

Egalitate algoritmică în sisteme automate de luare de decizii

Marius Miron
www.mariusmiron.com

Earth Species Project
Universitat Pompeu Fabra, Barcelona

Why machine learning may lead to unfairness

Songül Tolan¹, **Marius Miron¹**, Emilia Gomez^{1,2}, Carlos Castillo²

¹European Commission's Joint Research Centre

²Universitat Pompeu Fabra

Învățare automată, procesare de semnal

Ușor de formulat matematic

Focus pe rezolvarea de probleme

Rezultatele negative nu se publică

Formularea problemei, Rezolvare, Evaluare, Repetă

Ştiinţe sociale

Dificil de formulat matematic

Formulările/datele reprezintă procese sociale complexe

Datele sunt rezultatul unor procese care nu mai pot fi observate

Datele sunt reprezentări fluctuante, incomplete ale realităţii

E important să înțelegem de ce nu funcţionează un sistem

Învățare automată pentru luare de decizii



Kahneman - gândire rapidă și gândire lentă

System 1

Fast
Intuitive
Associative
Unconscious



System 2

Slow
Logical
Lazy
Takes effort

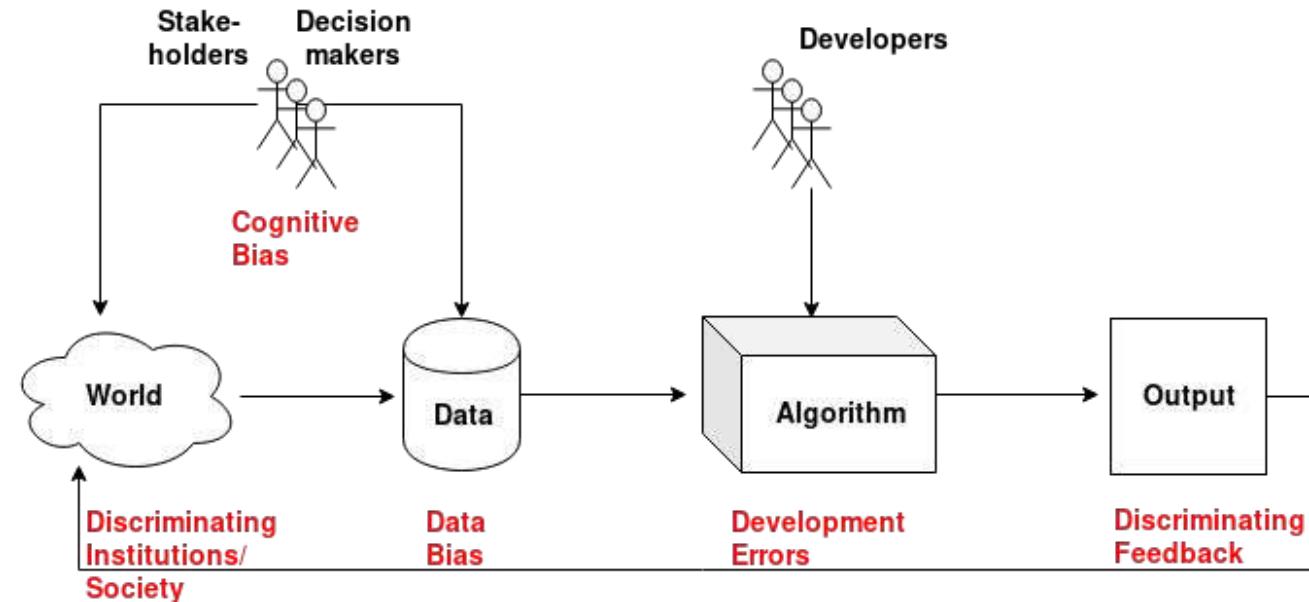
Justiția penală



**Compromis:
performanță - egalitate socială
(ne-descriminare)**

**Discriminare sistemică:
Modelul M discriminează
statistic grupul X în comparație
cu grupul Y**

Discriminarea algoritmică - cauze



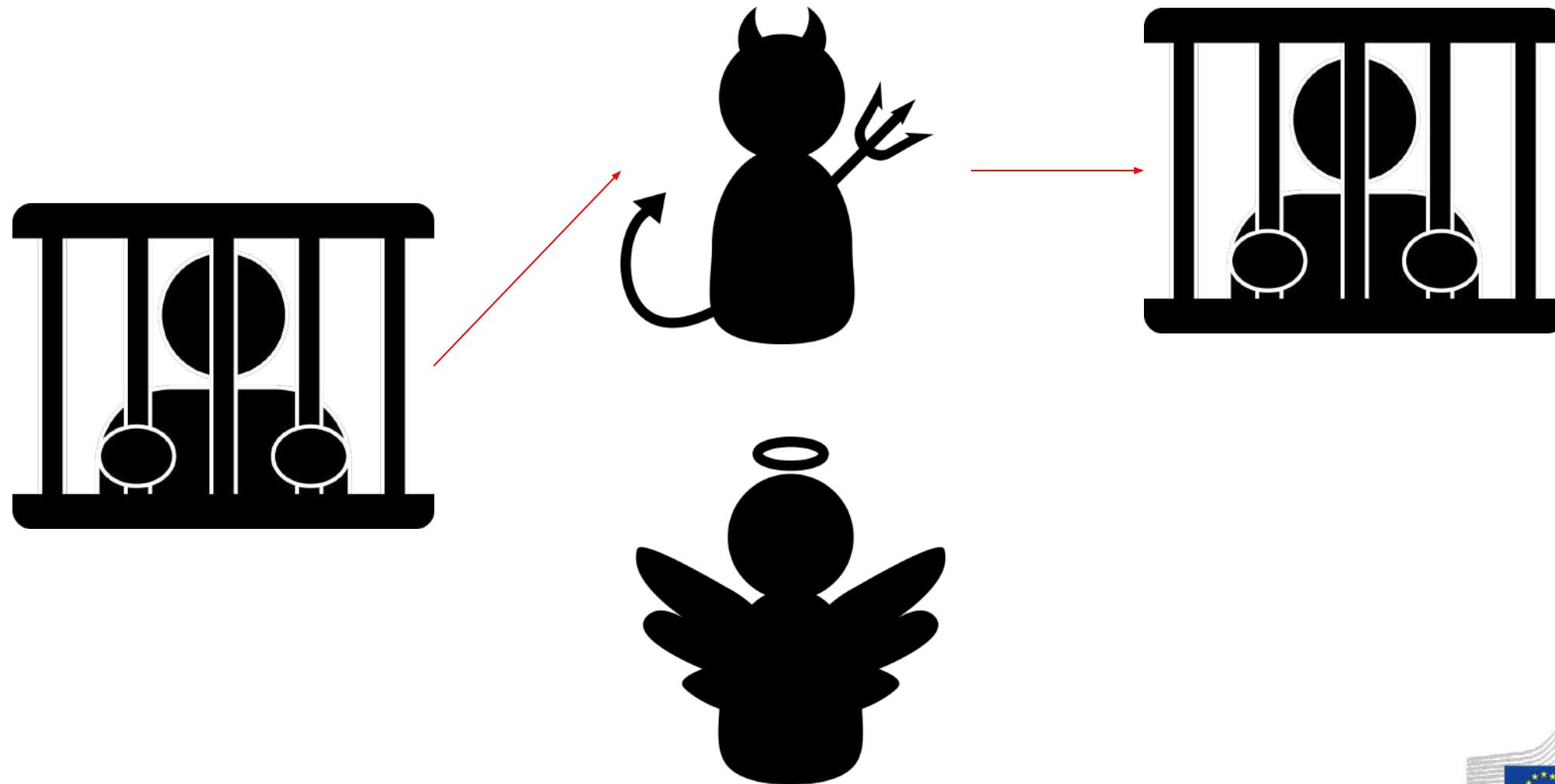
Termeni

Predictive performance = performanță

Fairness = egalitate (socială, algoritmică) / dreptate / imparțialitate

Bias = bias / prejudecată

Recidiva penală



Predictia recidivei penale



Prisoner



**Human
expert**



**Decision
/ Sentence**

Predictia recidivei penale



Prisoner



**Human
expert**



**Decision
/ Sentence**



Outcome



Predictia recidivei penale



Prisoner



**Human
expert**



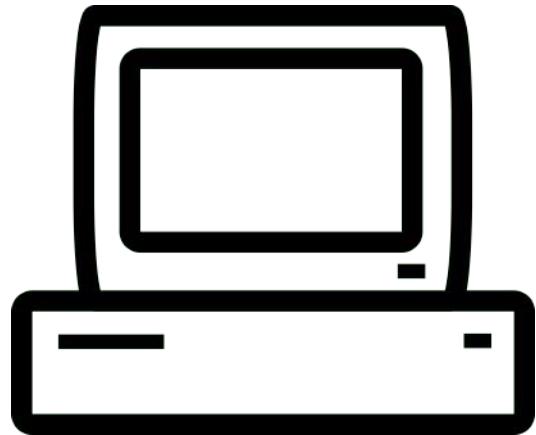
**Decision
/ Sentence**



Outcome



Predictia recidivei penale



Features

**Machine
learning model**



Prediction



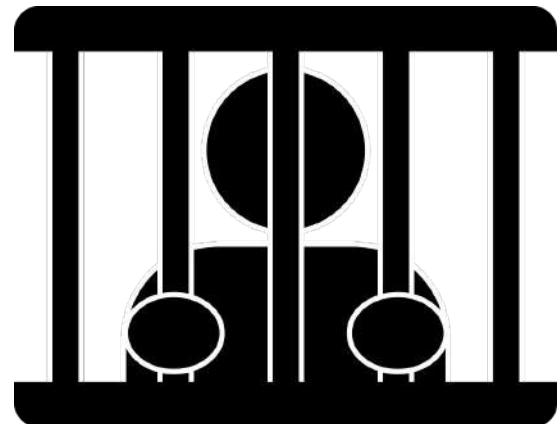
Outcome



European
Commission

Predictia recidivei penale

**Caracteristici
statice/demografice:**



- **Age at crime**
- **Sex**
- **Nationality**
- **Previous number of crimes**
- **Sentence**
- **Year of crime**
- **Probation**

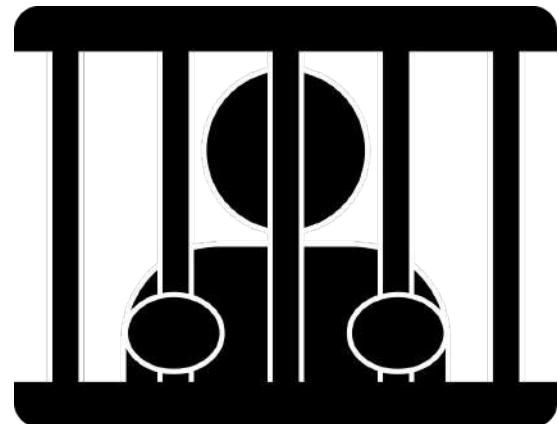
Egalitate / Dreptate / Imparțialitate



O decizie e imparțială dacă nu discriminează cetățenii pe criterii legate de rasă, etnie, gen, sex etc (grupuri protejate)

