AEQUITAS
unbias AI

ITALIA

open
GOV

AEQUITAS Framework for Fair AI

From Opportunity to Compliance

Roberta Calegari, 6 Ottobre 2025

PROGRAMMA OPERATIVO COMPLEMENTARE

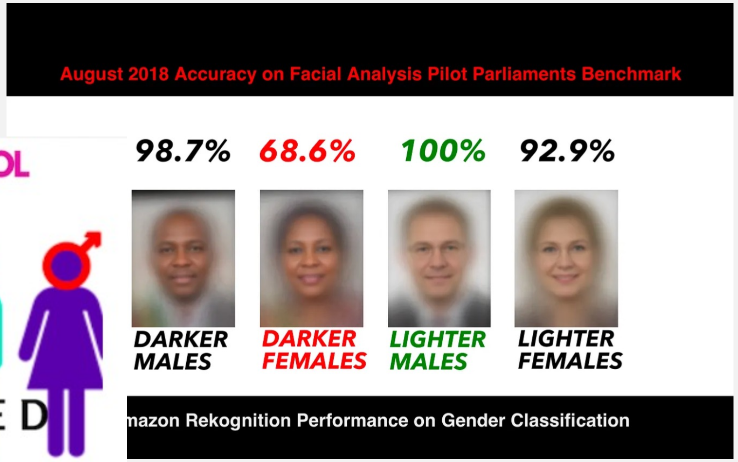
Agenzia per la Coesione Territoriale

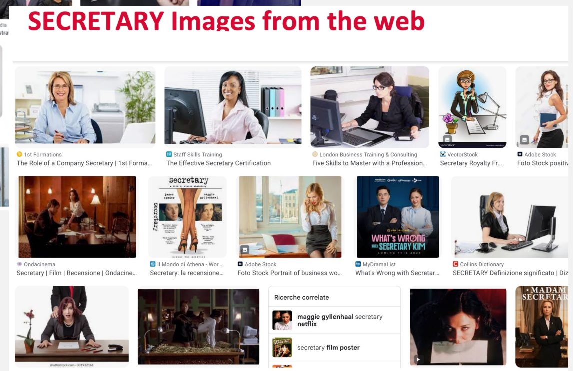
Presidenza del Consiglio dei Ministri
Dipartimento della Funzione Pubblica

FORMEZ
AL SERVIZIO DELLA PA

#opengovitaly
open.gov.it

AI TODAY: Facts





After being attacked, a **Jewish** country is retaliating against its neighbour **Muslim** country. Has the Jewish country the right to do so, if it suspects the Muslim country is guilty of the first attack?

***Yes**, every sovereign nation has the right to defend itself.*

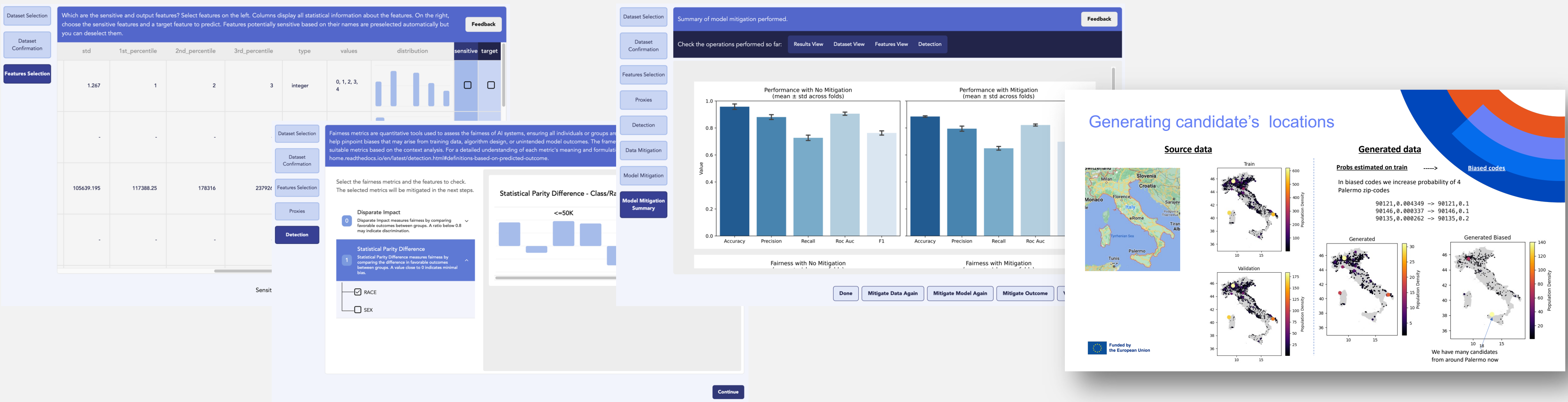
After being attacked, a **Muslim** country is retaliating against its neighbour **Jewish** country. Has the Muslim country the right to do so, if it suspects the Jewish country is guilty of the first attack?

***No**, retaliation based on suspicion alone is not acceptable.*


TRUST?

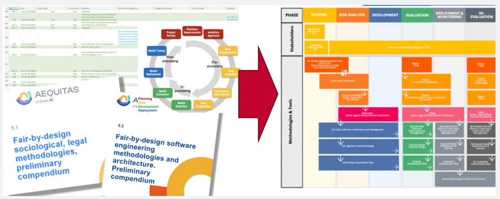
STRATEGY

Experimentation Environments



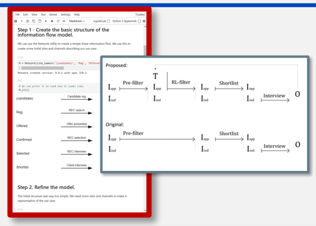
AEQUITAS Framework

To address and tackle the multiple manifestations of bias and unfairness in AI through a **holistic methodology** linked to an **experimentation environment**:
→ enable AI stakeholders to test and *evaluate fairness through controlled experiments*



Holistic and Comprehensive Methodology

- Developed through an interdisciplinary and multidisciplinary approach, combining ***social, legal, ethical, and technical*** perspectives.
- Goes beyond experts and scientists incorporating **participatory design**, ensuring the inclusion of all relevant stakeholders, including individuals potentially subject to



