Evaluation of Bias and Fairness in the University of L'Aquila and in the Italian Academic Data

Andrea D'Angelo, Giordano d'Aloisio, Diana Di Marco, Giovanni Stilo

This document reports the quantitative analysis conducted in the FAIR-EDU project. The aim of this analysis is to assess the amount of gender bias in the local (University of L'Aquila) and National level academic domain and to estimate whether a machine learning classifier is a fair predictor of the academic roles (i.e., researcher, associate professor, and full professor) of men and women starting from the academic data collected in the previous phase of this project.

This document is structured as follows: section 1 discusses the experimental setup, including the data processing pipeline and the proposed experiments, detailing all chosen technologies (e.g., classifiers, debiasers, and others). Then, in section 2, we discuss and visualize the results of the bias metrics. Finally, in section 3, we discuss the experiment's results, highlighting what conclusions can be inferred and what future work remains to be explored.

1 Experimental Setting

In this section, we document how the data was processed, starting from the initial database, in order to produce the final datasets used for the experiments. Then, a summary of said experiments is depicted, and the individual batches are explained in detail.

1.1 Data Pre-processing

The conducted experiments have been performed using the academic data gathered in the previous phases of this project. These data have been collected inside a relational database. Hence, the first step of our pre-processing pipeline has been to aggregate the different tables into a single dataset D' containing all the relevant data. In particular, for each record (i.e., professor or researcher), we

stored the overall and the year-by-year statistics obtained by Scopus, including the specific type of publication (i.e., journals, conferences, Book series, and others). To further represent the quality of a researcher/professor's produced work, we saved the mean Citescore, Sjr, and Snip of the journals in which the publications were made, along with how many years of activity they have in the Italian University system. Finally, we store each person's gender and current role (i.e., researcher, Associate Professor, or Full Professor). This aggregated dataset D' was then thoroughly anonymized to protect the University employees' privacy. As a result, no references to names, surnames, or other sensitive or personal data are stored, as they are neither relevant nor useful for computing bias metrics.

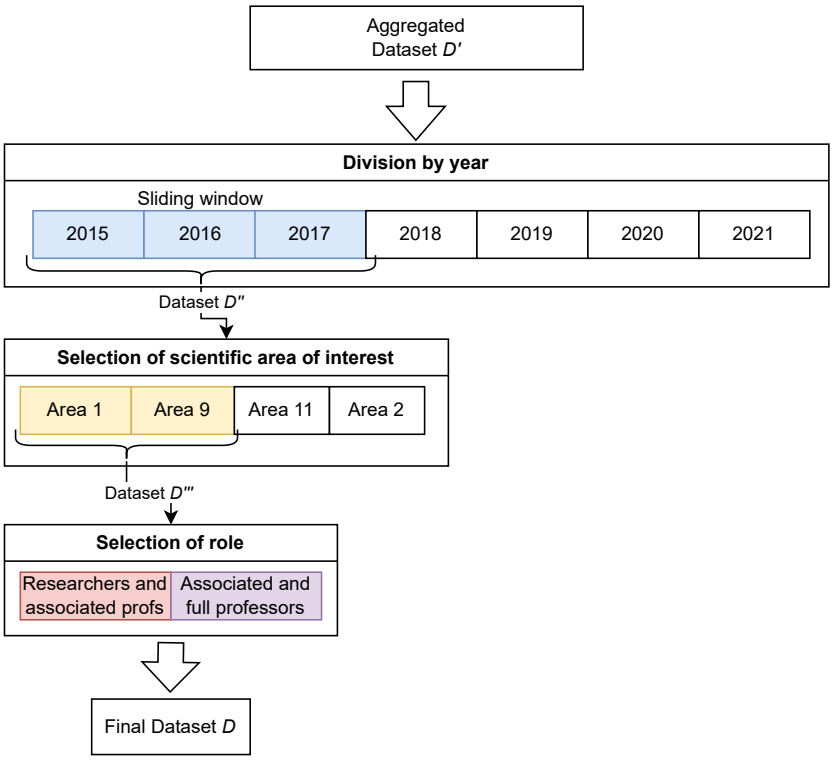

Figure 1: Processing pipeline of the dataset.

Then, starting from the anonymized dataset D', we performed a set of filtering operations to obtain the set of final dataset D that we used to compute bias metrics yearly. The filtering procedure is also depicted in Figure 1.

Since we are interested in the evolution of bias in academic promotions year by year, the anonymized dataset D' was split according to a sliding time window of fixed size. In particular, we considered a sliding window of three years, starting from 2015. Hence, to gather metrics for 2019, with the sliding window size set to 3, we would slice D' to obtain only the columns referencing data collected from

Batch	Classifier	Debiaser	Positive Label	Gathered Metrics
0	None	None	Associate Professor	Disparate Impact
1			Full Professor	
2	Logistic Regression	None	Associate Professor	Disparate Impact, Equalized Odds, Average Odds, Accuracy
3			Full Professor	
4		DEMV	Associate Professor	
5			Full Professor	

Table 1: Summary of the experiments.