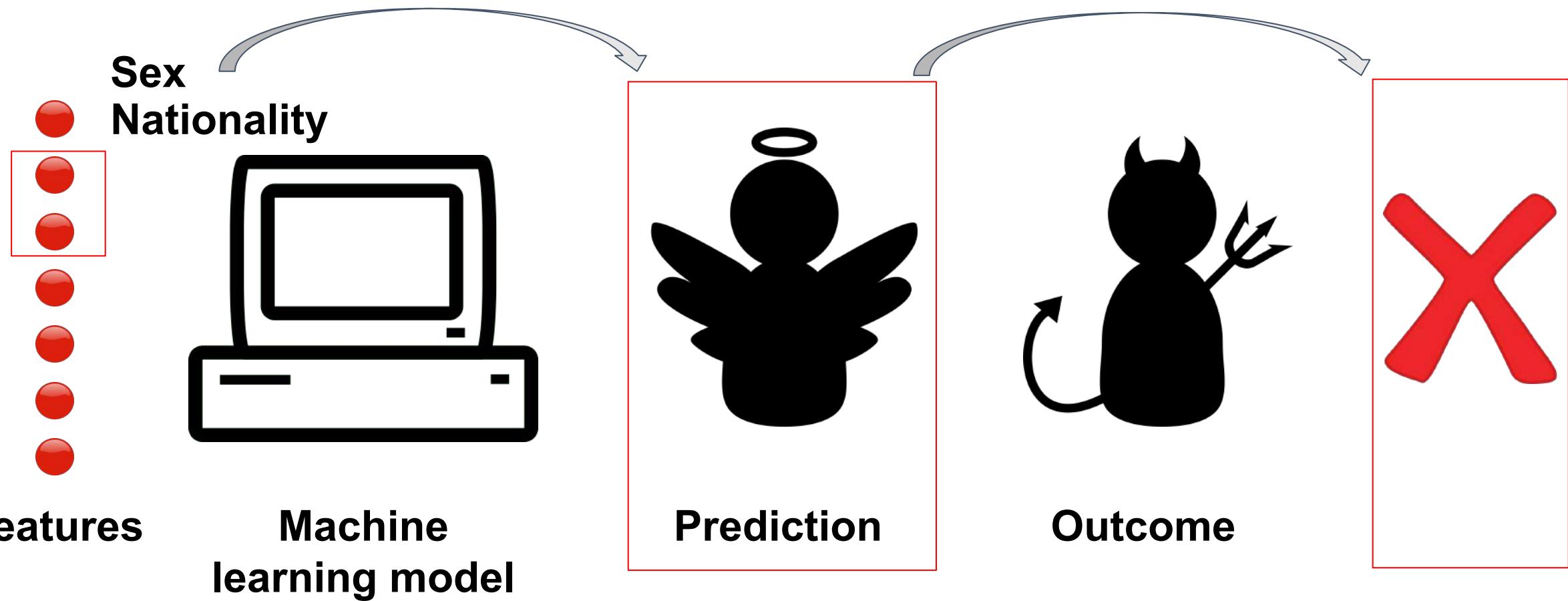
Egalitate / Dreptate / Imparțialitate

**Caracteristici
statice/demografice:**

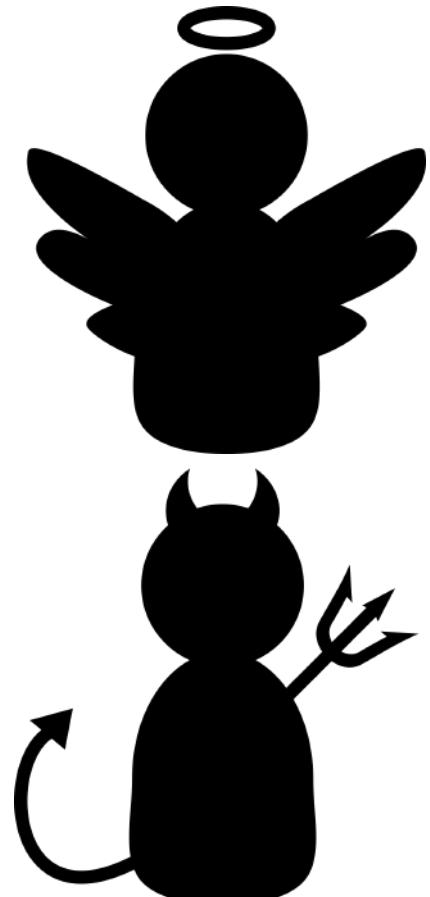


- **Age at crime**
- **Sex**
- **Nationality**
- **Previous number of crimes**
- **Sentence**
- **Year of crime**
- **Probation**

