access to code and examples

1 Repository on GitHub

1. Clone the repository
2. Create your virtual env
3. Install packages
4. Access code and data

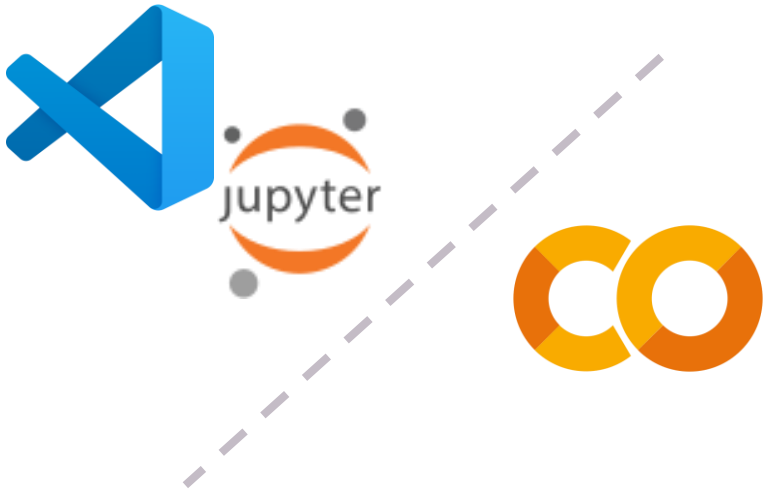
2 Notebooks on GoogleColab

1. Example 1: Data analysis
2. Example 1: ML + XAI
3. Example 2: Data analysis
4. Example 2: ML + XAI

https://github.com/aurorarq/euads-genderbias

The Python toolkit

Environment



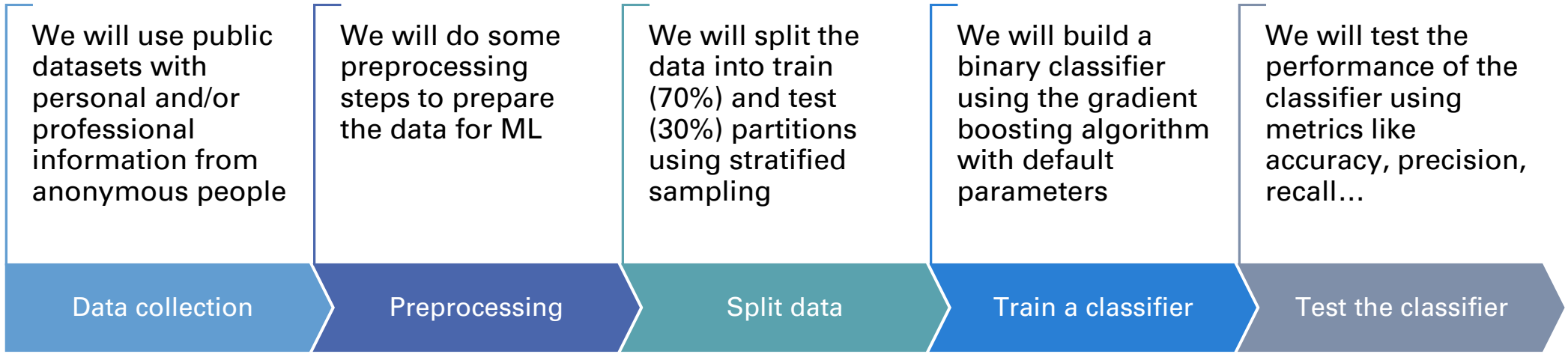
Packages

Source: Official webpages/Wikipedia

SOME INITIAL CONCEPTS

Some initial concepts: ML

- Machine learning task: build a predictive model for binary classification
- Our (basic) pipeline