Sensitive Attribute	**Unprivileged Group**	**Label**	**Positive Label**
Sex	Women	Current Academic Role	Associate Professor/ Full Professor

Table 2: Attributes for the experiments

2016 to 2019. This allowed us to obtain, visualize, and analyze a time series of bias metrics from which we could recognize the evolution of the data and draw apt conclusions. After this operation, we obtain a partially filtered dataset D''

The subsequent step was selecting only specific scientific areas from D''. Because different domains have different promotion criteria, it would be incorrect to consider them all together. In our study, we only focused on Areas 1 and 9 of the MIUR scientific areas classification, which refers broadly to Science, technology, engineering, and mathematics. From this further filtering, we obtain a dataset D'''. Finally, we split D''' into two further versions: one without records representing researchers and one without Full Professors. The former will be the final dataset for the computation of the bias existing within the context of promotion from researcher to Associate Professor. At the same time, the latter will be used to compare Associate Professors with Full Professors. In the rest of this report, we will refer to these datasets as D.

1.2 Experiment Description

Once the final yearly datasets D have been constructed, the experiments can occur. A general overview of the conducted experiments is depicted in table 1.

We chose {sex: Woman} as the unprivileged group, and for each year, the label is the person's role for that year (i.e., researcher, associated professor, or full professor). Table 2 contains a schema of the selected attributes.

Referring to table 1, the first two batches of experiments are performed on the datasets D without using a classifier or debiaser. The aim is to measure the inherent bias in the data, specifically an intrinsic imbalance between the selected groups. We chose the Disparate Impact as a measure of bias because it can be

calculated directly on the dataset considering the distribution of the true labels Y [3].

The later batches of experiments use a Logistic Regressor to predict the label. We first work on the dataset as-is, without any debiaser method. The main idea here is to try to predict the individual's role solely based on their academic performance, measured by Scopus metrics. By selecting *Women* as the unprivileged group, we can infer whether the classifier places an unfair emphasis on the individual's attribute *Sex* when determining which records are of lower status. Finally, experiments 4 and 5 use DEMV, a pre-processing debiaser method [2] for classification tasks. For this batch of experiments, we computed the following metrics on the classifier: Disparate Impact [3], Equalized Odds [4], Average Odds (i.e., equal accuracy) [1], and Accuracy [5].

All the experiments depicted in table 1 were conducted twice with a slight change of scope: initially within the context of the University of L'Aquila and later expanding the range to the whole of Italy. In both circumstances, we only preserved data for workers that were employed either at the University of L'Aquila or at an Italian university, respectively, for the entire duration of the reference time window. In addition, studies on the classifier's Feature Importance and the number of men and women in each academic role were conducted to elaborate on these findings.

2 Experiment Results

In this section, we report on the results of the conducted experiments for both Univaq and Italy. We first report in subsection 2.1 a description of the population, reporting, in particular, the distribution of roles globally and by sex for each considered year. Next, the results for different batches of experiments, with and without a classifier, are depicted respectively in subsections 2.2 and 2.3. These batches are intrinsically different since the former will only compute the inherent bias present in the datasets. At the same time, the latter will focus on how this bias is inherited by a classifier. As for table 1, different bias metrics are reported in each section.

2.1 Population Description

In this section, we report the distribution of the population studied for both Univaq and Italy. We recall that, as already said in section 1, our study focuses for this pilot only on researchers and professors of scientific areas 1 and 9 (STEM).

In particular, figures 2 and 3 report respectively the aggregated number of people and the number of men and women for each academic role in Univaq

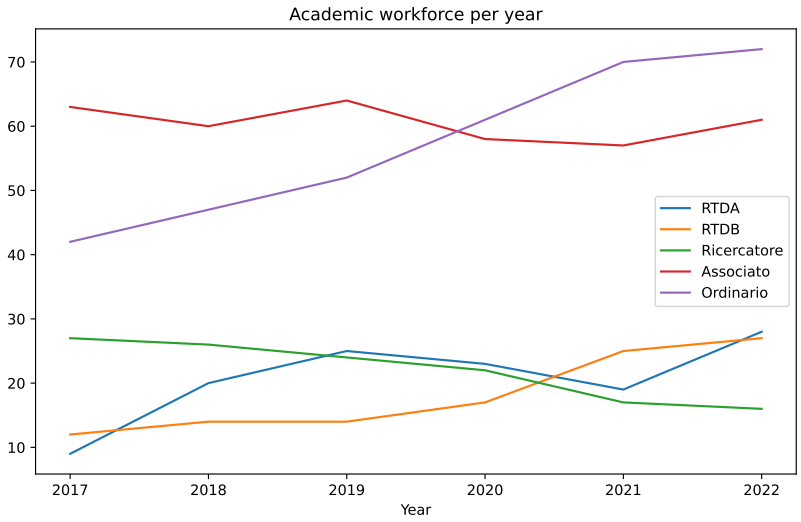

Figure 2: Aggregated number of people for each academic role in Univaq per year.