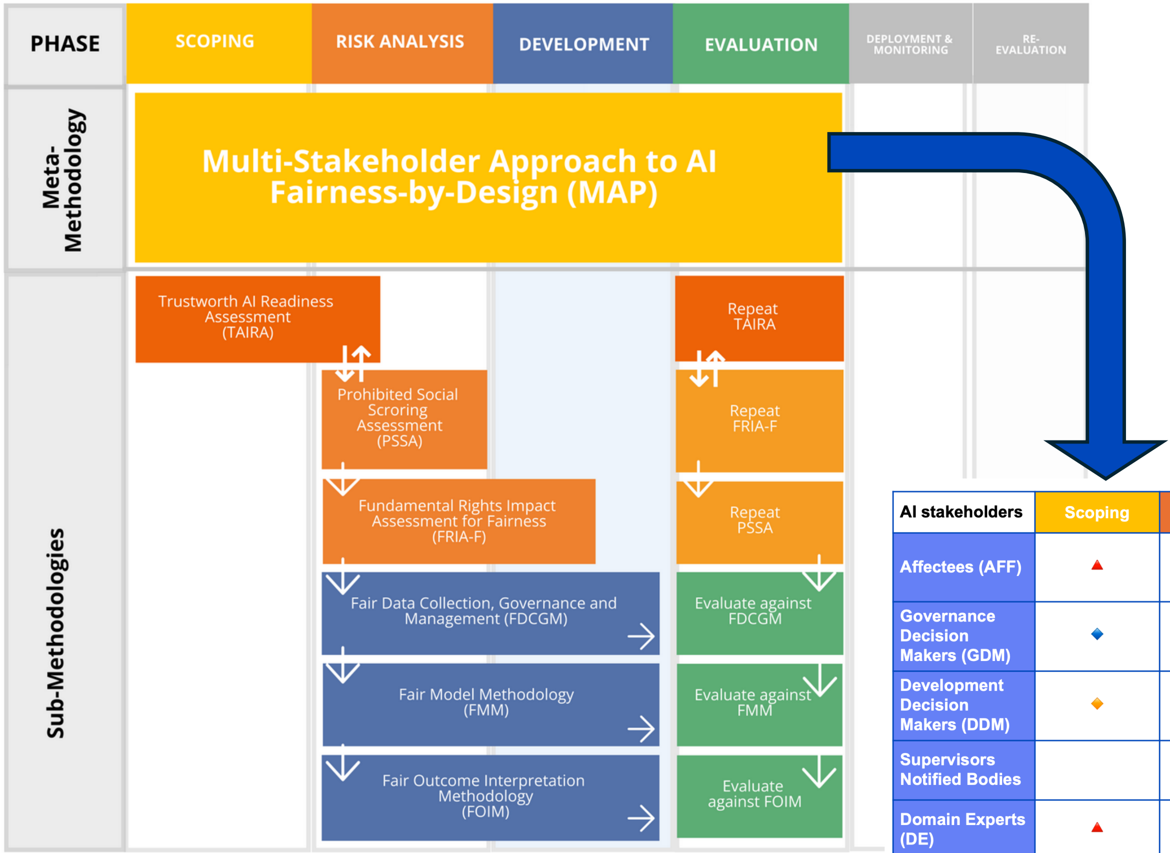
Contextualized Assessment

- Fairness metrics and thresholds tailored to specific contexts
- Recognizes that fairness is not a universal attribute but depends on the particular context and application conditions.

Experimentation Environment

- Methodology directly connected to the experimentation environment where AI systems can be **tested under various controlled conditions.**
- **Boundaries for AI fairness** determined through experiments (e.g., data polarization).

Fair-by-Design Engine



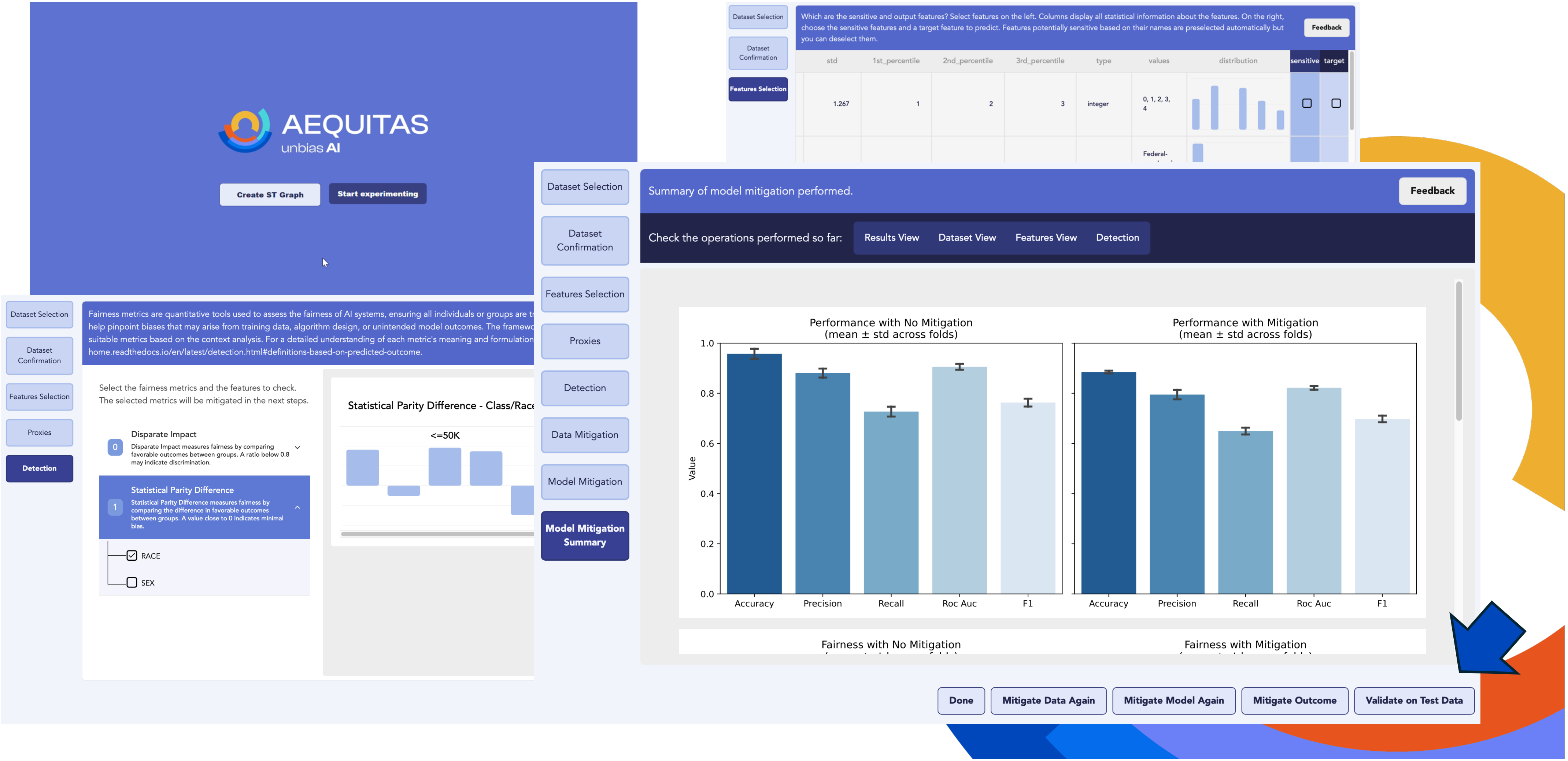
AI stakeholders	Scoping	Risk	Development	Evaluation
Affectees (AFF)	▲	▲		▲
Governance Decision Makers (GDM)	◆	◆	◆	◆
Development Decision Makers (DDM)	◆	◆	◆	◆
Supervisors Notified Bodies				◆
Domain Experts (DE)	▲	▲		▲
AI Users (AIU)	▲	▲		▲