Some initial concepts: XAI

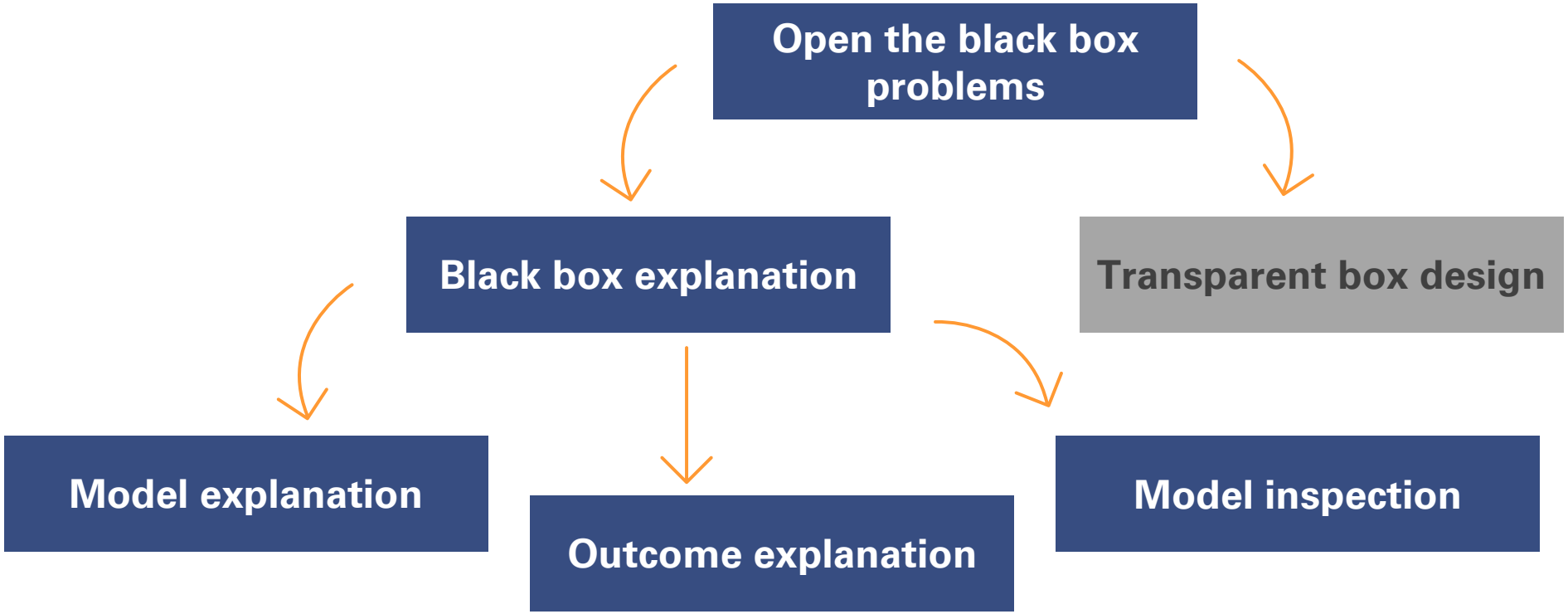
- Definition by Barredo Arrieta et al. [Bar20]

> *Given an audience, an explainable artificial intelligence is one that produces details or reasons to make its functioning clear or easy to understand*

- Use cases of XAI in the context of machine learning (ML) [Dwi23]:
  - Before model deployment: debugging AI models, **detecting bias**, scientific understanding, hypothesizing about new knowledge, more robust models
  - After model deployment: better decision making, **avoid discrimination**, justifiability

Some initial concepts: XAI

- A high-level overview of what XAI methods can do [Gui18]



Some initial concepts: XAI

- **Model inspection** (or dataset-level exploration) [Bie20]
  - Objective: understand how the model behaves from a representative sample of observations
  - In practice: detect which variables are more relevant ("important") in the decision model and how they influence the predictions in general
- **Technique used in this tutorial:** Permutation-based variable importance
  - Calculate the variance of a performance measure based on data permutation
  - For each variable, make a new version of the data (column permutation) and obtain the new predictions from the model
  - Quantify the variance of the performance measure due to the change

Some initial concepts: XAI

- **Outcome explanation** (or instance-level exploration) [Bie20]
  - Objective: understand why the model makes a particular prediction for a given observation
  - In practice: assign scores to each variable by analysing how the model behaves in the vicinity of the given observation (local)
- **Techniques used in this tutorial:** Break Down [Gos19, Bie20] and SHAP [Lun17]
  - Break Down: decomposes the prediction score into positive and negative scores that are assigned to the variables depending on their contribution
  - SHAP: calculates the average of the importance attributed to each variable across all possible data orderings