for each considered year. As can be seen in figure 2, the number of full and associated professors is much higher than the other roles, and we also observe a constant increase of full professors, which became, after the year 2020, even higher than associated professors. However, we can observe from figure 3 how the number of men is always higher than women in each considered role. In particular, looking again at associated and full professors, we notice how the difference is much higher than in other roles, with an average delta of more than 30 units for both roles.

The observations made for Univaq are confirmed by figures 4 and 5, which report instead the aggregated number of people and the number of men and women for each academic role in Italy. However, differently from Univaq, we observe how the number of associated professors is constantly higher than the number of full professors, while the number of researchers is still always lower. The delta between men and women is instead confirmed to be higher for each role, meaning that Univaq is in trend with Italian behavior.

Finally, we note how the role of *Researcher* (*Ricercarore* and *Ricercatore non confermato* in the charts) is constantly decreasing. This trend can be explained by a change in the Italian regulation, which has removed the *Researcher* role and replaced it with *RTD-A* and *RTD-B* roles.

While the visualizations depicted in this section show the population for Areas 1 and 9 with no further filters, for our experiments within UNIVAQ and Italy we selected only people that were consistently part of the University of L'Aquila and Italian University System, respectively, throughout all of the years in the reference

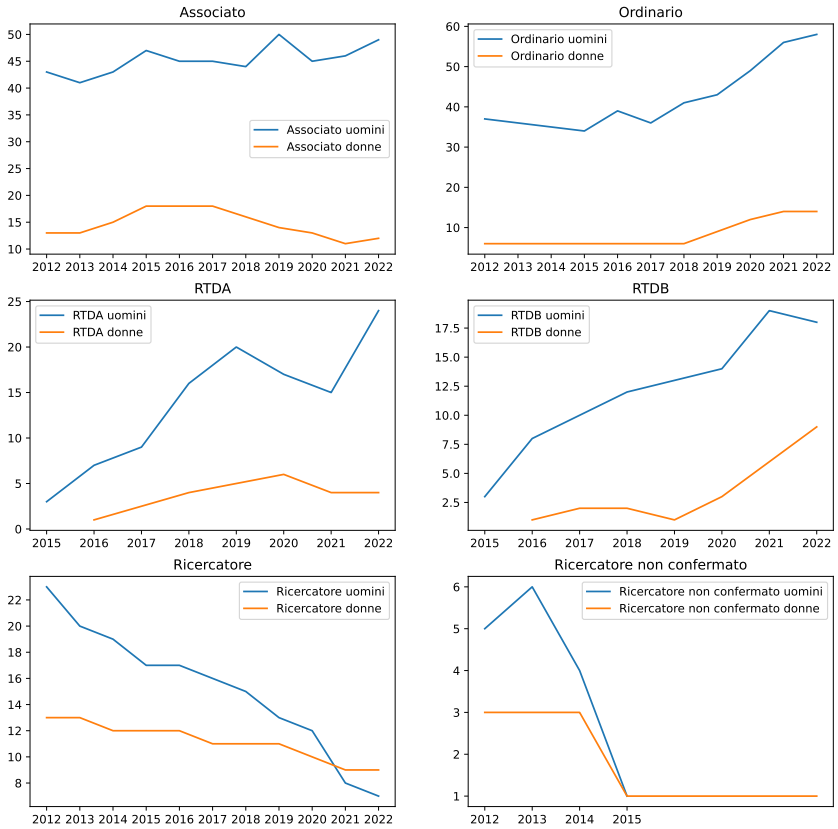

Figure 3: Men and women for each academic role in Univaq.

period. The datasets' sizes were thus reduced to around 60 records for UNIVAQ and 4.000 records for Italy.

2.2 Bias in Academic data

In this section, we start discussing the performed experiments. In particular, here we focus on the first couple of batches of experiments where no classifier was used. We computed the inherent bias present in the dataset through the means of Disparate Impact (DI). We recall that, in the case of fairness, the DI should have a value close to 1.