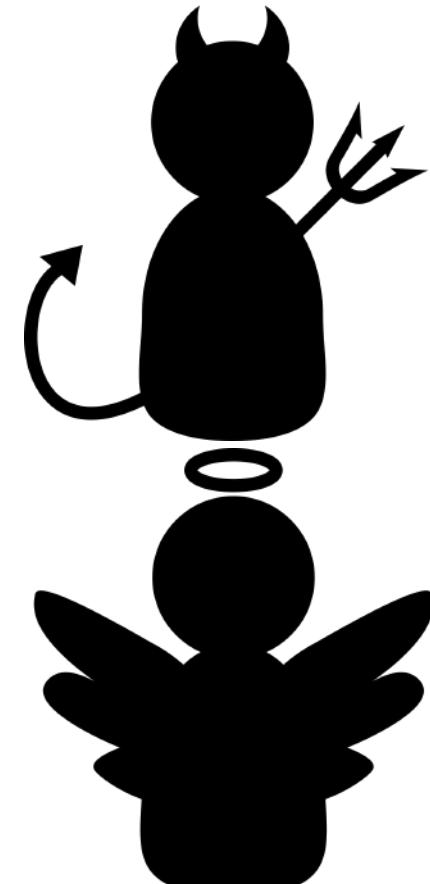
Evaluarea discriminării



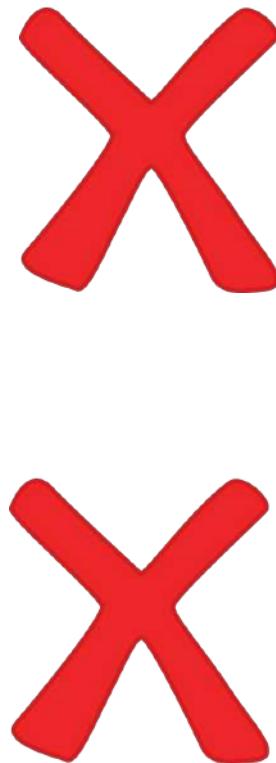
Evaluarea discriminării



Prediction



Outcome

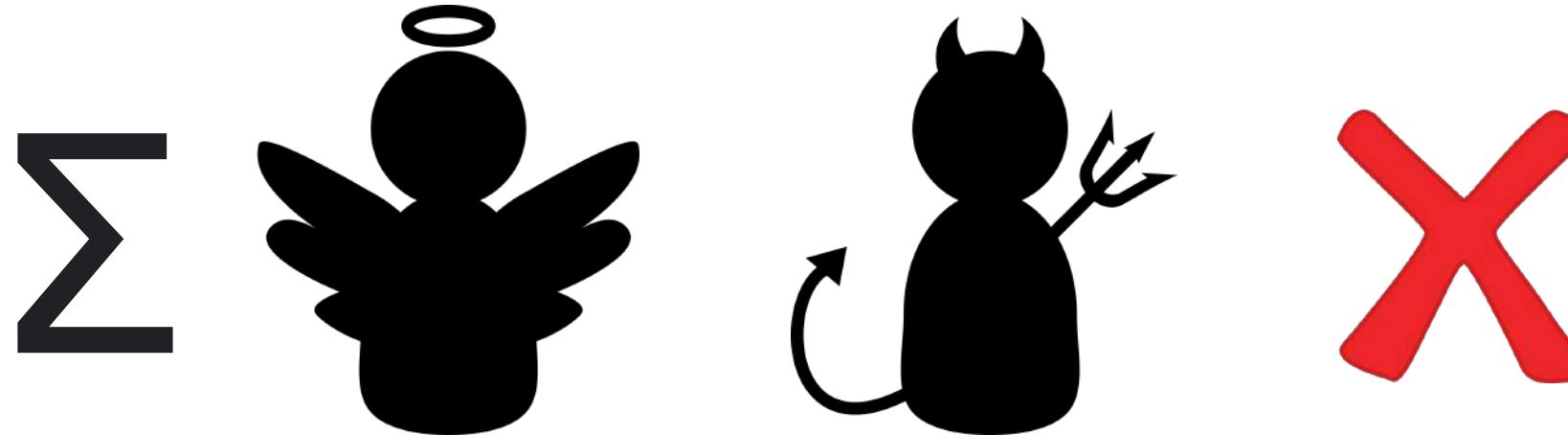


False negative

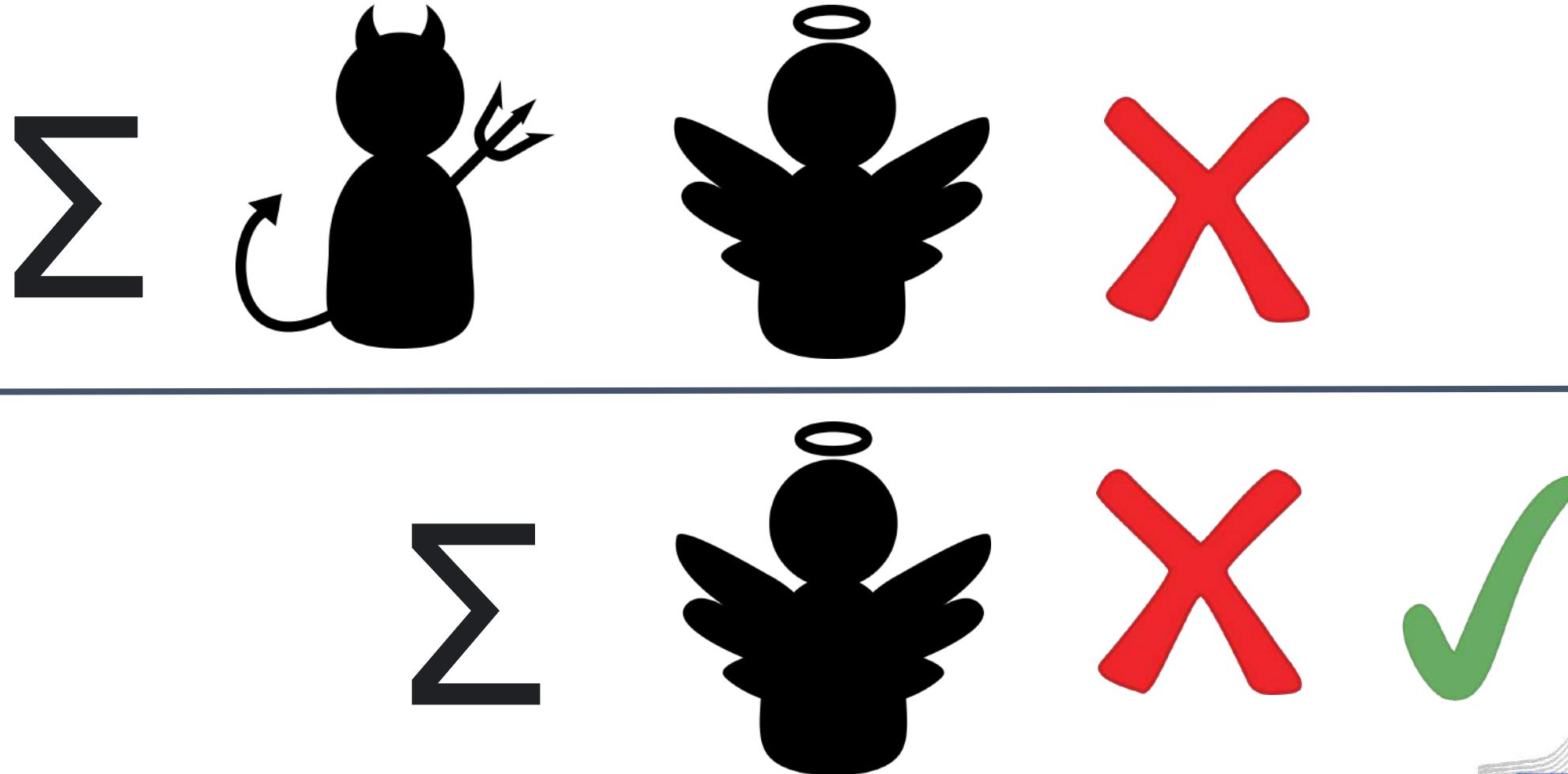


False positive

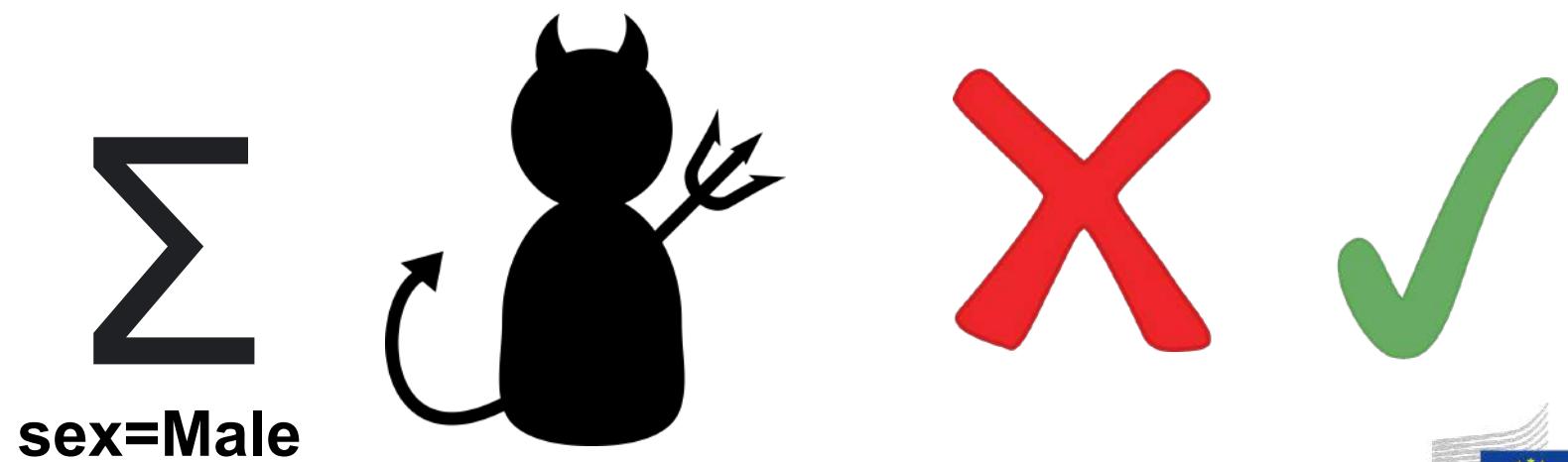
FNR Rata fals negative - “cazuri nedetectate”



FPR Rata fals pozitive - “alarme false”



Egalitatea/discriminarea pe criterii de grup - sex



Diferența între FNR-urile grupurilor



$$FNR_{\text{disparity}} = \frac{FNR_{\text{female}}}{FNR_{\text{male}}}$$

= cineva e etichetat în mod fals ca ne-recidivist.

Diferența între FPR-urile grupurilor



$$FPR_{\text{disparity}} = \frac{FPR_{\text{female}}}{FPR_{\text{male}}}$$

= cineva e etichetat în mod fals / pe nedrept ca recidivist.

Headache?



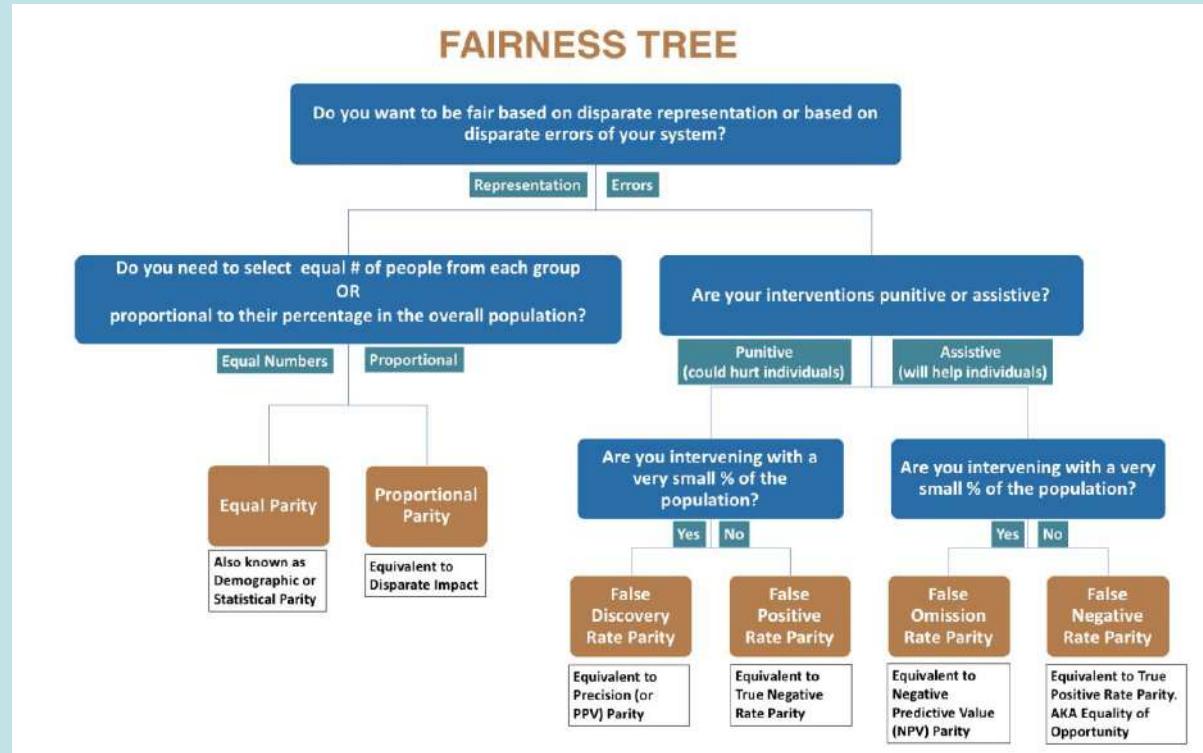
European
Commission

Prea multe definiții și moduri de a evalua



The fairness in machine learning literature comprises at least 21 disparity metrics.

Cum alegem măsurătoarea potrivită?



Recidiva juvenilă



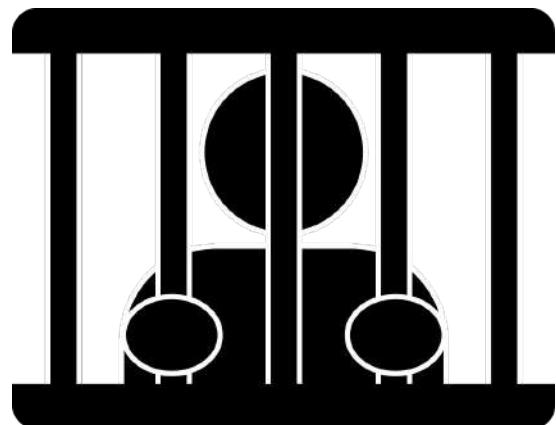
Sisteme algoritmice pentru detectarea riscului

Structured Assessment of Violence Risk in Youth (SAVRY)

- Experții sunt implicați - psihologi
- Sumă simplă de factori (in comparison with COMPAS)
- 24 factori de risc - low, medium, high

SAVRY

Example factori SAVRY:



- **Early violence**
- **Self-harm history**
- **Home violence**
- **Poor school achievement**
- **Stress and poor coping**
- **Substance abuse**
- **Criminal parent/caregiver**

Predictia recidivei penale

SAVRY
features

Σ
SAVRY sum



Expert



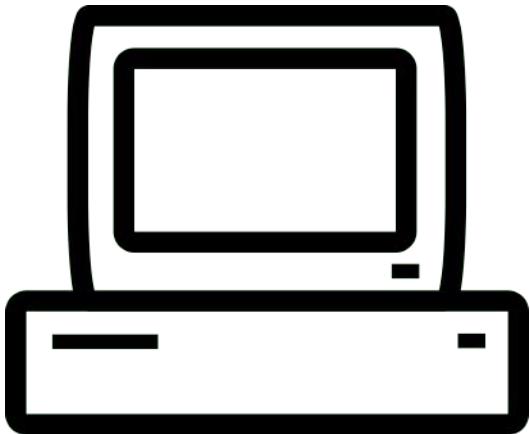
Final expert
evaluation



Outcome



Static ML



Features

Machine
learning model



Prediction

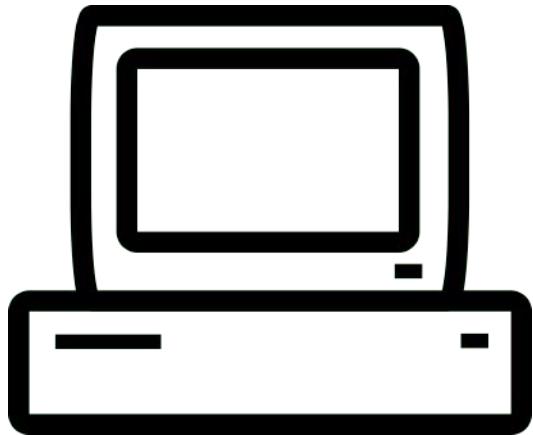


Outcome



European
Commission

SAVRY ML



**SAVRY
features**

**Machine
learning model**



Prediction

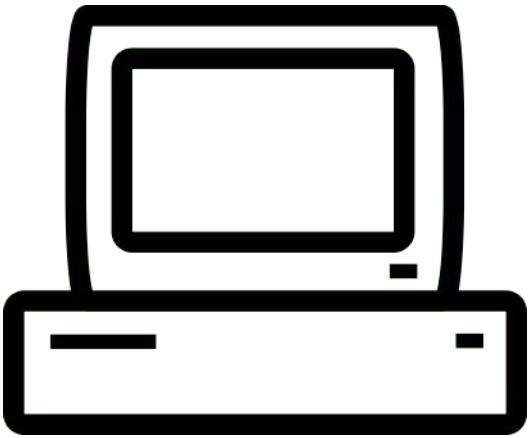
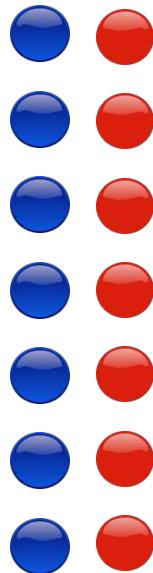