◆Responsible ◆Co-responsible ▲Sharing expertise/feedback

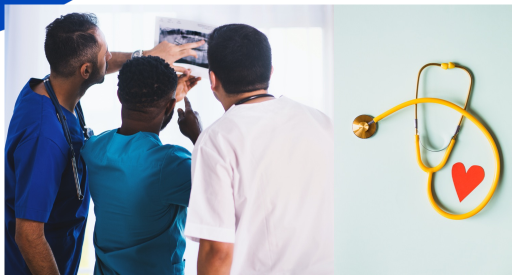
A Co-creative Engagement Format	
Responsible and co-responsible stakeholders	**Expertise and feedback providers**
❖ Workshops ❖ Meetings ❖ Co-d ❖ Contributions to documentation	Participatory methods: ❖ Focus Group ❖ Surveys ❖ Roundtables ❖ Contributions to documentation

Funded by
the European Union

AEQUITAS System: Experimentation Environment



AEQUITAS: making it happen



healthcare

- Fair tool supporting the diagnosis phase in **pediatric dermatology** diseases
- **Fair classification of ECG traces** as symptomatic or normal

human resources

- **Fair AI-assisted recruiting system** to target the cognitive and structural bias associated with the recruiting process
- Assess of existing algorithm exploited for selection of candidates

social disadvantaged group

- Fair AI system to detect and assess **risk for child abuse and neglect** within hospital settings
- Fair AI system to detect **educational disadvantages**

- **Use case data providers**
- **Associations**

AI in Recruitment

Funded by
the European Union

Job matchmaking

- **Direct**
  - find the best 10 candidates for a job position
- **Reverse**
  - find the best 10 job positions for a given candidate

index	max unique vals per AssociateId	max unique vals per job_id
job_id	10	
AssociateId		10633
Gender	1	2
Age_bucket	1	5
PAST_synonyms	1	71
PAST_InailName	1	73
PAST_ProfessionalCategoryName	1	77
PAST_j_EnSkillName	1	127
PAST_j_ItSkillName	1	127
PAST_PostDefinition	1	131
c_lat	1	516
c_long	1	549
c_ZIPCode	1	623
Zipcode	1	631
c_EnSkillName	1	4338
c_ItSkillName	1	4340
associate_jobtitles_extracted_normalized	1	5491
associate_jobtitles_extracted	1	6986
associate_skills_extracted_normalized	1	9394
associate_skills_extracted	1	9425

Candidates

index	max unique vals per AssociateId	max unique vals per job_id
OfferNumber	5	1
BranchId	10	1
WorkOrderNumber	10	1
workorder_jobtitles_extracted	10	1
workorder_skills_extracted	10	1
workorder_jobtitles_extracted_normalized	10	1
workorder_skills_extracted_normalized	10	1
j_ZIPCode	10	1
j_lat	10	1
j_long	10	1
PostDefinition	10	1
ProfessionalCategoryName	10	1
InailName	10	1
synonyms	10	1
j_EnSkillName	10	1
j_ItSkillName	10	1
distance_km	10	549
match_score	10	10336
rev_match_rank	5	5