Some initial concepts: XAI

- **Outcome explanation** (counterfactuals) [Gui22]
  - Objective: understand how the prediction could change (what-if scenario)
  - In practice: generate new samples close to the given observation with feasible changes in feature values so that the prediction changes
- **Technique used in this tutorial:** DiCE [Mot20]
  - Counterfactual generation as an optimisation problem (minimum changes, maximum diversity)
  - Highly configurable: number of counterfactuals, features to change and their desired ranges

TIME FOR PYTHON PROGRAMMING! (EXAMPLE 1)

Example 1: Employee promotion

- **Dataset:** Employee promotion (available on Kaggle)
  - 12 variables (including gender) + target (is promoted?)
  - 54808 observations

- **Notebooks**
  - Data analysis: A look at the data to analyse the distribution of variables
  - ML training and testing: A first ("manual") approach to detecting gender bias
    1. Training and testing a classifier using the full dataset
    2. Training and testing classifiers using data split by gender
    3. XAI techniques for model inspection, outcome explanation and counterfactual generation

Figure from Pixabay by VIN JD

FAIR ML TO MITIGATE GENDER BIAS

Bias in ML

- What is an unfair algorithm? [Meh21]

> *Fairness is the absence of any prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics. Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people*

- Bias comes in many forms [Meh21]
  - Data → Algorithm: what/how we measure affects the representativeness of the observations
  - Algorithm → User: how the algorithm handles the data (popularity or evaluation bias)
  - User → Data: how the users that generate data behave in different contexts or over time

Gender bias in ML

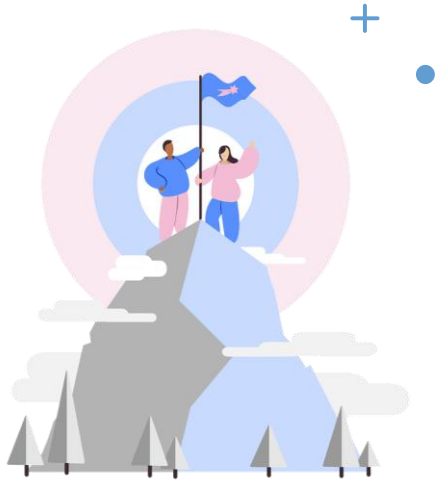
- Gender/sex is one of the common “sensitive” attributes on which an ML algorithm might base its automatic (biased) decisions
- Some possible causes of gender bias:
  - Failure to account for inherent biological characteristics (e.g., health care)
  - Minority groups in certain situations (e.g., management positions in companies)
  - Underrepresentation for historical reasons (e.g., finance autonomy)
- Some examples [Pra20, Bir23]
  - Medicine: Different reaction to treatments or need for specific diagnostic methods
  - Finance: Women found it more difficult to be approved for loans or credits
  - Translation: Assumptions of professional roles when translating from gender-neutral languages

Fair ML

- Many public datasets are known to lead algorithms to make biased decisions regarding sensitive attributes such as race, age, gender [LeQ22]
- Several metrics have been proposed to quantify fairness [Che23]
  - A recent approach to quantifying unfairness uses counterfactual explanations [Kur22]
- Methods fall into three categories [Meh21, Che23]:
  - Pre-processing: transforming the data to eliminate potential causes of discrimination
  - In-processing: modifying the training process (opt. function, constraints) to reduce bias
  - Post-processing: redefining the prediction function to account for bias

TIME FOR PYTHON PROGRAMMING! (EXAMPLE 2)

Example 2: Dutch census



- **Dataset:** Dutch census (available on GitHub [Kur22])
  - 9 variables (including gender) + target (has a good job?)
  - 60420 observations
- **Notebooks**
  - Data analysis: A look at the data to analyse the distribution of variables
  - ML training and testing: Using fairlearn to mitigate gender bias
    1. Training and testing a classifier using the full dataset
    2. Training and testing classifiers with fairlearn to “protect” the gender attribute
    3. XAI techniques for model inspection, outcome explanation and counterfactual generation

