clearbox AI

DataBeers Torino
Talent Garden, Feb 27 2024

Know you Bias: Affrontare i Bias nei Dati attraverso i Dati Sintetici

Simona Mazzarino

humans && machines

Cosa sono i bias?

- Nel campo della data science e della statistica, il bias è una tendenza sistematica in cui i metodi utilizzati per raccogliere dati e generare statistiche presentano una rappresentazione inaccurata, distorta o tendenziosa della realtà.

- Un bias può presentarsi in numerose fasi del processo di raccolta e analisi dei dati.

Che tipi di bias esistono?

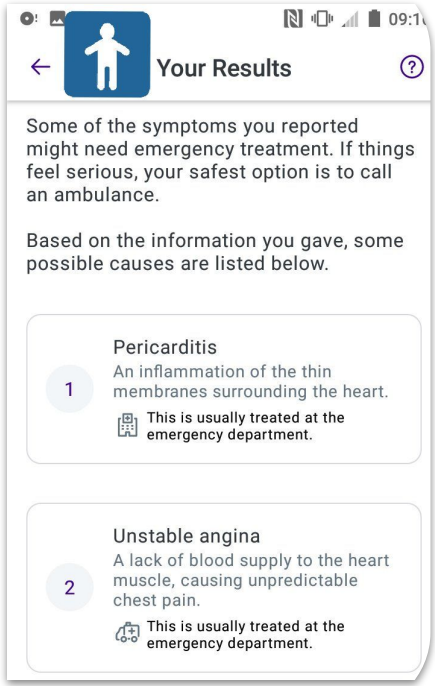
- **Bias dovuti alla selezione dei dati**
  Il bias si verifica a seguito di una determinata scelta dei dati utilizzati per addestrare un modello di machine learning;

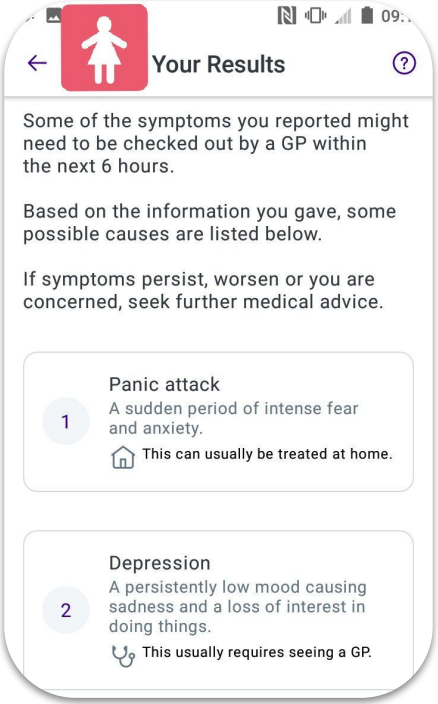
- **Bias sociali**
  Derivano da pregiudizi o stereotipi presenti nella società;

- **Bias statistico/computazionale**
  Originato dall'uso e dall'interpretazione dei modelli di intelligenza artificiale.

Fonte: Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence. NIST special publication, 1270(10.6028).

Bias e fairness





Bias e fairness

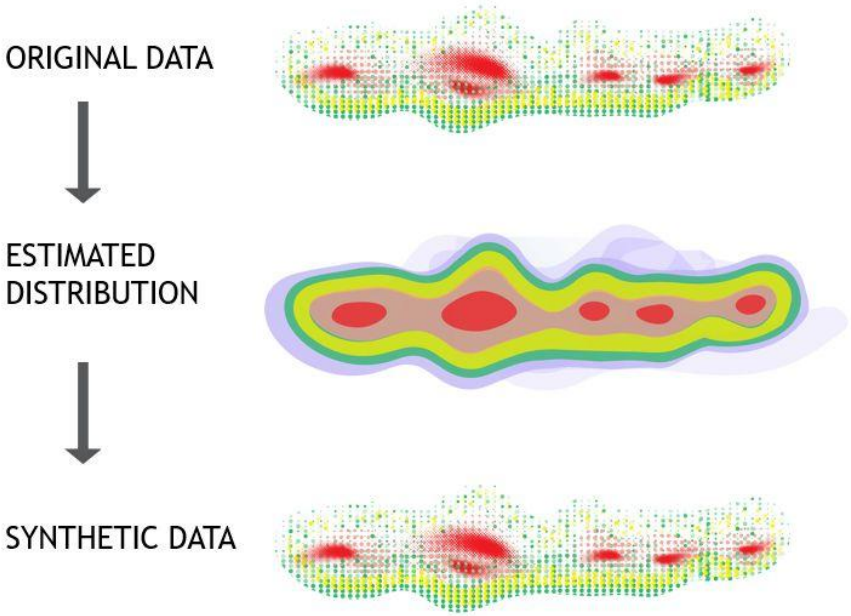
Fonte: Apple's 'sexist' credit card investigated by US regulator, BBC News. https://www.bbc.com/news/business-50365609

Bias e fairness nell'AI generativa

"An impressionist painting of a data scientist working on their laptop"

