AutoFair Toolkit: An automated tool for evaluating fairness in AI/ML models

Angelos Poulis
sdi1900230@di.uoa.gr
National and Kapodistrian University of Athens
Athens, Greece

Dimitrios Tomaras, Vana Kalogeraki
{tomaras,vana}@aueb.gr
Athens University of Economics and Business
Athens, Greece

Dimitrios Gunopulos
dg@di.uoa.gr
National and Kapodistrian University of Athens
Athens, Greece

ABSTRACT

The great proliferation of Artificial Intelligence in a wide variety of application domains has produced an imminent need to interpret the output of such models, especially for evaluating their fairness. Such domains may include human resources management for evaluating individual and group fairness, advertising and financial technologies, where fairness can have significant impact. However, existing tools have focused on partial aspects of evaluating fairness, either looking into the problem of explainability or bias detection with no tool available to serve multiple purposes at at time. Moreover, existing tools suffer from poor scalability and developers cannot directly benefit from a variety of available infrastructure resources. In this work, we present the AutoFair toolkit, an automated toolkit able to detect bias in AI/ML models, examine trade-offs between different measures of fairness, certify their fairness a priori and provide explanations for their models.

ACM Reference Format:
Angelos Poulis, Dimitrios Tomaras, Vana Kalogeraki, and Dimitrios Gunopulos. 2024. AutoFair Toolkit: An automated tool for evaluating fairness in AI/ML models. In . ACM, New York, NY, USA, 1 page.

1 INTRODUCTION

Artificial Intelligence has revolutionized various industries, including financial risk analysis, the hiring sector and marketing campaigns among others. However, there is an imminent need to understand and explain the output of such models to a wide range of people, such as possible stakeholders, end-users, policy makers and governing agencies [1]. Fair AI models can e.g. improve hiring processes, provide unbiased ads to end-users and support better explanations regarding financial decisions. Existing tools have limited focus on partial aspects of evaluating fairness. To this end, we build the AutoFair toolkit, a toolkit where users can seamlessly input data, perform risk-free simulations and evaluate the fairness of their AI models. The toolkit builds upon a containerized environment to ensure scaling capabilities with regards to the types of models and amount of data that need to be examined.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

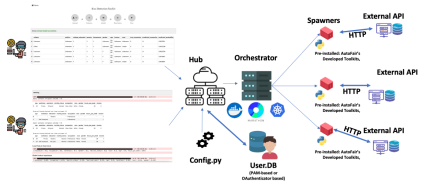

Figure 1: AutoFair toolkit

Figure 2: Bias Detection

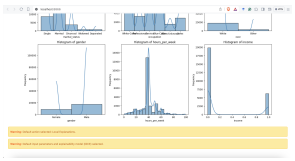

Figure 3: Explainability

2 RESEARCH PROBLEM

AutoFair toolkit (Figure 1) is able to provide insights from the data examined along the model and help practitioners to identify features of importance from the data, which can be further investigated in the model evaluation process. Another important problem we aim to solve is, that, existing tools do not provide an end-to-end pipelined solution so that a user can have a detailed overview of the performance of an AI model. The AutoFair toolkit supports both stand-alone execution (i.e. detect bias in a model solely) and pipeline execution. AutoFair provides a taxonomy of fairness measures for quantifying bias in AI models (in **Bias Detection** activity, Figure 2), a risk-free simulation environment to explore trade-offs in the fairness measures without touching live data, ideal for non-technical people such as policy makers (in **Trade-off Exploration** activity), certification of fairness a priori in order to learn the trade-off between measures of fairness (in **Certification a priori** activity) and finally, by supporting several definitions of fairness, the toolkit provides a means to explore and explain differences and discrepancies between individual and group fairness (in the **Explainability** activity, Figure 3). Our evaluation has shown the practicality and the significant benefits of the AutoFair toolkit.

ACKNOWLEDGMENTS

This research has been financed by the European Union through the the H2020 AutoFair project (No. 101070568).

REFERENCES

[1] [n. d.]. AutoFair Project. https://humancompatible.org/