Figure from Pixabay by Mediamodifier

SOME FINAL THOUGHTS

Recap & discussion

- Explainable methods for gender bias detection
  - Basic data exploration is quite useful to detect possible biases before building the model
  - Global and local methods can be used to confirm our suspicions
  - Gender bias may exist even if the gender attribute is not highly relevant to the model predictions
  - It is important to set up counterfactual methods to suggest feasible changes (very common for age, race, gender)
- Fair ML to mitigate gender bias
  - Do not assume that the female gender is always the one that suffers discrimination
  - Mitigation methods may reduce the overall influence of the gender attribute, but local gender-based explanations may persist
  - Mitigation of bias may imply a decrease in the accuracy of the model

Acknowledgement

EuADS Summer School – Data Science for Explainable and Trustworthy AI

Thanks to the organisers, sponsors and attendees to the Summer School.

University of Córdoba – Annual Research Plan (2022)

This tutorial has been developed in the context of the GENIA project ("Studying gender bias in machine learning models using explainable artificial intelligence"), funded by the Annual Research Plan of the University of Córdoba (2022).

References

[Bar20] A. Barredo Arrieta et al., “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI”. Information Fusion 58:82-115, 2020.

[Bie20] P. Biecek, T. Burzykowski. “Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models”. CRC Press. 2020. Available on: https://ema.drwhy.ai/

[Bir23] P. Birzhandi, P. Cho. “Application of fairness to healthcare, organizational justice, and finance: a survey. Expert Systems with Applications 216:119465, 2023.

[Che23] Z. Chen et al., “A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers”. ACM Transactions on Software Engineering and Methodology, 2023.

[Dwi23] R. Dwivedi et al., “Explainable AI (XAI): Core Ideas, Techniques, and Solutions. ACM Computing Surveys 55(9):194, 2023.

[Gos19] A. Gosiewska, P. Biecek. “iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models”, 2019. Available on: https://arxiv.org/abs/1903.11420v1

[Gui18] R. Guidotti et al., “A Survey of Methods for Explaining Black Box Models”. ACM Computing Surveys 51(5): 93, 2018.

References

[Gui22] R. Guidotti, “Counterfactual explanations and how to find them: literature review and benchmarking”. Data Mining and Knowledge Discovery, 2022.

[Kur22] A. Kuratomi et al., “Measuring the Burden of (Un)fairness Using Counterfactuals”. Proc. Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), pp. 402-417, 2022.

[LeQ22] T. Le Quy et al., “A survey on datasets for fairness-aware machine learning”. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 12(3): e1452, 2022.

[Lun17] S.M. Lundberg, S-I Lee, “A Unified Approach to Interpreting Model Predictions”. Proc. 31st International Conference on Neural Information Processing Systems, pp. 4768-4777, 2017.

[Meh21] N. Mehrabi et al., “A Survey on Bias and Fairness in Machine Learning”, ACM Computing Surveys, 54(6): 115, 2021.

[Mot20] R. K. Mothilal, A. Sharma, C. Tan, “Explaining machine learning classifiers through diverse counterfactual explanations”. Proc. Conf. Fairness, Accountability, and Transparency, 2020.

[Pra20] M.O.R. Prates, P. Avelar, L. Lamb, “Assessing gender bias in machine translation: a case study with Google Translate”. Neural Computing and Applications, 32:6363-6381, 2020.

DO WE STILL HAVE TIME LEFT?

Proposed exercises

- Notebook 1: employee promotion
  - Calculate fairness metrics for the initial classifier
  - Include a classifier with mitigation mechanism using fairlearn
- Notebook 2: dutch census
  - Compare fairlearn classifiers with classifiers trained with data split by gender
  - Use other fairlearn mitigation methods (e.g., preprocessing)

08/06/2023

THANK YOU!

Aurora Ramírez

aramirez@uco.es

https://www.uco.es/users/aramirez/en

@aurora_rq