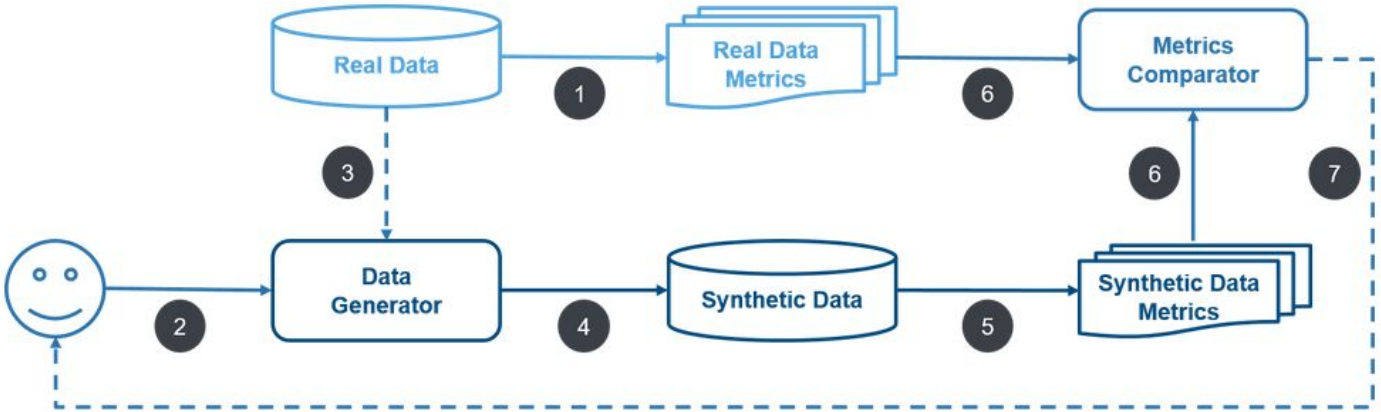
"An impressionist painting of a person sweeping the floor"

Cosa sono i dati sintetici?



- I dati sintetici vengono generati artificialmente, utilizzando algoritmi di intelligenza artificiale su campioni di dati reali.
- Essi possiedono le stesse proprietà statistiche e capacità predittive dei dati reali su cui sono stati generati.

Come vengono generati i dati sintetici?



Fonte: J. P. Morgan https://www.jpmorgan.com/technology/artificial-intelligence/initiatives/synthetic-data

Come affrontare i bias con i dati sintetici?

Dataset giocattolo

Adult Census Income, classificazione binaria con XGBoost model.

Obiettivo

Far sì che il modello prenda decisioni sullo stipendio senza tener conto di alcune categorie protette come le colonne sex, race o relationship.

	age	work_class	education	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_country	income
0	39	State-gov	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	HS-grad	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	11th	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...
32556	27	Private	Assoc-acdm	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
32557	40	Private	HS-grad	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
32558	58	Private	HS-grad	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
32559	22	Private	HS-grad	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
32560	52	Self-emp-inc	HS-grad	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

Come affrontare i bias con i dati sintetici?

Per capire se un modello non prende decisioni basandosi su categorie protette come il sesso, l'etnia o lo stato matrimoniale, esistono due importanti metriche:

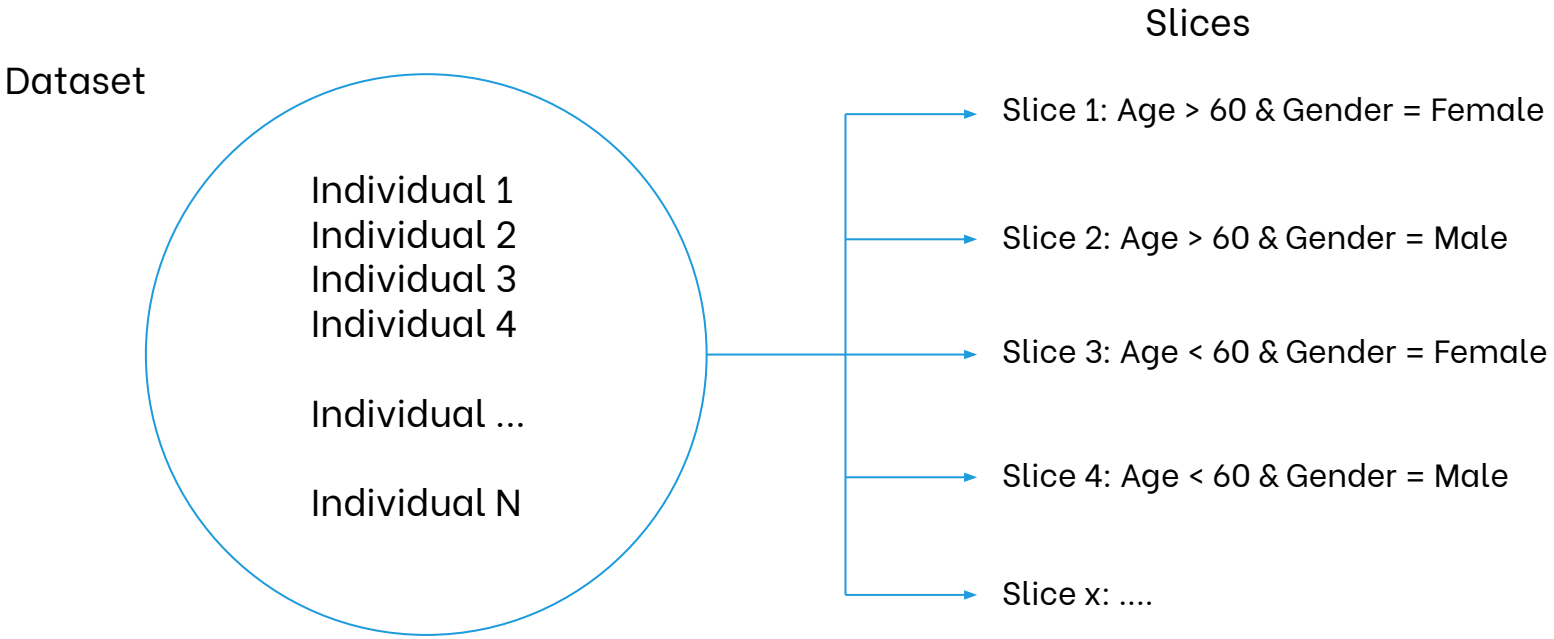
Equalised odds

Le predizioni del modello sono indipendenti delle variabili sensibili.
TPR (true positive rate, tasso di veri positivi) e **FPR** (false positive rate, tasso di falsi positivi) sono uguali tra i gruppi protetti.

Equalised opportunity

Le aspettative rispetto all'etichetta positiva non cambiano tra i gruppi.

Come affrontare i bias con i dati sintetici?



Come affrontare i bias con i dati sintetici?

```
Slice  1 : `age`>=17.0 and `age`<=41.333 and `sex`=='Male' , Positive predictions[%]:  0.15 , TPR:  0.54
Slice  2 : `age`>=17.0 and `age`<=41.333 and `sex`=='Female' , Positive predictions[%]:  0.05 , TPR:  0.48
```

```
Slice  3 : `age`>41.333 and `age`<=65.667 and `sex`=='Male' , Positive predictions[%]:  0.39 , TPR:  0.69
Slice  4 : `age`>41.333 and `age`<=65.667 and `sex`=='Female' , Positive predictions[%]:  0.1 , TPR:  0.5
```

```
Slice  5 : `age`>65.667 and `age`<=90.073 and `sex`=='Male' , Positive predictions[%]:  0.17 , TPR:  0.65
Slice  6 : `age`>65.667 and `age`<=90.073 and `sex`=='Female' , Positive predictions[%]:  0.0 , TPR:  0.0
```

Come affrontare i bias con i dati sintetici?

- Con l'utilizzo dei dati sintetici è possibile migliorare le metriche di equità del modello generando punti sintetici per popolare specifiche porzioni dei dati con esempi della classe positiva.

- Per questo particolare esempio, aumentiamo il dataset creando esempi sintetici per le donne con reddito elevato nell'intervallo di età compreso tra 42 e 90 anni.

Come affrontare i bias con i dati sintetici?

Le metriche di equità sono migliorate con il dataset aumentato:

Original Dataset

Slice#	% Pos.	TPR
1 Male 17-41	15	0.54
2 Female 17-41	5	0.48
3 Male 41-65	39	**0.69**
4 Female 41-65	10	**0.5**
5 Male >65	17	0.65
6 Female >65	0.	0.

Augmented Dataset

Slice#	% Positive	TPR
1	12	0.49
2	5	0.4
3	40	**0.7**
4	20	**0.73**
5	18	0.62
6	5	0.2

Come affrontare i bias con i dati sintetici?

Performance del modello

Original Dataset

	precision	recall	f1-score
False	0.88	0.96	0.92
True	0.83	0.60	0.70
Weigh. avg	0.87	0.87	**0.86**

Augmented Dataset

	precision	recall	f1-score
False	0.88	0.95	0.92
True	0.81	0.60	0.69
Weigh. avg	0.86	0.87	**0.86**

Conclusioni

I dati sintetici, aumentando il numero di istanze della classe minoritaria di un dataset, permettono di ottenere un modello più robusto e meno soggetto a bias.

Quando si pensa all'implementazione di metriche di equità nei flussi di dati e nei processi di apprendimento automatico, dovresti sapere:

- Quali sono i tuoi obiettivi o quelli della tua azienda in materia di equità?
- Considerare il coinvolgimento di più stakeholder ed esperti per stabilire quali siano bias accettabili e quali no.
- Focalizzarsi sulla qualità dei dati.
- Continuare a testare: cercare di verificare con diverse suddivisioni e modalità di integrazione la presenza di bias nei dati.

clearbox AI
humans && machines

Thanks for the attention

Feel free to contact us:


www.clearbox.ai


simona@clearbox.ai


@ClearboxAI