Outcome



European
Commission

Static + SAVRY ML



Features

**Machine
learning model**



Prediction



Outcome



European
Commission

Dataset

Juvenile offenders in Catalonia¹

- 855 people
- crimes between 2002 -2010, release in 2010
- age at crime between 12 and 17 years old
- status followed up on 2013 and 2015

1. Open data: <http://cejfe.gencat.cat/en/recerca/opendata/jjuvenil/reincidencia-justicia-menors/index.html>

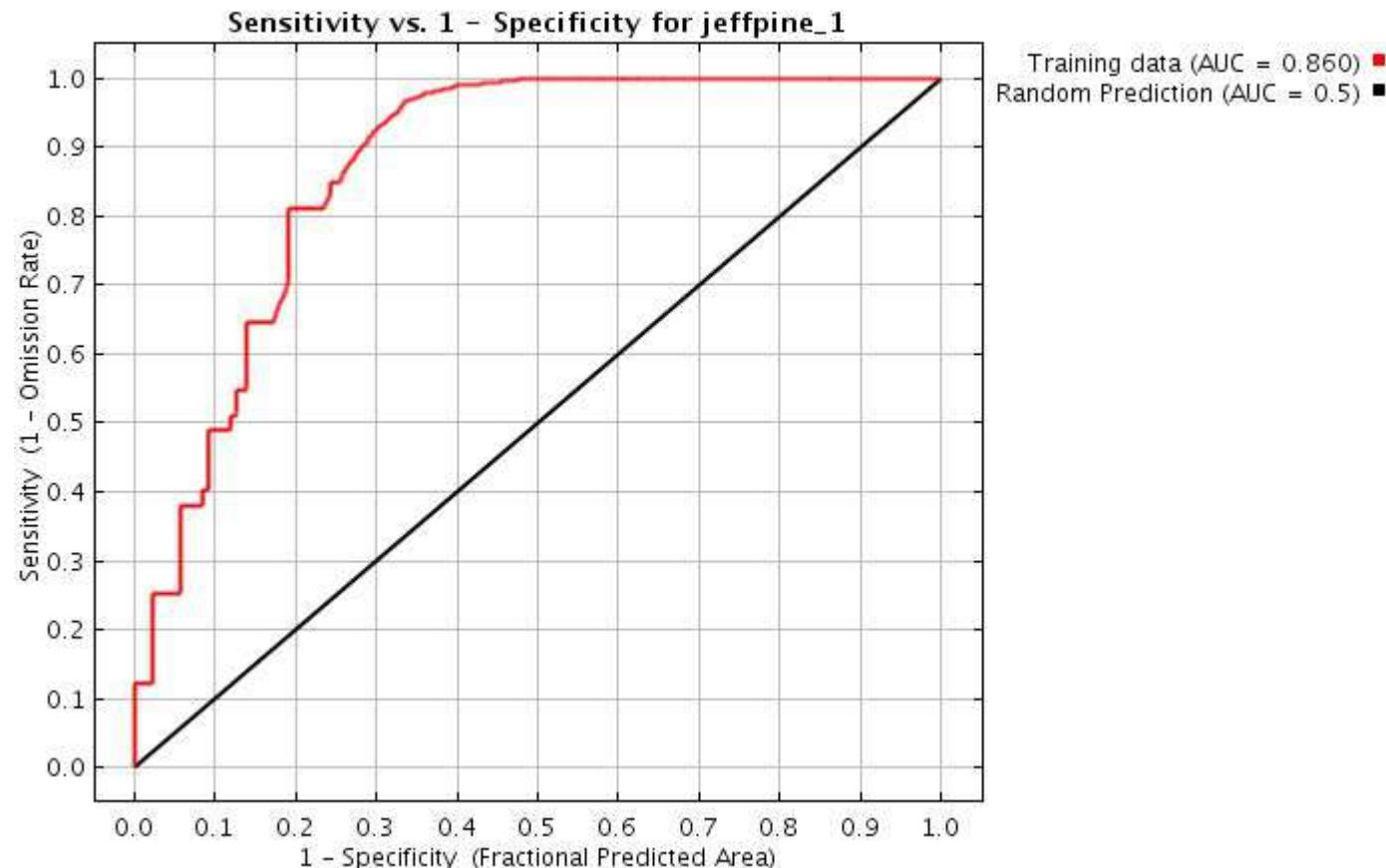
Experimental setup

Training a set of ML methods

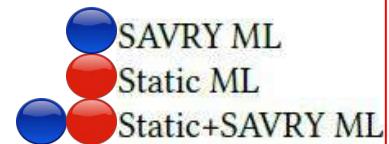
- logistic regression (logit), multi-layer perceptron (mlp), support vector machines (lsvm), k-nearest neighbors (knn), random forest (rf), naive bayes (nb)
- k-fold cross validation with $k=10$ (10% test, 10% validation, 80% training)
- we run 50 different experiments with different initial conditions
- **we compute feature importance with LIME¹**

1. LIME <https://github.com/marcotcr/lime>

Predictive performance - AUC ROC



Results, predictive performance AUC

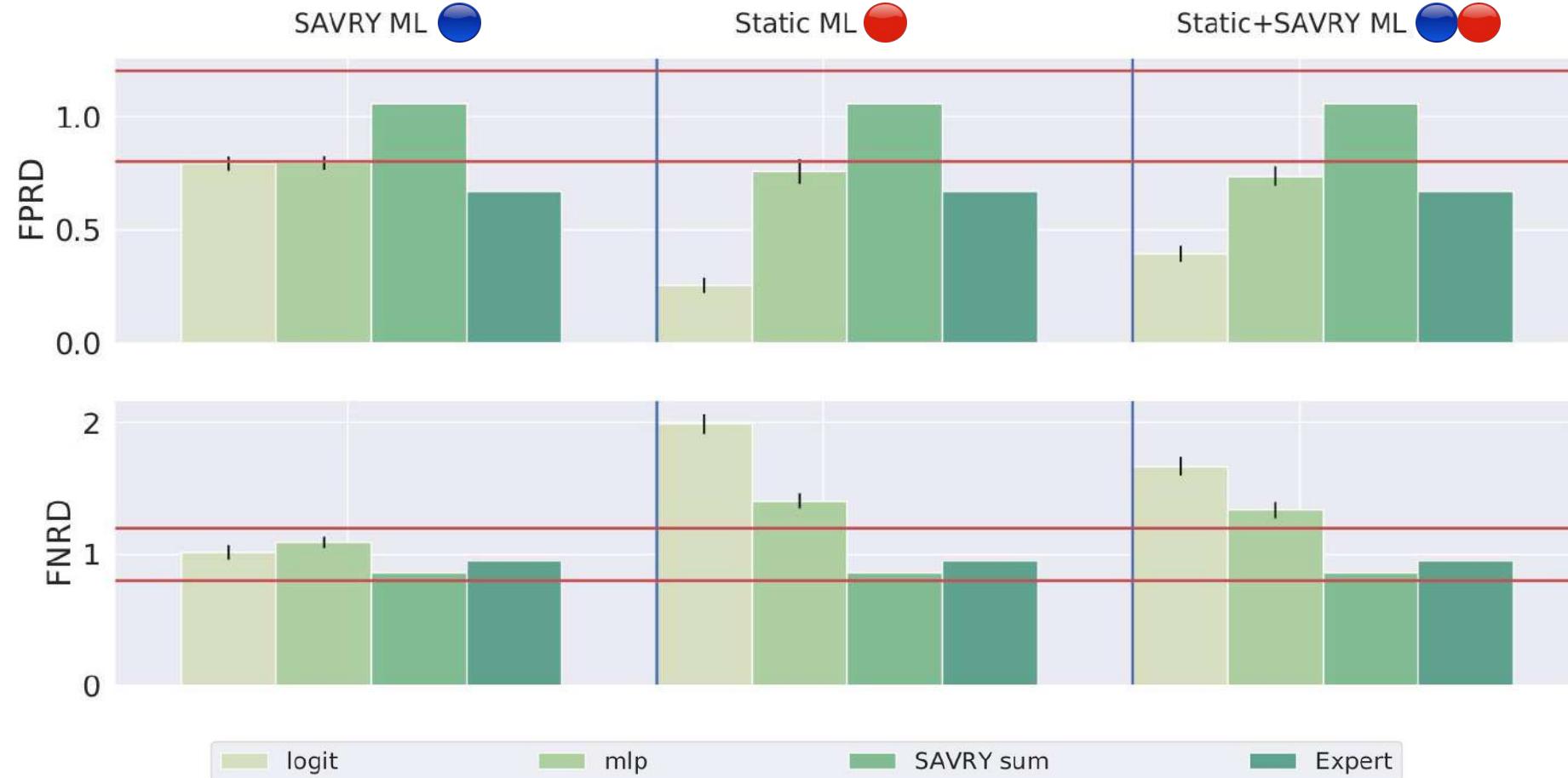


	logit				mlp				knn		lsvm		rsvm		nb		rf	
	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.
SAVRY ML	.66	.0058	.66	.0058	.60	.0121	.65	.0082	.52	.0197	.65	.0015	.65	.0110				
Static ML	.70	.0055	.70	.0068	.62	.0122	.61	.0119	.56	.0149	.69	.0040	.66	.0110				
Static+SAVRY ML	.71	.0064	.70	.0053	.64	.0129	.71	.0074	.50	.0058	.69	.0018	.69	.0121				