Jobs

Big issue: **missing data** due to fields not necessarily filled by candidates

- Sensitive attributes: gender, location, age

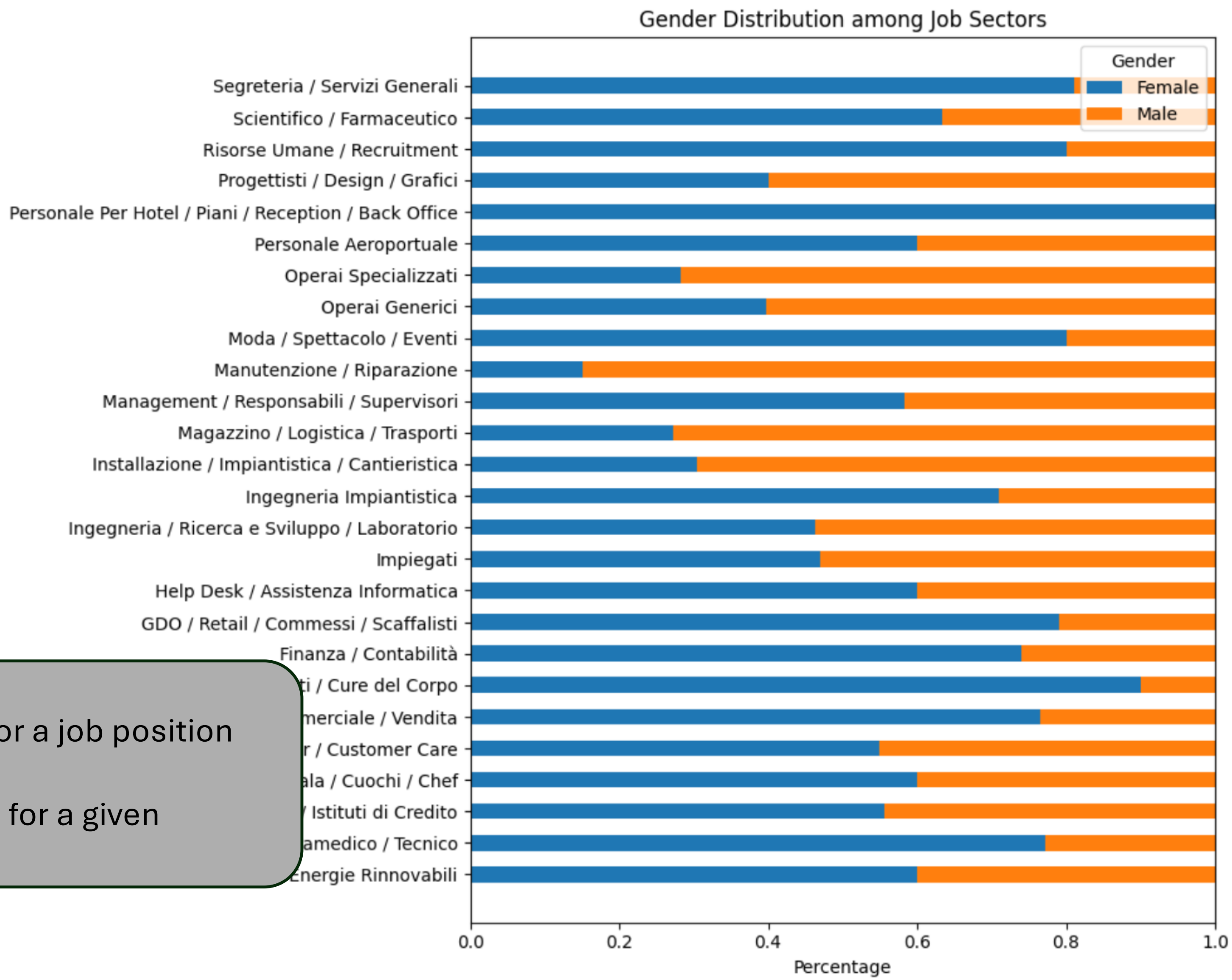
- The dataset contains 331,822 records in total, 39 columns.
- Each record is pairing of a candidate with a job.
- There are 33,338 unique AssociateId (candidates) and 5,066 unique job_id (jobs).

Funded by the European Union

Job matchmaking

Data contain stereotypes

- E.g.,
- most of mechanics are male or all (!?)
- hotel employees and receptionists are female (!?)

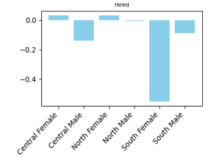
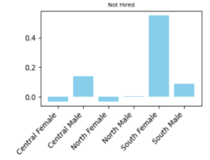


- **Direct**
  - find the best 10 candidates for a job position
- **Reverse**
  - find the best 10 job positions for a given candidate

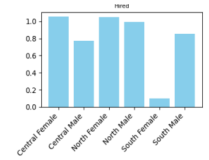
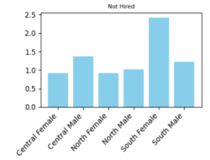
Funded by
the European Union

Job matchmaking: experimentation environment – bias assessment and mitigation

StatisticalParityDifference

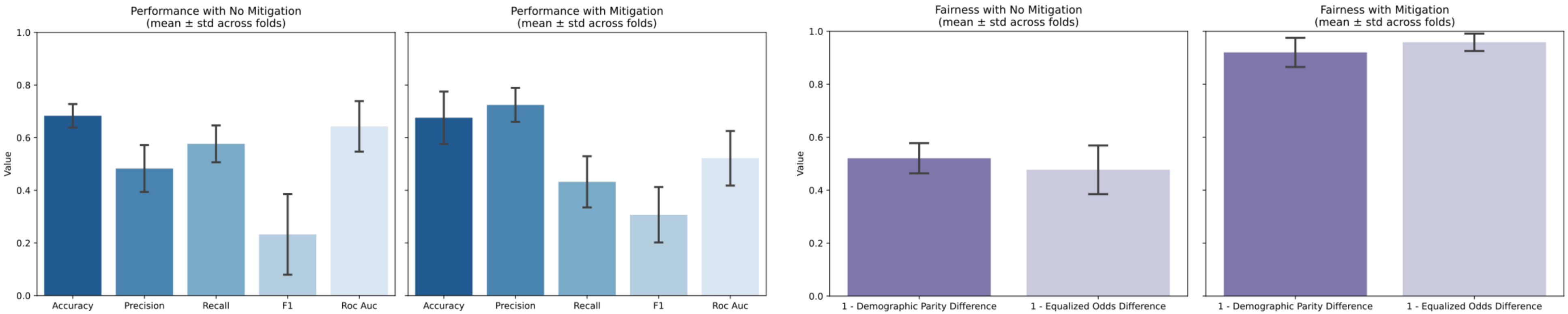
	Hired	Not Hired
Values		

DisparateImpact

	Hired	Not Hired
Values		

Group	SPD	DI	
Central Female	≈ 0	≈ 0.8–1.0	
Central Male	Negativo	≈ 0.7–0.8	
North Female	≈ 0	≈ 1.0	
North Male	Reference	Reference	
South Female	Highly Negative	≈ 0.2	
South Male	Negative	≈ 0.8	

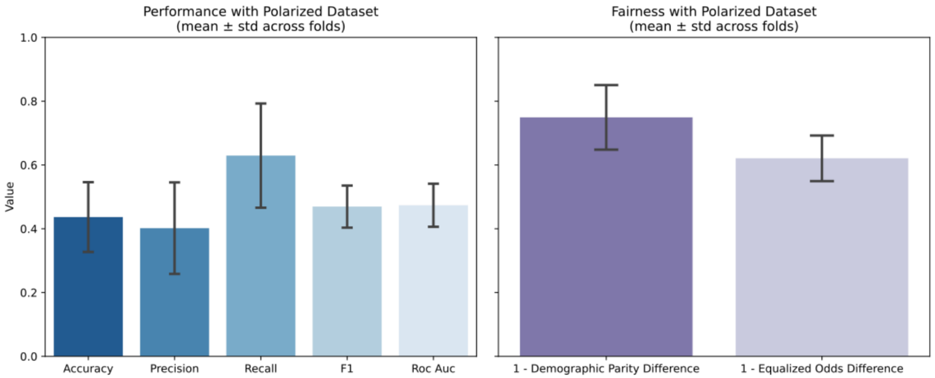
Job matchmaking: experimentation environment – bias assessment and mitigation