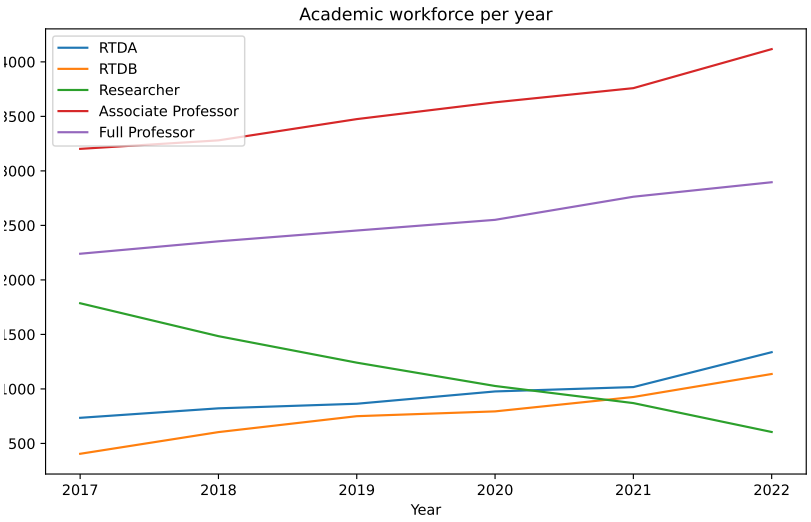

Figure 4: Aggregated number of people for each academic role in Italy, per year.

Figures 7 and 6 show results for the University of L'Aquila. As can be seen, the figures expose an opposite trend of DI. In particular, in figure 6, it can be seen how the DI decreases over the years, meaning an increase of bias between researchers and associated professors. This increase in bias can be explained by the fact that some associated professors became full professors over the years, as also highlighted in figure 7. In fact, the trend exposed in the latter figure highlights an increase in the DI over the years, meaning that the bias between men and women full professors has decreased.

Figures 9 and 8 depict instead the results for Italy as a whole. In this case, both trends are positive, meaning that gender equality has improved over the years. However, the level of bias between men and women full professors, although has improved over the years, is still quite high. In fact, the values represented in figure 9 indicate how the DI is increased by only around 0.02 units in five years, meaning that there is still bias between men and women full professors in Italy.

2.3 Algorithmic Bias in Local and National Academic System

In this section, we discuss later batches of experiments making use of a classifier to predict the academic role of each researcher based on their Scopus metrics. Because a classifier is now involved, we are able to monitor multiple metrics other than Disparate Impact (DI). In particular, we measure Equalized Odds (EO), Average Odds (AO), and Accuracy of the classifier.

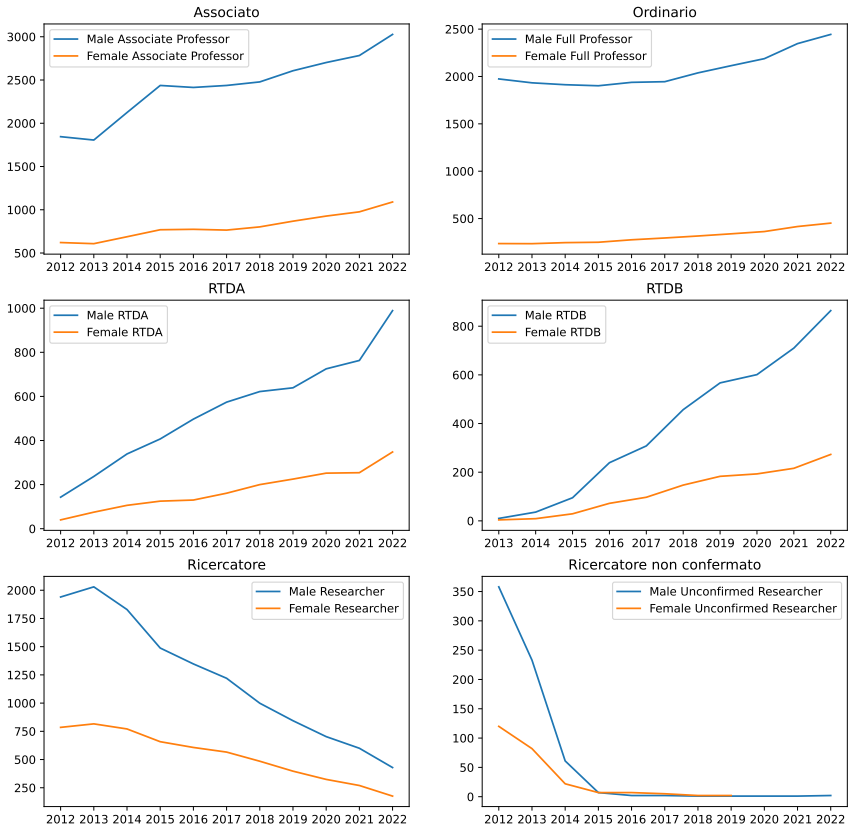

Figure 5: Men and women for each academic role in Italy.