SAVRY Sum - 0.64 AUC
Expert - 0.66 AUC

Results: disparity, sex

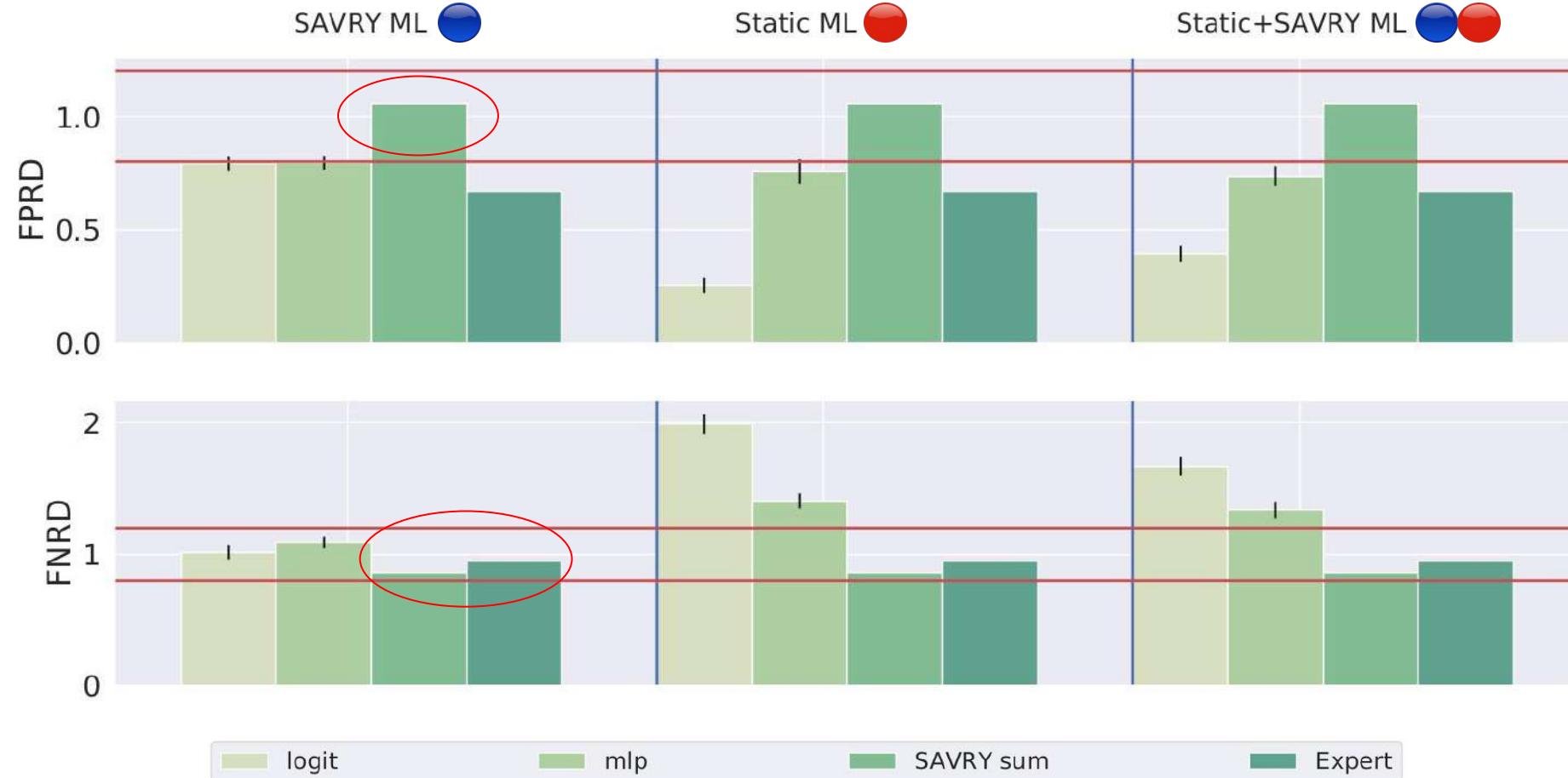
**Alarme
false**



**Cazuri
nedetec-
tate**

Results: disparity, sex

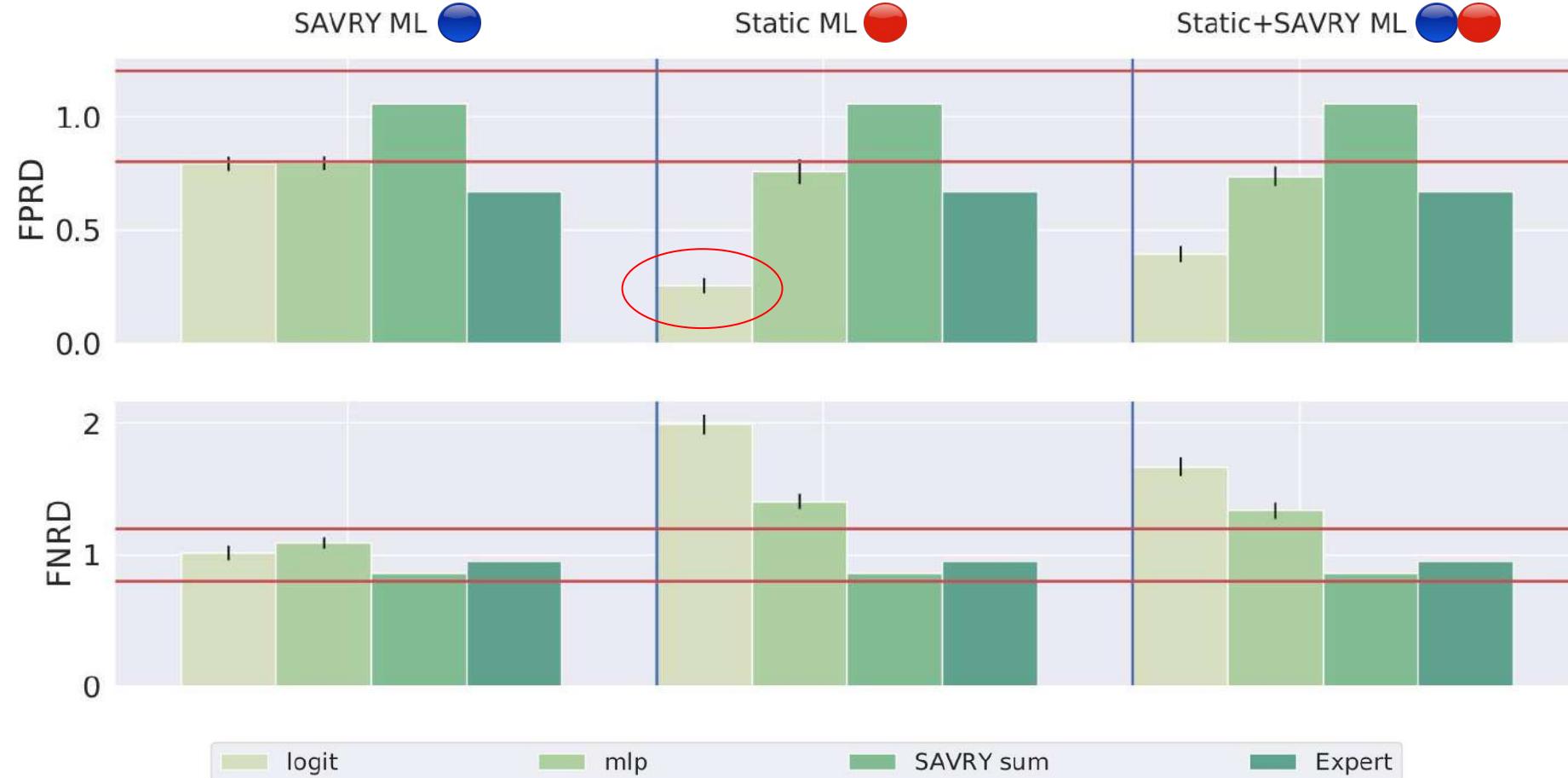
**Alarme
false**



**Cazuri
nedetec-
tate**

Results: disparity, sex

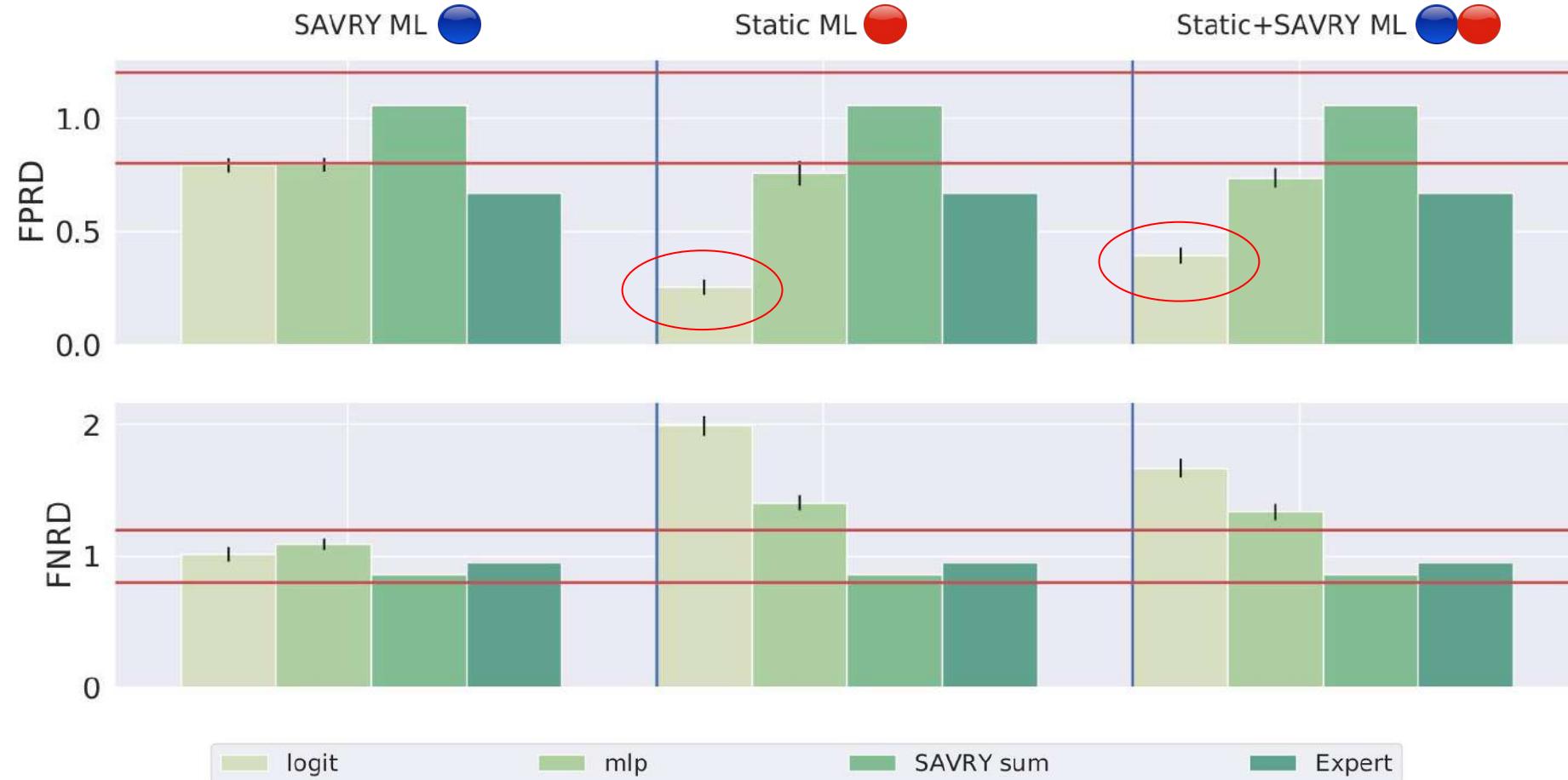
**Alarme
false**



**Cazuri
nedetec-
tate**

Results: disparity, sex