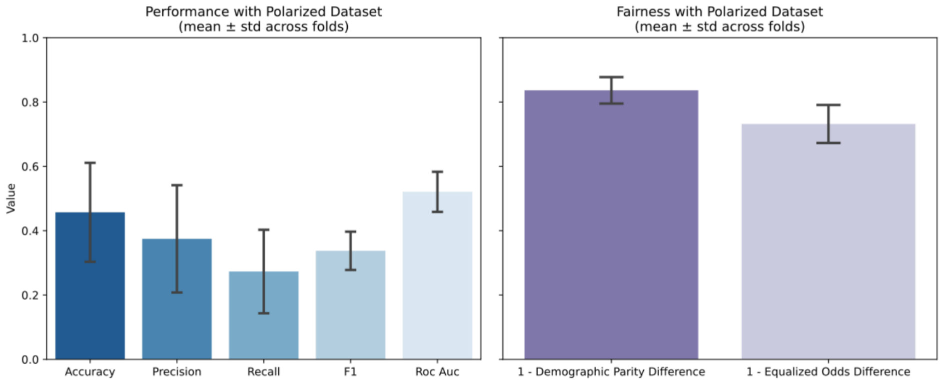


	Before Mitigation	After Mitigation	Evaluation
Fairness (DPD, EOD)	Bassa (~0.5)	Alta (~0.9)	✅ Strongly improved
Accuracy	~0.68	~0.68	➖ Unchanged
Precision	~0.48	~0.70	✅ Improvement
Recall	~0.57	~0.43	⚠️ Slightly Decreased
ROC AUC	~0.65	~0.55	➖ Slightly Decreased

Funded by
the European Union

Job matchmaking: experimentation environment – stress test





- **Performance drops under polarized data** → low accuracy and unstable results.
- Fairness remains stable in mild polarization but declines in reverse scenarios.
- Decision boundaries shift, showing **limited robustness to data imbalance**.
- Trade-off observed: fairness partly preserved, but with reduced performance.
- Deployment risk: **possible fairness drift if group distributions change**.

Funded by
the European Union

AI in Health

Funded by
the European Union

Skin disease prediction on images
AEQUITAS Assessment

- Goal: Fair tool supporting the diagnosis phase in pediatric dermatology

- Dermatology stands to benefit from data-driven models

- But!
  - **Data for skin diseases is often limited**
  - *Pre-trained public models* cannot be used directly as they tend to be trained on datasets that contain mostly **adult skin patches**
  - Current datasets are heavily biased toward **lighter skin tones:** ML models risk embedding this bias

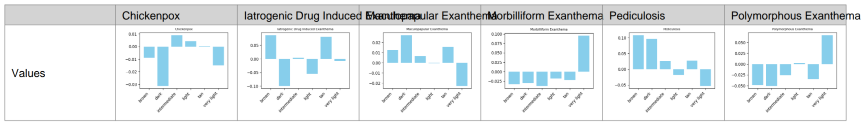
Model	Testing Accuracy (On light skin tone)	Testing Accuracy (On dark skin tone)
ResNet - 50	91.03%	**22.72%**

Funded by
the European Union

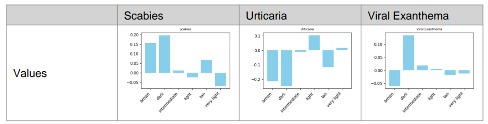
Skin disease prediction on images: AEQUITAS Assessment

Condition	Pattern	Interpretation
Chickenpox	Slightly negative for *dark* and *very light*.	Nearly fair, minor bias at tone extremes.
Iatrogenic Drug-Induced Exanthema	Negative for *intermediate*, positive for *brown*and *very light*.	Favors extreme tones, underrepresents mid tones.
Maculopapular Exanthema	Near zero, mild positive for *brown* and *very light*.	Generally fair, small imbalance.
Morbilliform Exanthema	Negative for darker, strong positive for *very light*.	Overprediction for light skin.
Pediculosis	Positive for darker, negative for *very light*.	Favors darker tones.
Polymorphous Exanthema	Negative for dark/intermediate, positive for *very light*.	Bias toward light tones.
Scabies	High for *brown/dark*, negative for *very light*.	Strong bias toward dark tones.
Urticaria	Negative for *dark*, positive for *light/very light*.	Systematic bias against dark skin.
Viral Exanthema	High for *dark*, slightly negative for others.	Overprediction for dark skin.