Figures 10 and 11 depict the trend of these metrics for UNIVAQ. It must be noted that the size of the reference datasets for these batches of experiments since we only focused on the University of L'Aquila, was very small and thus unreliable for classification. The oscillation that this metrics show is certainly a byproduct of this issue. Nonetheless, the trend of the Disparate Impact metric is similar to the trend shown by the original dataset in section 2.2. Due to the low sample size, the classifier cannot reach high accuracy (0.7 at best), while the other metrics show an overall balance and fairness in both cases.

Figures 12 and 13 depict instead the metrics for the classifiers trained on much larger data, referring to Italy as a whole. As a result, the metrics' variation

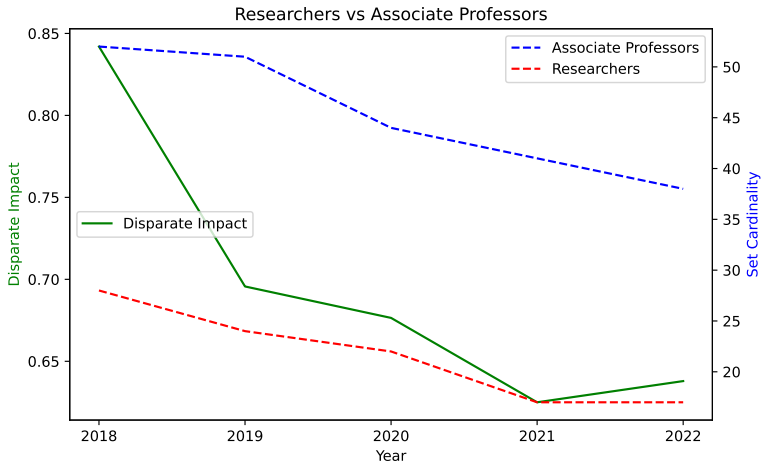

Figure 6: Yearly disparate impact within UNIVAQ between Researchers and Associate professors, no classifier.

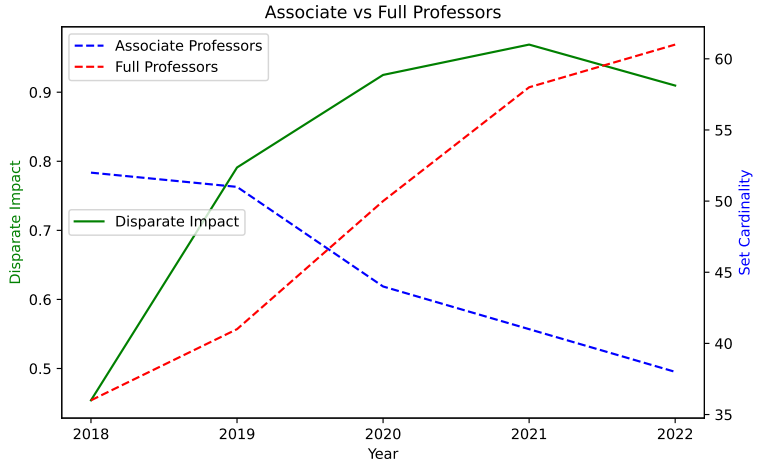

Figure 7: Yearly disparate impact within UNIVAQ between Associate and Full Professors, no classifier.

is much smaller, and the accuracy increases to a peak of slightly above 0.8 in the prediction between researchers and associated professors and is constantly higher than 0.7 in predicting associated and full professors.

A stark contrast between the two plots is instead illustrated in the bias metrics of the classifier's predictions. In particular, the comparison between Researchers and Associate Professors shown in figure 12 is reasonably fair, with DI close to

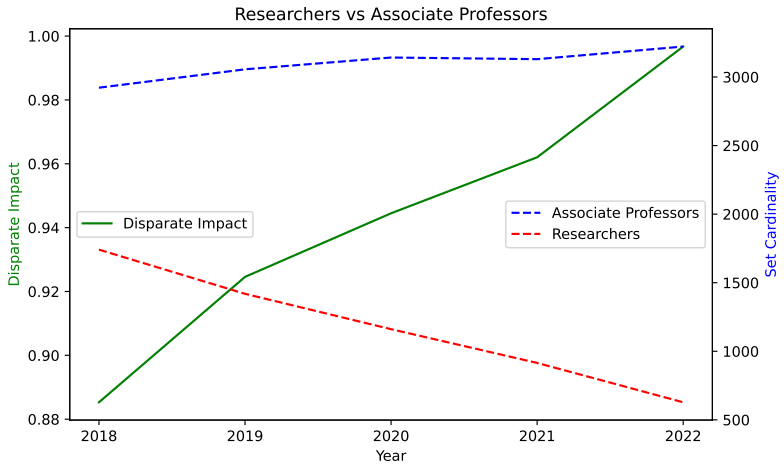

Figure 8: Yearly disparate impact within Italy between Researchers and Associate professors, no classifier.