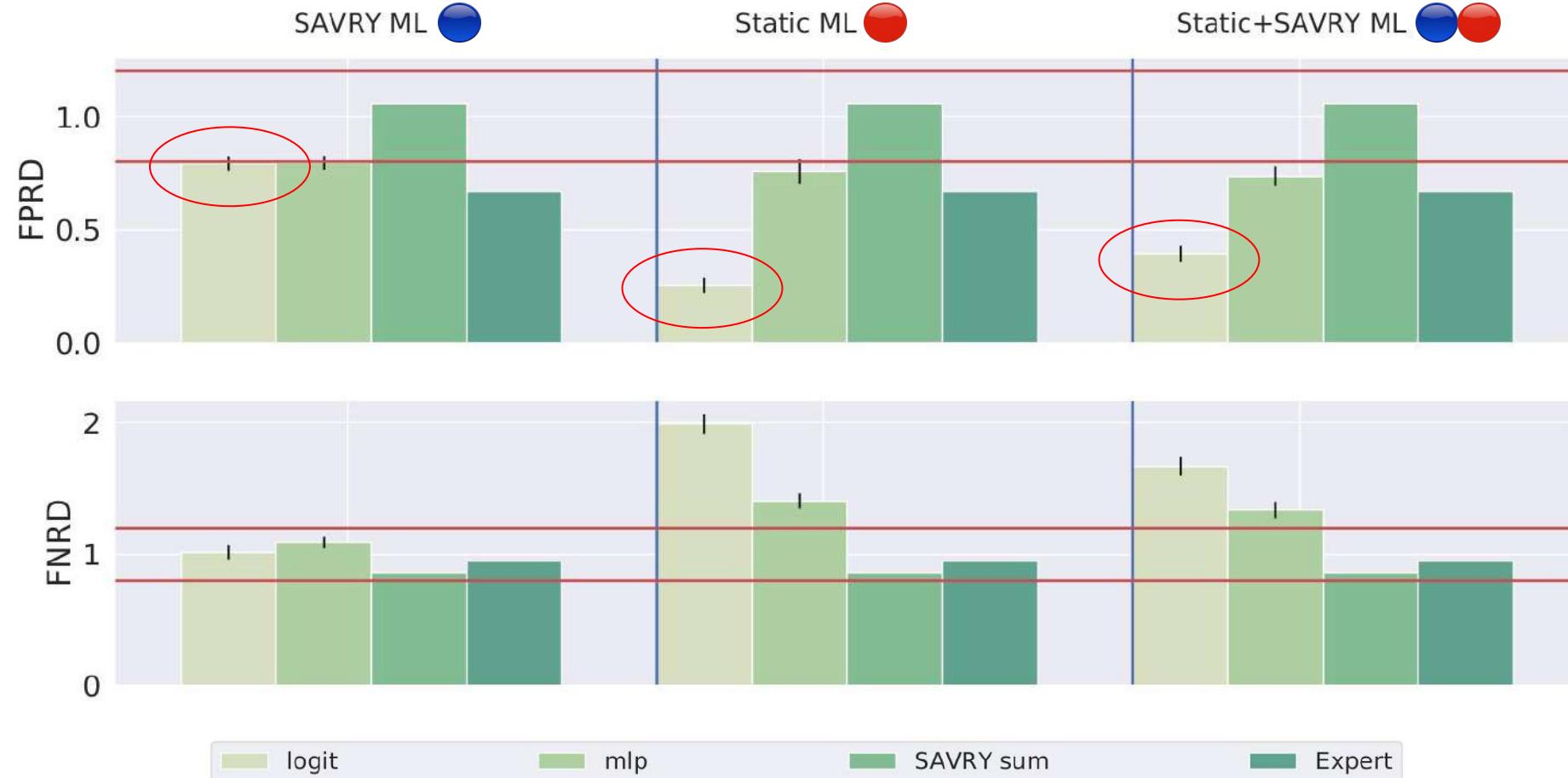
**Alarme
false**



**Cazuri
nedetec-
tate**

Results: disparity, sex [higher for men]

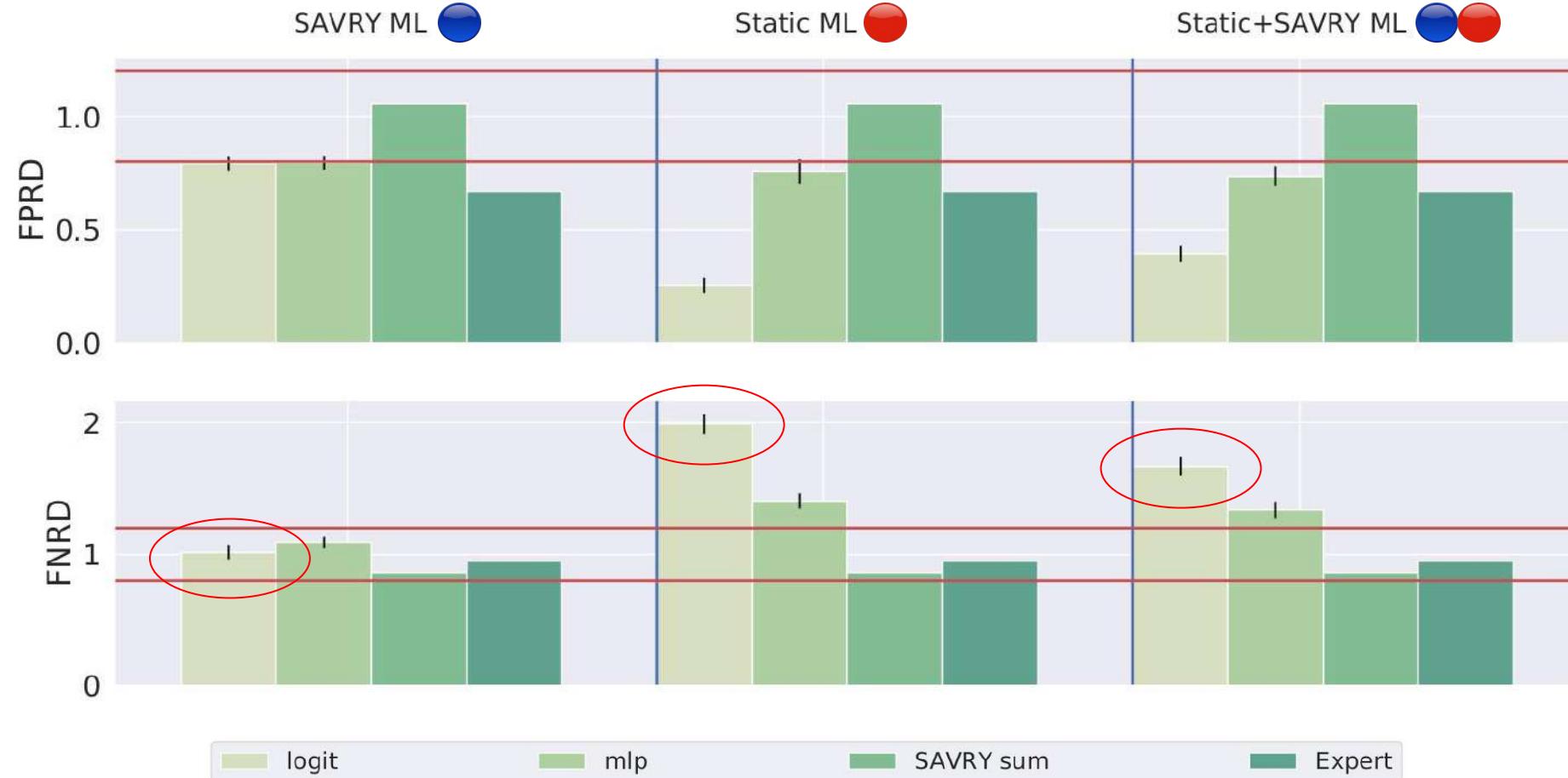
**Alarme
false**



**Cazuri
nedetec-
tate**

Results: disparity, sex

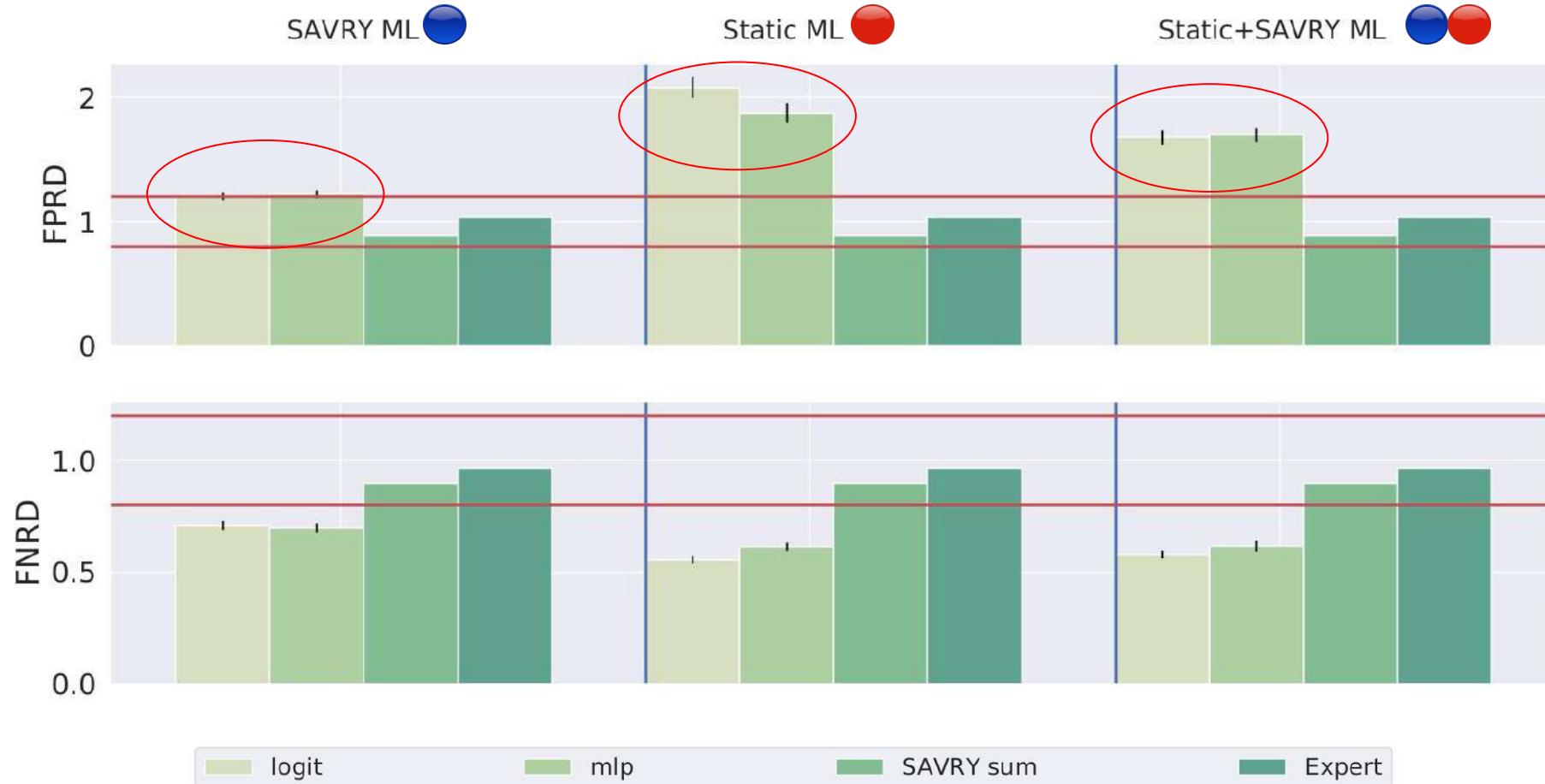
**Alarme
false**



**Cazuri
nedetec-
tate**

Results: disparity, nationality [higher for foreigner]