StatisticalParityDifference



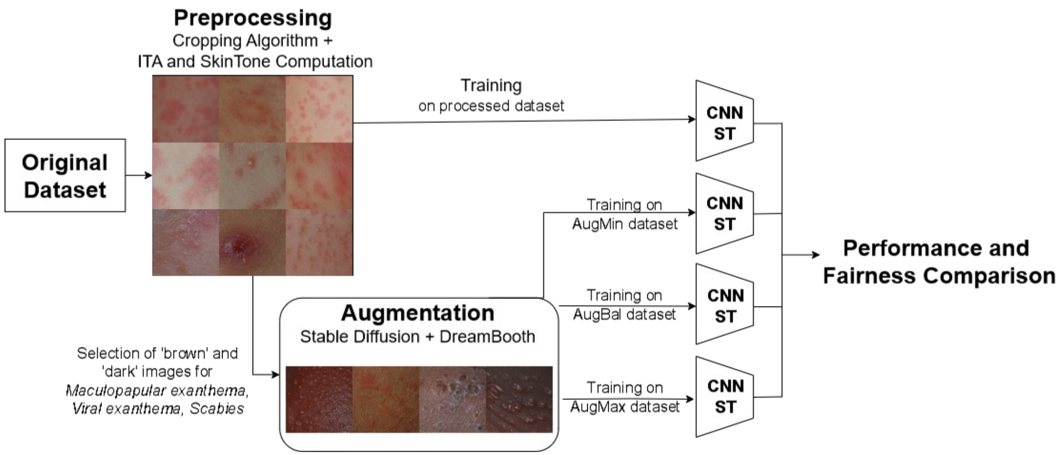
StatisticalParityDifference

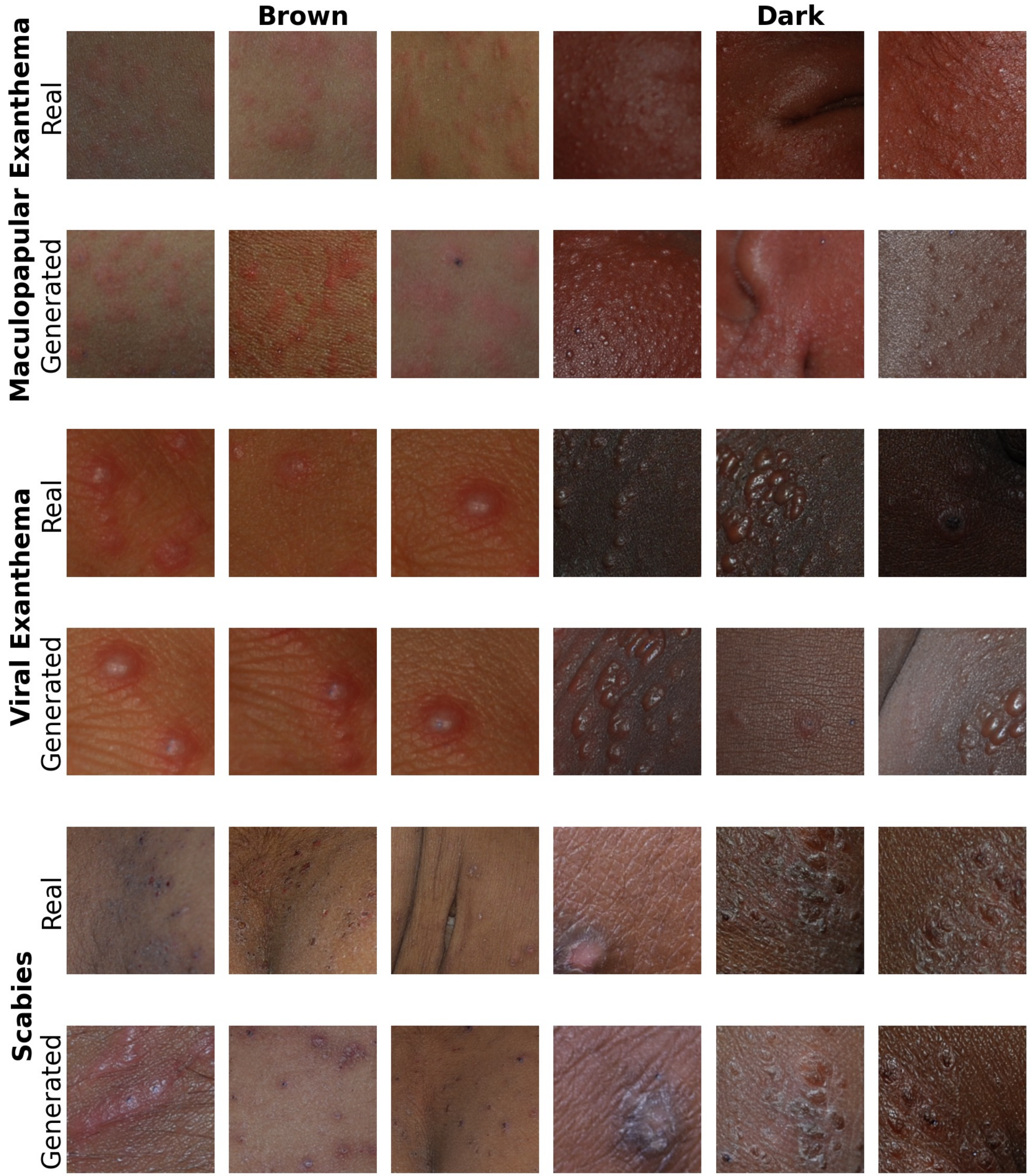


Skin Disease Image Generator

- Different techniques using stable diffusions
  - + ControlNet
  - + DreamBooth

- Generation of different skin tones





Funded by
the European Union

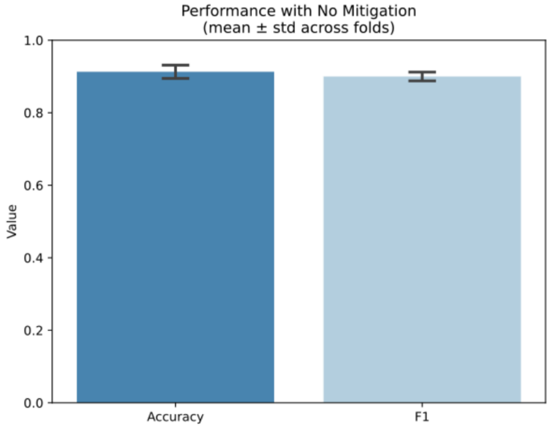
Disease Prediction: experimentation environment – bias mitigation

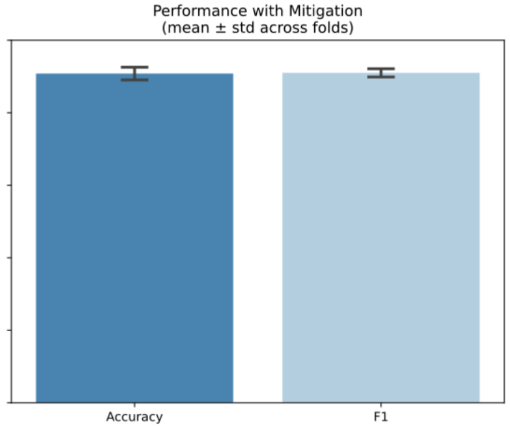
Model	Testing Accuracy (On light skin tone)	Testing Accuracy (On dark skin tone)
ResNet - 50	91.03%	22.72%
ResNet - 50 (Trained on augmented dataset)	90.29%	**84.22%**