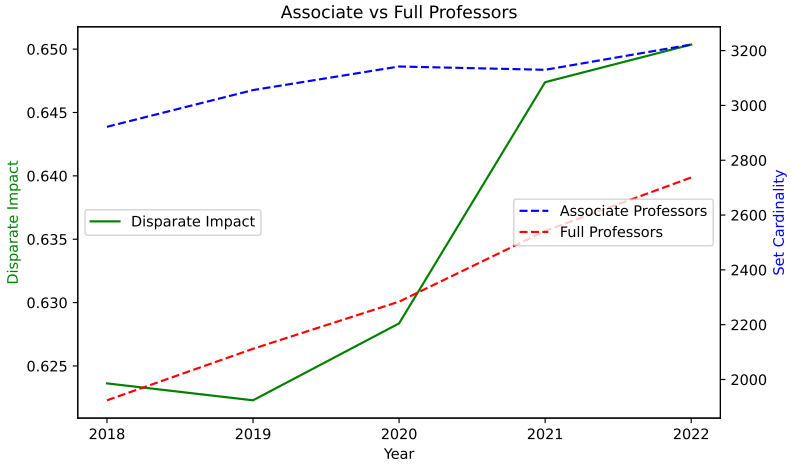

Figure 9: Yearly disparate impact within Italy between Associate and Full Professors, no classifier.

1 and AO being close to 0. The only exception is seen in the EO metric, which constantly decreases up to -0.2. This decrease can be explained by an increase in the number of male Associated professors over the years with respect to women, as also highlighted in figure 5. However, an increase in the accuracy and the AO almost constantly around 0 highlights how the classifier is still able to predict the role almost correctly for males and females. The classification among associated

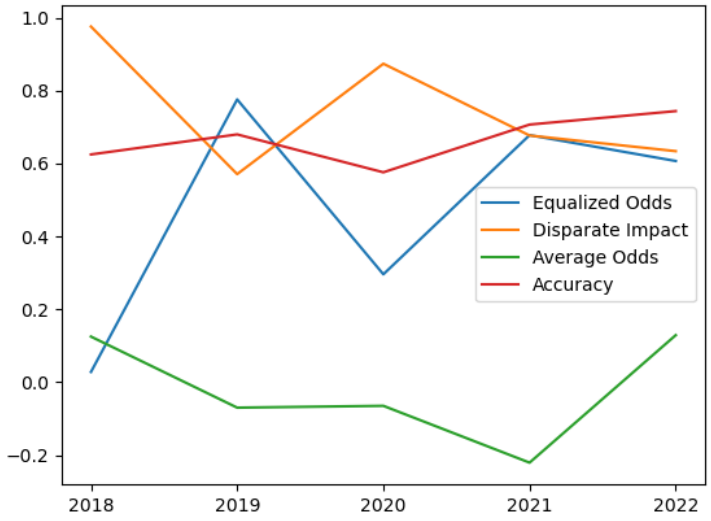

Figure 10: Yearly disparate impact within UNIVAQ between Researchers and Associate professors with Logistic Regression classifier.

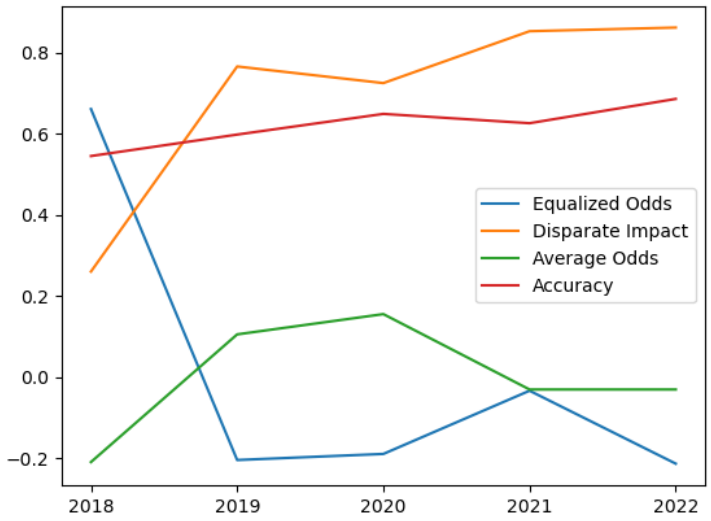

Figure 11: Yearly disparate impact within UNIVAQ between Associate and Full professors with Logistic Regression classifier.