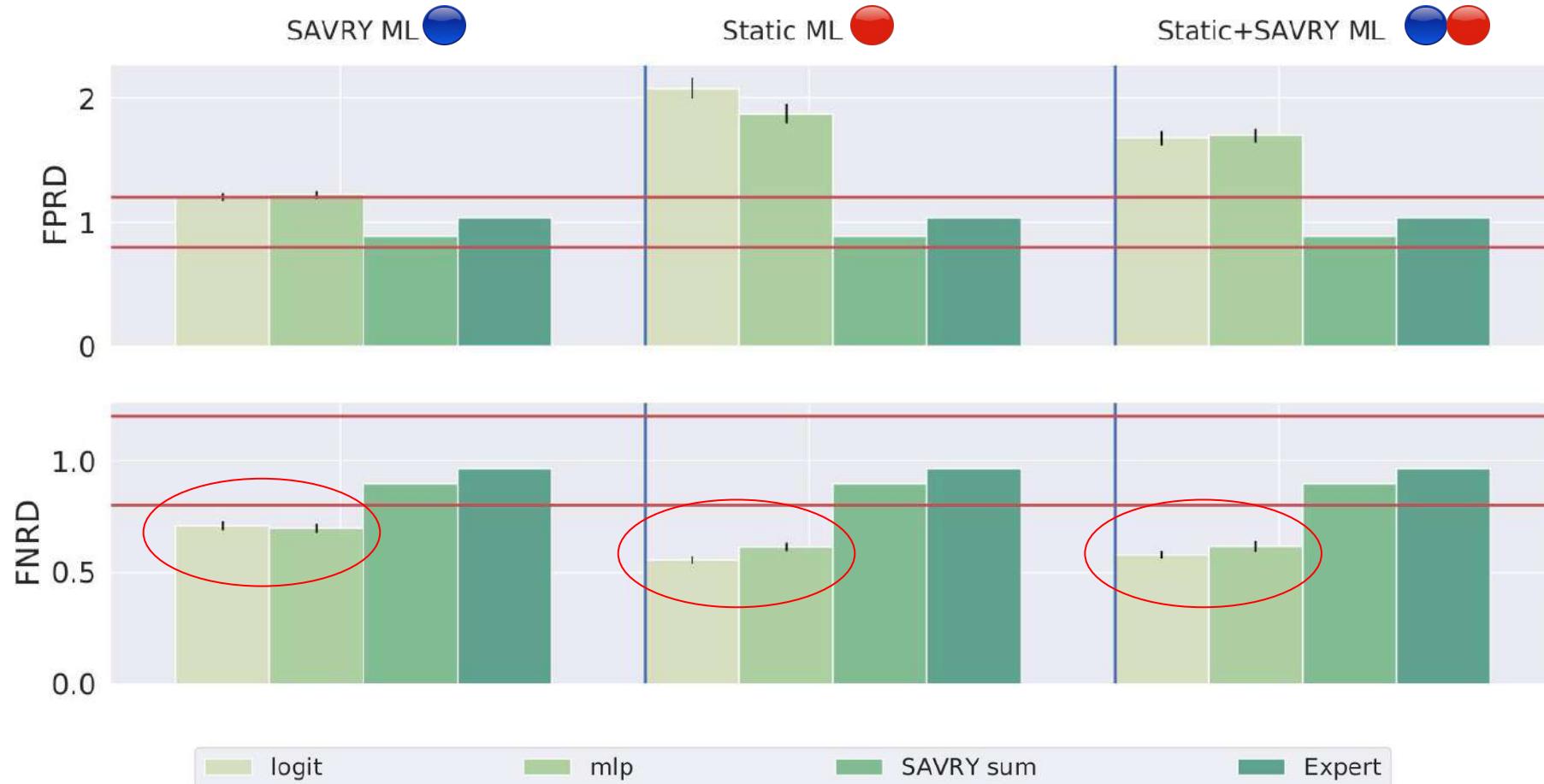
**Alarme
false**



**Cazuri
nedetec-
tate**

Results: disparity, nationality

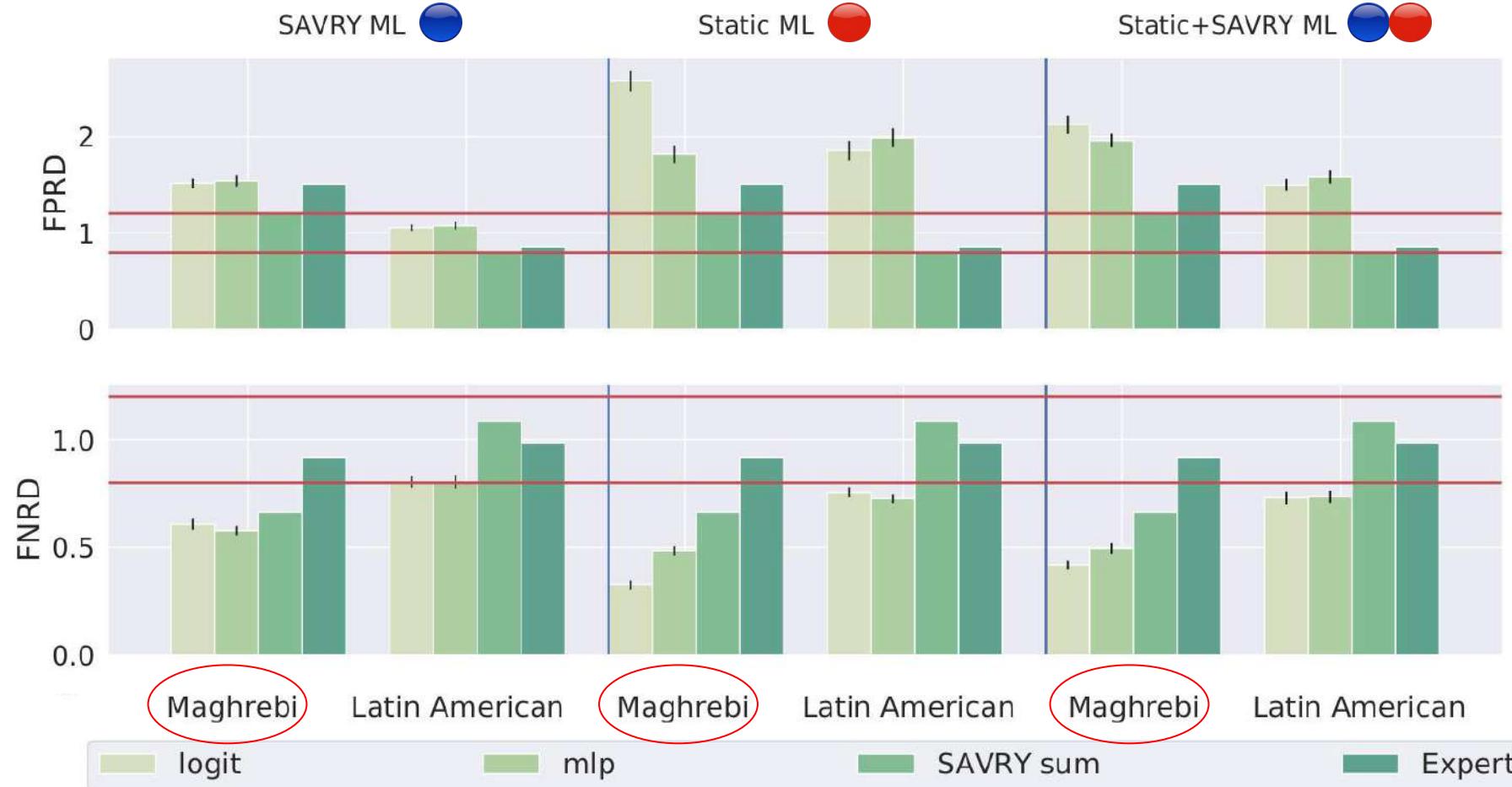
**Alarme
false**



**Cazuri
nedetec-
tate**

Results: disparity, nationality

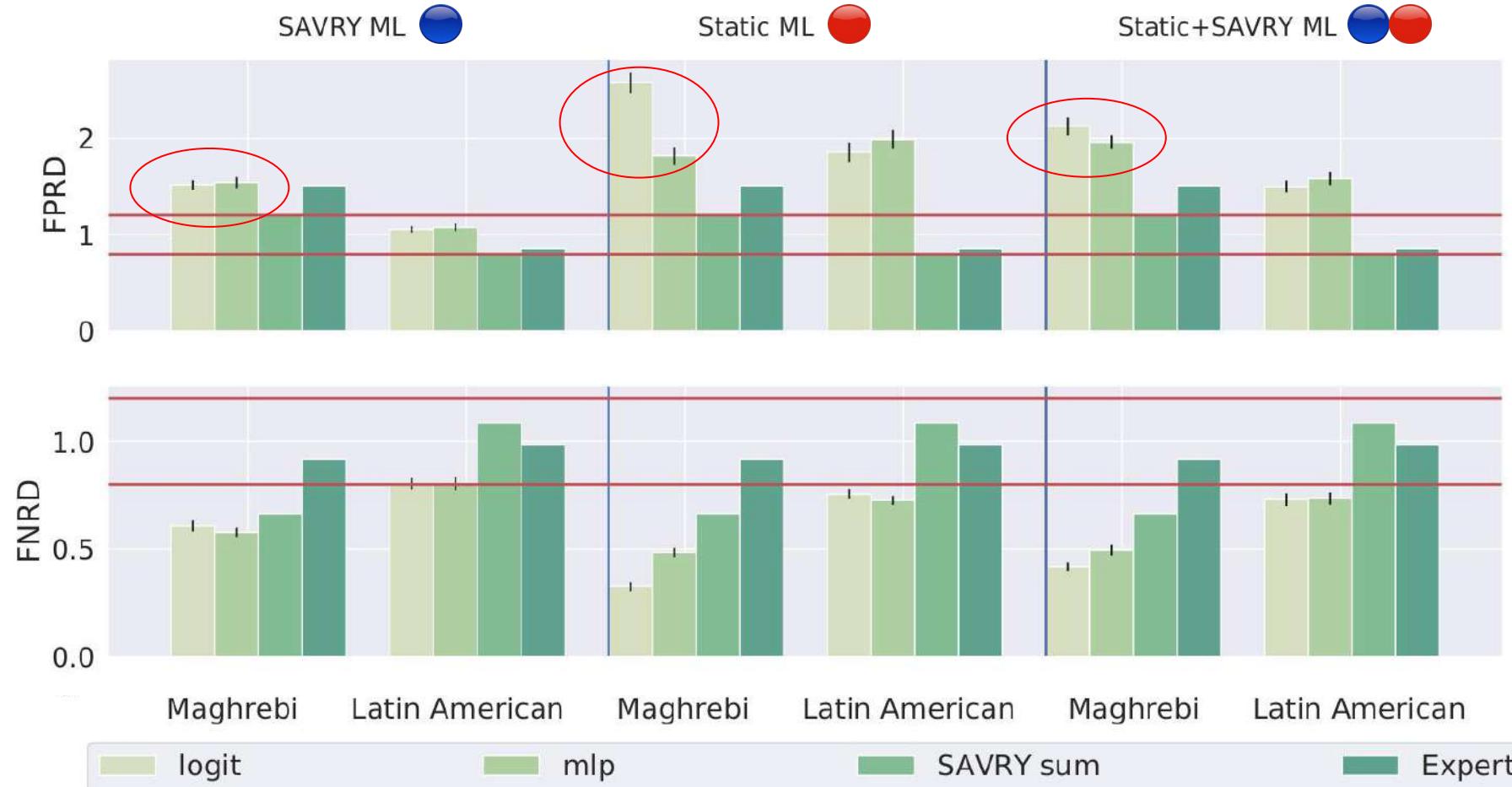
**Alarme
false**



**Cazuri
nedetec-
tate**

Results: disparity, nationality

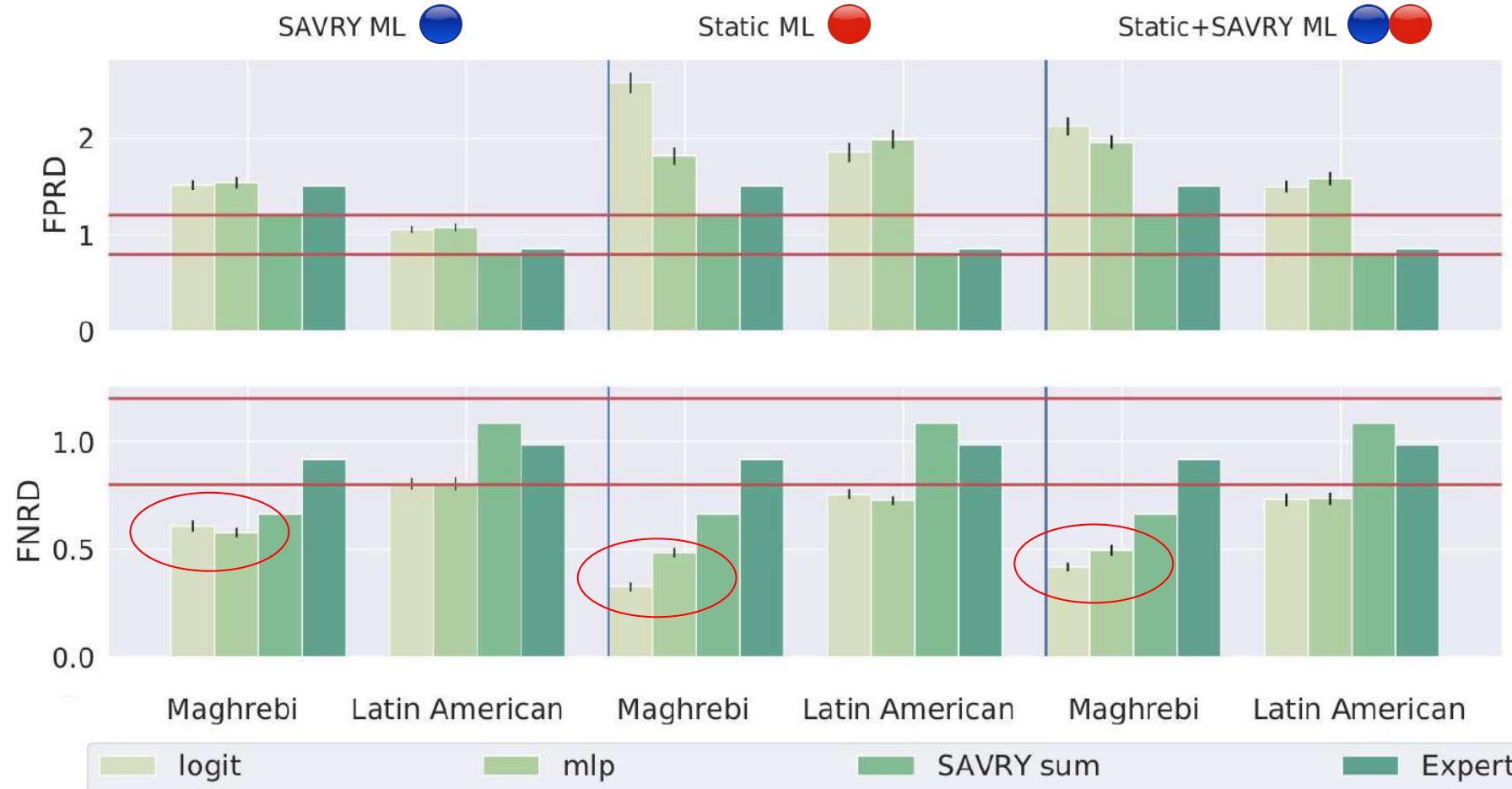
**Alarme
false**



**Cazuri
nedetec-
tate**

Results: disparity, nationality

**Alarme
false**



**Cazuri
nedetec-
tate**

Results: feature importance for logit

	SAVRY ML	Static	Static+SAVRY		
final expert evaluation	0.370*** (0.076)	✓ crime in 07-08 ✓ crime in year 09 ✓ age maincrime ✓ days to program start (norm) ✓ crime in year 10 ✓ days in program (norm) ✓ prog: enforcement measure ✓ prior crimes frequency ✓ female ✓ Maghrebi ✓ Latin American ✓ prog: mediation/reparation	-0.298** (0.118) -0.259** (0.121) -0.109*** (0.021) -0.105*** (0.040) -0.275*** (0.098) -0.087* (0.048) -0.248** (0.103) 0.059* (0.033) -0.187*** (0.046) 0.158*** (0.058) 0.105** (0.052) -0.178* (0.103)	✓ crime in years 07-08 ✓ crime in year 09 ✓ days to program start (norm) ✓ age maincrime final expert evaluation ✓ crime in year 10 ✓ female ✓ enforcement measure ✓ Maghrebi ✓ Latin American	-0.272** (0.133) -0.255* (0.132) -0.117*** (0.044) -0.115*** (0.022) 0.291*** (0.091) -0.256** (0.115) -0.196*** (0.053) -0.206* (0.122) 0.152** (0.069) 0.135** (0.060)
SAVRY sum	0.183 (0.910)				
personality	-1.362 (7.061)				
treatment susceptibility	-1.340 (6.336)				
total score (social)	-0.141 (0.909)				
total score (protective)	0.191 (0.902)				
previous violent offenses	-0.601 (2.533)				
total score (historic)	0.056 (0.045)				
home violence	-0.543 (1.816)				
past intervention failures	-0.598 (2.530)				

Results: feature importance for mlp

	SAVRY ML			Static ML			Static+SAVRY ML	
feature	importance		feature	importance		feature	importance	
	Mean	StdDev		Mean	StdDev		Mean	StdDev
probation/internment	147.43	24.85	✓ province of residence	219.21	28.44	✓ foreigner	199.80	11.37
total score (social)	117.93	9.71	✓ age maincrime	202.83	25.72	✓ sex	188.07	8.35
total score (personality)	117.63	9.83	✓ foreigner	178.38	19.06	✓ national group	117.40	23.09