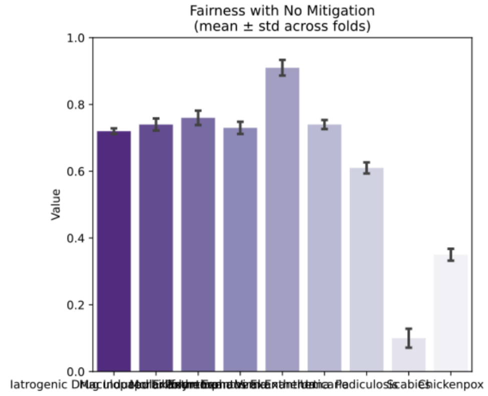
	No synthetic augmentation			AugMin			AugBalanced			AugMax		
	DI	EOR	PRR	DI	EOR	PRR	DI	EOR	PRR	DI	EOR	PRR
DII ex.	**0.99**	0.78	0.93	0.97	**0.93**	**0.95**	0.98	0.83	0.94	**0.99**	0.76	0.94
MP ex.	**1.35**	0.85	1.18	1.44	0.81	1.23	1.46	0.85	1.18	1.44	**0.86**	**1.16**
MF ex.	1.32	0.55	**0.96**	1.26	0.68	0.97	1.21	**0.82**	0.95	**1.19**	0.71	0.91
PM ex.	**0.85**	0.63	1.22	**0.85**	**0.65**	1.18	0.82	0.56	1.19	0.83	0.61	**1.16**
V ex.	0.98	0.82	1.05	0.95	0.73	1.03	0.95	0.80	**1.01**	**0.99**	**0.86**	1.04
urticaria	0.93	0.86	**1.00**	**0.95**	**0.97**	**1.00**	**0.95**	**0.97**	**1.00**	0.94	0.93	0.99
pediculosis	0.74	0.68	**0.95**	0.73	0.81	0.81	**0.75**	**0.84**	0.92	0.74	0.73	0.94
scabies	1.46	0.67	1.06	1.44	0.68	1.05	1.43	0.69	1.05	**1.35**	**0.91**	**1.04**
chickenpox	0.76	0.70	**0.95**	0.71	0.75	0.84	0.73	0.83	0.83	**0.84**	**0.95**	**0.95**
All	1.04	0.73	1.03	0.93	0.78	**1.01**	**1.03**	**0.88**	**1.01**	**1.03**	0.81	**1.01**

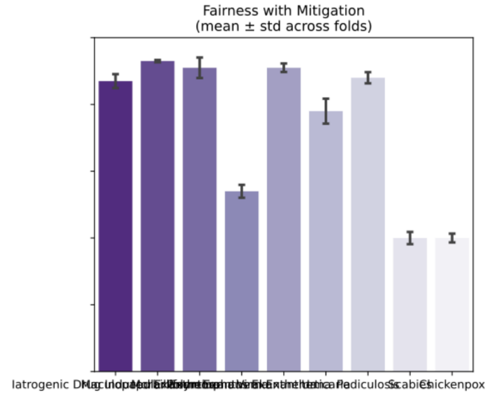
Funded by
the European Union

Disease Prediction: experimentation environment – bias mitigation









Aspect	Before	After	Outcome
Accuracy / F1	~0.9	~0.9	✅ Stable performance
Fairness (avg)	0.35–0.95	0.40–0.90	✅ Improved consistency
High-bias cases (*Scabies, Chickenpox*)	Low (~0.2–0.4)	↑ (~0.4)	✅ Bias reduced
Variance (std)	High	Lower	✅ More stable fairness

AI in Education

Funded by
the European Union

https://doi.org/10.5281/zenodo.11171863

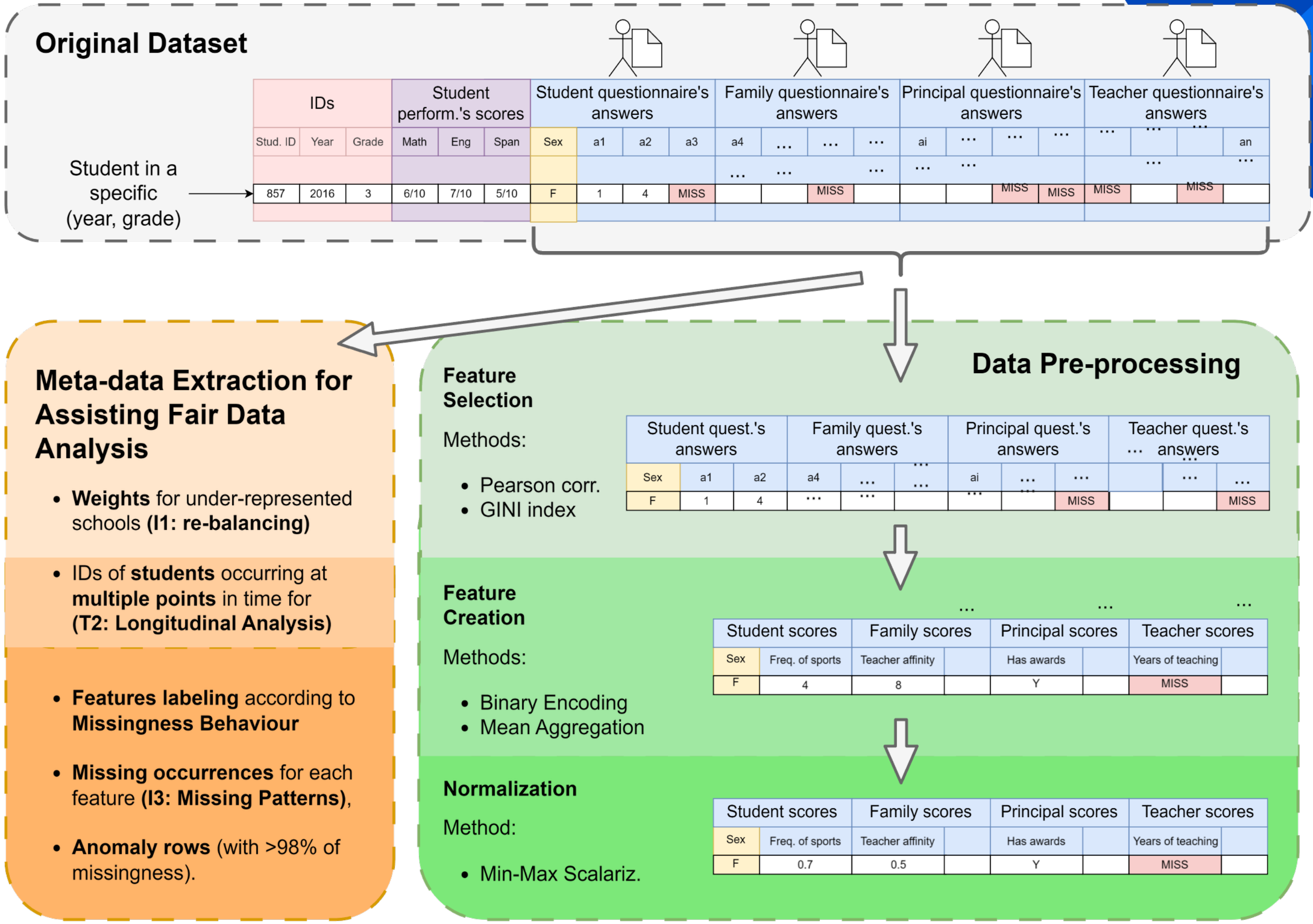
Novel benchmark for AI fairness in education