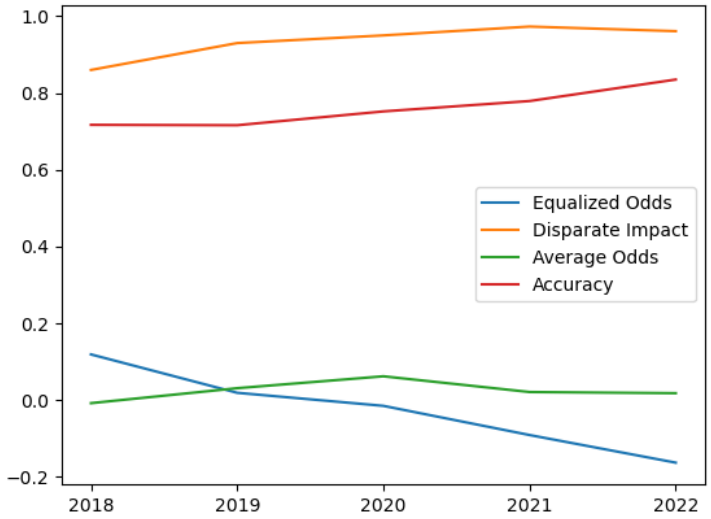

Figure 12: Yearly disparate impact within Italy between Researchers and Associate professors with Logistic Regression classifier.

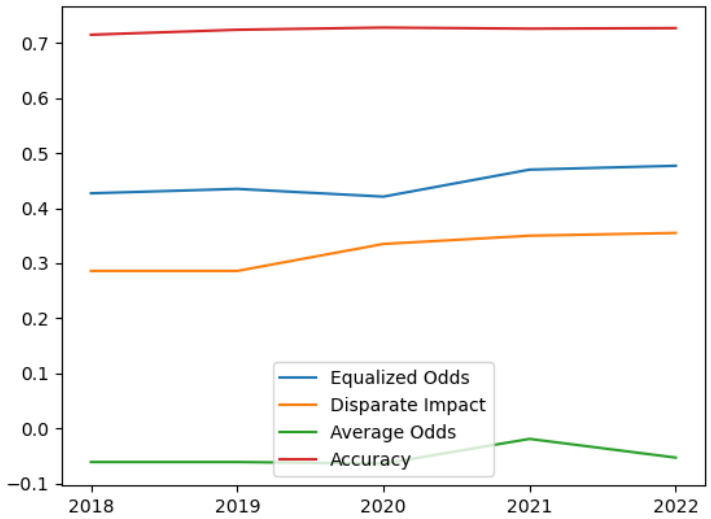

Figure 13: Yearly disparate impact within Italy between Associate and Full professors with Logistic Regression classifier.

and full professors shown in figure 7 exposes instead a higher bias, highlighted by a low Disparate Impact and high Equalized Odds.

2.3.1 Application of debiaser method

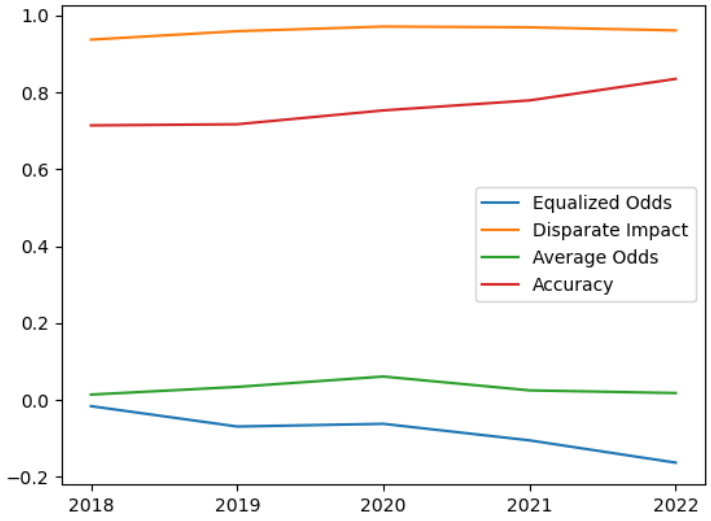

Figure 14: Yearly disparate impact within Italy between Researchers and Associate professors with DEMV and Logistic Regression classifier.

Figures 14 and 15 show the accuracy and fairness metrics for the Logistic Regression classifier trained on Italian data after the application of the *Debiaser for Multiple Variables (DEMV)* pre-processing method[1]. In both figures, it can be seen how the application of a debiaser method improves the fairness of the classifier. The improvement is much evident in figure 15 showing the classification between associated and full professors. The bias of the classifier trained with raw data was very high, with a DI around 0.3, an EO higher than 0.4, and an AO around -0.1 (see figure 9). After the application of DEMV, bias has been mitigated while always slightly present. In fact, the DI ranges around 0.6 and 0.7, EO is around 0.1, and AO is still around -0.1 but closer to 0.

