Politecnico di Torino
Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering
XXXV cycle

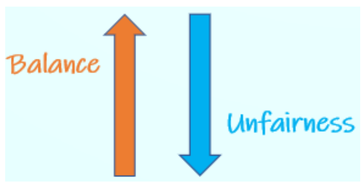
Supervisor
Marco Torchiano

Effects of Data Imbalance on Software Application Bias and Mitigation Strategies, Techniques and Tools

PhD Candidate: *Mariachiara Mecati* - mariachiara.mecati@polito.it

1. Context

Nowadays automated decision–making (ADM) systems affect many aspects of our life. The spread of machines that output decisions or recommendations is facilitated by the availability of a big amount of data (often personal). When the data used to train ADM systems is affected by poor quality and bias, this issue may propagate and cause biased outputs with possible discriminatory consequences toward certain individuals or groups: among the causes of software bias, ***data imbalance*** is one of the most significant issues.



2. Goals

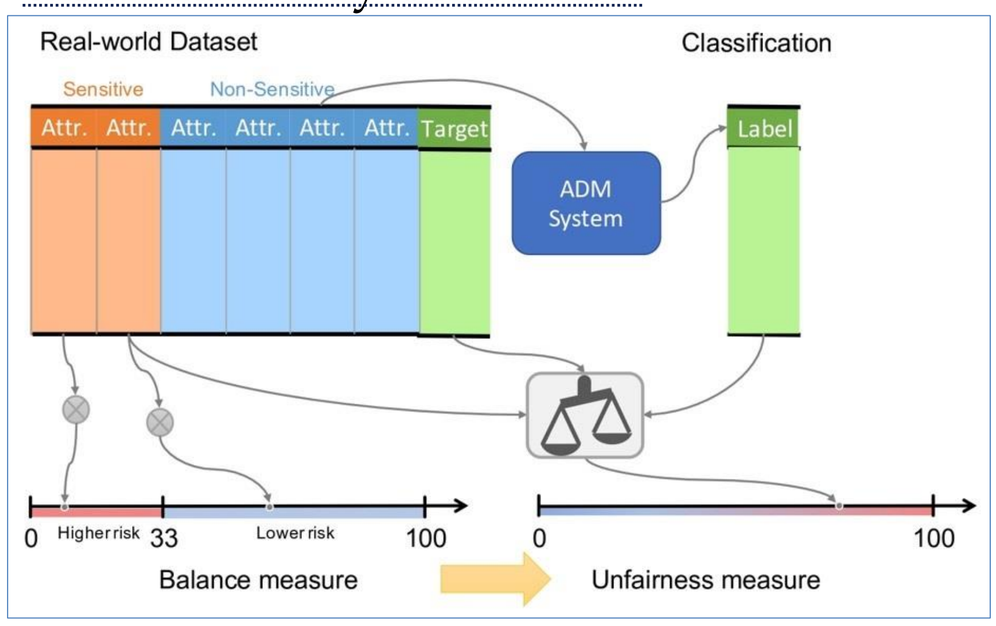
Studying the impact of data imbalance on automated decisions made by software applications: we built a conceptual and operational data measurement framework for detecting imbalance of protected attributes in input data -through four different measures of balance- and we searched for a relational link between the imbalance issue (especially in input data) and biased outcomes.

3. Method

We assume ***data imbalance*** as a ***risk factor*** for systematic discrimination caused by ADM systems: we measured the ***(im)balance*** of protected attributes in training data through 4 balance measures (the Gini, Shannon, Simpson and Imbalance Ratio indexes) and we measured the ***unfairness*** of the classification outcomes through 3 fairness criteria, in order to understand how the balance in input data can be used to detect the risk of algorithmic unfairness.

4. Results

- The balance measures properly detect unfairness of software output, even though the choice of the metric has a relevant impact on the detection of discriminatory outcomes.



- A negative correlation between balance and unfairness measures holds, as the higher the ***balance*** measures (indicating low imbalance in the input data), the lower the ***unfairness*** values (meaning high fairness in the classification outcomes).
- Imbalance in intersectional attributes and target variable play an important role in the detection of biased outputs.

4. Conclusions

Overall the results showed that our approach is suitable for the proposed goal, however further work is necessary to deepen knowledge of the thresholds to consider as risky and the intersectionality topic.

6. References

1. Mecati, Mariachiara; Vetrò, Antonio; Torchiano, Marco. *Detecting Risk of Biased Output with Balance Measures*. Journal of Data and Information Quality. https://doi.org/10.1145/3530787
2. Vetrò, Antonio; Torchiano Marco; Mecati, Mariachiara. *A data quality approach to the identification of discrimination risk in automated decision-making systems*. Government Information Quarterly. https://doi.org/10.1016/j.giq.2021.101619