total score (protective)	115.76	8.56	✓ year of maincrime	168.96	13.86	✓ maincrime category	150.90	16.44
total score (historic)	116.59	10.25	✓ prior crimes	175.11	22.56	✓ prior crimes frequency	151.53	18.26
history non-violent offending	112.17	7.44	✓ national group	181.68	32.23	✓ maincrime program sentence	143.29	10.50
positive/resilience characteristics	111.62	7.32	✓ prior crimes frequency	156.15	20.98	✓ year of maincrime	141.88	9
previous violence	113.22	8.93	✓ maincrime category	144.27	18.26	✓ maincrime violent	148.92	16.23
early violence	111.42	7.17	✓ maincrime violent	137.20	14.95	✓ province of execution	146.07	13.76
pro-social activities	109.82	5.57	✓ prior crimes	131.53	12.66	✓ prior crimes	146.97	14.71

Results: difference in base rates, prevalence

	Base rate	Not Recidivated	Recidivated	Difference
protected features				
male	40.03%	0.839	0.931	0.093***
female	20.37%	0.161	0.069	-0.093***
Spanish	32.06%	0.667	0.523	-0.143***
foreign	46.22%	0.333	0.477	0.143**

Descriptive statistics of input features by recidivism status.

Results: difference in base rates, prevalence

	Base rate	Not Recidivated	Recidivated	Difference
protected features				
male	40.03%	0.839	0.931	0.093***
female	20.37%	0.161	0.069	-0.093***
Spanish	32.06%	0.667	0.523	-0.143***
foreign	46.22%	0.333	0.477	0.143**

Descriptive statistics of input features by recidivism status.

Results: difference in base rates, prevalentă

	Base rate	Not Recidivated	Recidivated	Difference
protected features				
male	40.03%	0.839	0.931	0.093***
female	20.37%	0.161	0.069	-0.093***
Spanish	32.06%	0.667	0.523	-0.143***
foreign	46.22%	0.333	0.477	0.143**

Descriptive statistics of input features by recidivism status.

Conclusions

- ML models have better predictive performance
- ML models tend to discriminate more
- static features outweigh SAVRY features as importance
- preliminary study: the cause may be in the data (base rates)

Contributions

We propose a methodology and a ML framework¹

- to easily train ML models on tabular data (csv files)
- to evaluate these models in terms of predictive performance and fairness
- to connect to interpretability frameworks
- to reproduce with ease results and research

1. Open framework: <https://gitlab.com/HUMAINT/humaint-fatml>



Multumesc!

Întrebări?

@nkundiushuti & mariusmiron.com