[1]We did not apply DEMV to the UNIVAQ data since they where too few, making DEMV infeasible

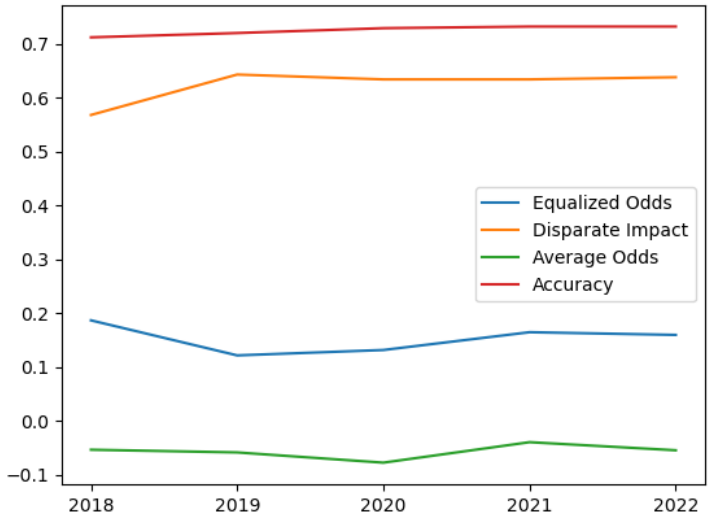

Figure 15: Yearly disparate impact within Italy between Associated and Full professors with DEMV and Logistic Regression classifier.

3 Discussion

From the experiments performed, we can derive the following conclusions.

While the gap between men and women in each role is staggering, this still refers only to areas 1 and 9, namely Science, technology, engineering, and mathematics. These areas are characterized by the low rate of women applications, starting from the first year of their respective bachelor's degrees. So, while this is a worrisome trend that needs to be tackled with urgency, the plots in 5 may not, by themselves, depict an inherent sex bias in the selection criteria for academic promotions in general.

The trend of academic promotions in UNIVAQ tends to be in line with the Italian one. In particular, the trend of academic promotions between Associated and Full professors exposes a higher bias with respect to the trend of promotions between researchers and associated professors. This bias can be explained by a difference in the number of men and women full professors. However, it is worth noting how, in selecting UNIVAQ data, we only selected people that were in UNIVAQ for all the considered range of years. Hence, we did not consider new acquisitions or people leaving UNIVAQ.

A classifier trained with such data can inherit this bias, and so, for instance, predict a male as a full professor with a higher probability than a woman. In

particular, we show how a Logistic Regression classifier trained with Italian Associated and Full professor data can be biased against women in predicting them as Full professors. However, applying a debiaser method (such as DEMV) can help reduce this bias and make the role prediction fairer.

4 Future works

The main goal of this study was to locate and analyze inherent bias in the selection process of promotions within the academic environment of both UNIVAQ and Italy. To achieve this goal, we used classifier methods to predict the academic role of each researcher. Thus, future works can be reasonably split into two lines of work: expanding the datasets' size, and further improving the classifiers we built. The datasets used for the training can be expanded upon with further academic performance metrics, such as the scores of the journals that published a researcher's work or the number of years of activity in which a researcher has actively published papers or has been cited. Moreover, rather than narrowing our attention to the combined Areas 1 and 9, we might collect data from many scientific fields and analyze the bias metrics between them.

By improving and extending the training datasets, we also aim to improve the classifier's accuracy, which in turn would increase the trustworthiness of the metrics we computed. Although a classifier based on logistic regression was utilized for this investigation, we may choose to substitute or complement the study with a more complex or appropriate model. Furthermore, we can analyze the classifier's feature importance with the aim of assessing which features are the most important ones for classification, which could potentially be another indicator of bias.

We find that tracking inherent bias in the selection process for such critical roles is essential. Rather than being a one-time study, this endeavor intended to establish the groundwork for a stable, growing system. Further work on the classifier would lead to a stable, more fine-tuned model for academic role prediction. Said classifier, after thorough testing, could potentially be used by institutions as an aid during the selection process or as a self-evaluation tool for aspiring professors and researchers alike.

References

[1] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[2] Giordano d'Aloisio, Andrea D'Angelo, Antinisca Di Marco, and Giovanni Stilo. Debiaser for multiple variables to enhance fairness in classification tasks. *Information Processing and Management*, 60(2), 2023. All Open Access, Hybrid Gold Open Access.

[3] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, Sydney NSW Australia, August 2015. ACM.

[4] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[5] G.H. Rosenfield and K. Fitzpatrick-Lins. A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetric Engineering and Remote Sensing*, 52(2):223–227, 1986.