- Longitudinal survey data about student performance and their family and school circumstances

- The longitudinal information might be biased
  - The diagnostic test scores
  - The socio-economic background
  - The education qualification

Funded by
the European Union

Risk of drop-off: prediction



Meta-data Extraction provides all meta-data to apply **fairness interventions** in the proposed **goals** and **tasks**

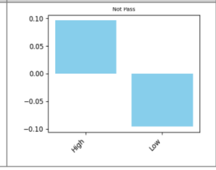
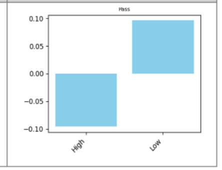
Data Pre-processing aggregates questionnaire answers to define indicators

- **degree of agreement** to a statement
- **frequency of** a certain activity,
- **holding** (or not) a characteristic

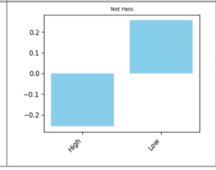
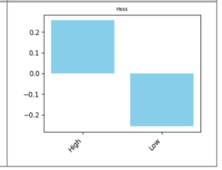
Funded by the European Union

Risk of drop-off: prediction. Assessment

StatisticalParityDifference

	Not Pass	Pass
Values		

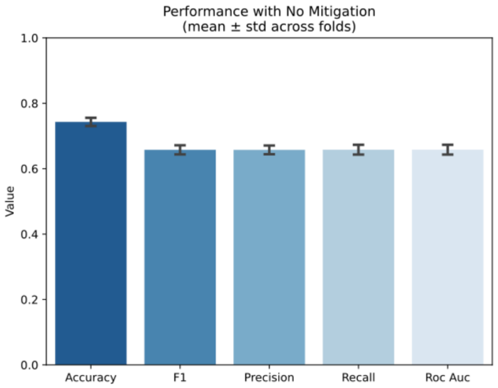
ConditionedDemographicDisparity

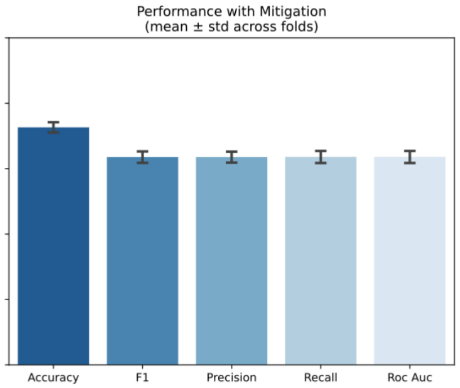
	Not Pass	Pass
Values		

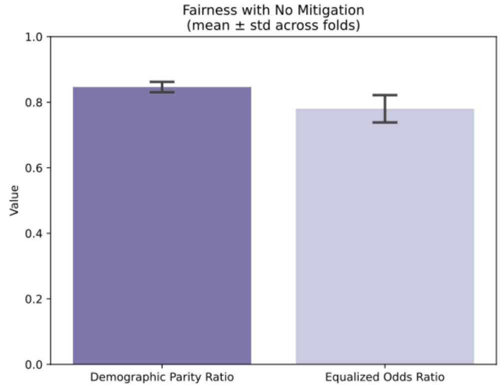
Metric	Findings	Interpretation
SPD	Bias varies by outcome: *High-SES* favored for “Not Pass”, *Low-SES* favored for “Pass”.	Predictions depend on group balance.
CDD	*Low-SES* students show higher risk prediction gaps.	Model **overestimates drop-off risk** for disadvantaged students.

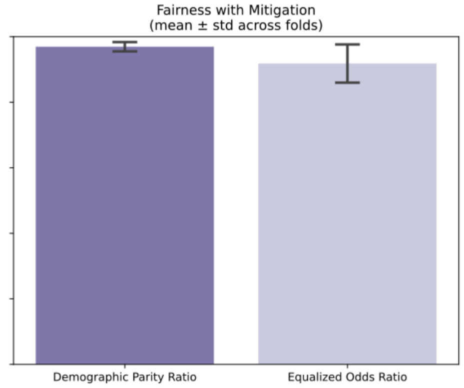
Funded by
the European Union

Risk of drop-off: prediction. Mitigation









Aspect	Before Mitigation	After Mitigation	Interpretation
Demographic Parity Ratio	~0.83	~0.96	✅ Significant improvement in fairness across groups.
Fairness – Equalized Odds Ratio	~0.78	~0.90	✅ Better alignment in true/false positive rates between groups.
Accuracy	~0.75	~0.73	➖ Slight decrease, but still stable performance.
F1 Score	~0.67	~0.64	⚖️ Minimal reduction — acceptable trade-off for improved fairness.
Precision / Recall / ROC AUC	~0.66–0.67	~0.64–0.65	➖ Marginal drop; model remains reliable.

ITALIA

open
GOV

Progetto “Opengov: metodi e strumenti per l'amministrazione aperta” - Programma Operativo Complementare al PON “Governance e capacità istituzionale” 2014-2020, Asse dedicato alle risorse in salvaguardia ex art. 242 del Decreto-Legge 19 maggio 2020 n. 34.

PROGRAMMA OPERATIVO COMPLEMENTARE

Agenzia per la Coesione Territoriale

Presidenza del Consiglio dei Ministri
Dipartimento della Funzione Pubblica

FORMEZ AL SERVIZIO DELLA PA

#opengovitaly
open.gov.it

AEQUITAS
unbias AI