



Opportunità, innovazione e rischi nell'era dell'Intelligenza Artificiale

Bias nei sistemi di Intelligenza Artificiale
Intervento

Ing. Alessandro Simonetta
Presidente Commissione Tematica di Intelligenza Artificiale

Roma, 22 maggio 2025

Breve Curriculum Vitae

LAVORO DIPENDENTE NELLA PA



- STAFF DELL'ISPETTORE GENERALE
- DIRETTORE DELLA DIREZIONE IT, LOGISTICA E TRANSIZIONE DIGITALE
13/10/2022 – 14/07/2024
- RESPONSABILE DELLA TRANSIZIONE DIGITALE – ART.17 CAD
12/12/2022 – 14/07/2024

01/09/22 – today



02/03/99 – 31/08/22

- PROFESSIONISTA INFORMATICO
- COORDINATORE DEL SETTORE - CONSULENZA PER L'INNOVAZIONE TECNOLOGICA

27 ANNI

LAVORO DIPENDENTE PRIVATO



01/02/95 - 30/06/95

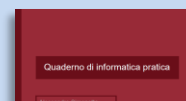


01/07/95 - 01/03/99

4 ANNI

PROFESSORE A CONTRATTO

- ALFABETIZZAZIONE INFORMATICA A.A. 2001/2002
- TEORIA DELL'INFORM. E TECNICHE DI ELABOR. DIGITALE
- TECNICHE INFORMATICHE
- FONDAMENTI DI PROGRAMMAZIONE
- CALCOLATORI E SISTEMI OPERATIVI
- ARCHITETTURE DEI SISTEMI DI ELABORAZIONE A.A. 2024/2025



24 ANNI



13/05/09 – today

UNI CT 504 «INGEGNERIA DEL SW» VICE PRESIDENTE - JTC1/SC7
UNI CT 510 «SICUREZZA» MEMBRO COMMISSIONE - JTC1/SC27
UNI CT 533 «INTELLIGENZA ARTIFICIALE» LIAISON OFFICER FROM UNI CT 510

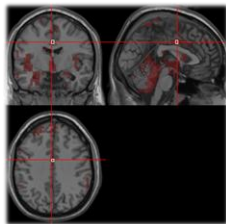


16 ANNI

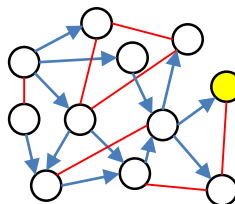
1994



Prof. Daniele Nardi



fMRI



BCI



INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION



INTERNATIONAL JOURNAL ON
ADVANCED SCIENCE
ENGINEERING
INFORMATION TECHNOLOGY

Editorial Team

30 ANNI



Ing. Alessandro Simonetta

Bias nei sistemi di Intelligenza Artificiale



Attività di ricerca recenti

BIAS / AI / SQUARE

- ISO/IEC STANDARDS AND DESIGN OF AN ARTIFICIAL INTELLIGENCE SYSTEM
- ETICA NEI SISTEMI DI INTELLIGENZA ARTIFICIALE
- APPLICAZIONE DELL'AI PER IL REINSERIMENTO SOCIALE E LAVORATIVO NELLA CONDUZIONE DI MACCHINARI COMPLESSI
- FAIRNESS METRICS AND MAXIMUM COMPLETENESS FOR THE PREDICTION OF DISCRIMINATION (*)
- THE USE OF MAXIMUM COMPLETENESS TO ESTIMATE BIAS IN AI BASED RECOMMENDATION SYSTEMS
- USING THE SQUARE SERIES AS A GUARANTEE FOR GDPR COMPLIANCE (*)
- INTEGRATING SQUARE DATA QUALITY MODEL WITH ISO 31000 RISK MANAGEMENT TO MEASURE AND MITIGATE SOFTWARE BIAS (*)
- METRICS FOR IDENTIFYING BIAS IN DATASETS (*)

(*) CITATI IN CEN/CLC/TR 18115:2024 "DATA GOVERNANCE AND QUALITY FOR AI WITHIN THE EUROPEAN CONTEXT"

SECURITY / SOFTWARE

- A FORENSIC METHODOLOGY FOR THE IDENTIFICATION OF ILLICIT DATA LEAKAGE
- CODE PROTECTION TECHNIQUES WHEN DISTRIBUTED IN SOURCE FORMAT: AN ADOBE CONNECT POD WRITTEN IN JAVASCRIPT

NUOVE ARCHITETTURE DI CALCOLO

- MULTI-VALUED LOGIC DIGITAL CIRCUITS FOR REALIZING A COMPLETE COMPUTER ARCHITECTURE
- A NEW APPROACH FOR DESIGNING OF COMPUTER ARCHITECTURES USING MULTI-VALUE LOGIC
- DESIGNING DIGITAL CIRCUITS IN MULTI-VALUED LOGIC

BENESSERE ORGANIZZATIVO E CS

- MIGLIORARE LE PERFORMANCE AZIENDALI INCREMENTANDO IL BENESSERE LAVORATIVO E SOCIALE
- UNA METODOLOGIA STATISTICO-INFORMATICA PER LE ANALISI DI BENESSERE ORGANIZZATIVO



Indice degli argomenti

- AI come leva per gli obiettivi del decennio digitale 2030
- Perché è nata l'esigenza di un Regolamento per la materia AI
- Cosa prescrive il Regolamento AI ACT in termini di analisi del rischio
- Panorama normativo internazionale per l'AI utile per l'analisi del rischio
- Case study: COMPAS dataset
 1. Identificazione di un unico indicatore per la misura di equità
 2. Indici di valutazione della qualità del dataset
 3. Relazione tra scarsa qualità nel training set e risultato non equo
 4. Annullare gli effetti nocivi del bias
- Conclusioni
- Bibliografia



Commissione
europea



Etica nei sistemi AI



Commissione
europea

Decennio digitale europeo: obiettivi digitali per il 2030

L'Europa mira a dare maggior forza alle imprese e ai cittadini in un futuro digitale incentrato sulla persona, sostenibile e più prospero.



https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_it



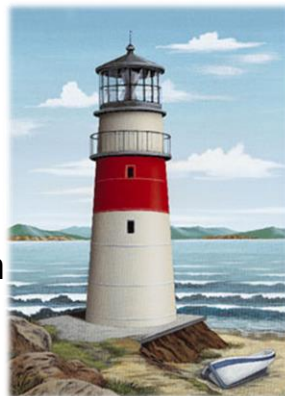
TRASFORMAZIONE DIGITALE DELLE IMPRESE

- 75% delle aziende UE che utilizzano Cloud, AI o Big Data
- Raddoppiare il numero delle startup unicorno in UE (1 miliardo di \$ senza essere quotata)
- 90% delle PMI che utilizzano le nuove tecnologie



COMPETENZE

- 20 milioni specialisti ICT e bilanciamento di genere
- 80% della popolazione con competenze digitali



DIGITALIZZAZIONE DEI SERVIZI PUBBLICI

- 100% servizi pubblici fondamentali online
- 100% dei cittadini con identità digitale eID e accesso alla cartella clinica digitale



INFRASTRUTTURE DIGITALI SICURE E SOSTENIBILI

- Connettività Gigabit per tutti
- Copertura mobile ad alta velocità (almeno 5G) ovunque
- EU arrivi al 20% della produzione mondiale dei semiconduttori
- 10.000 nodi edge cloud a impazzo zero
- Primo Computer Quantistico Europeo nel 2025

https://digital-strategy.ec.europa.eu/it/policies/europes-digital-decade#tab_3

Si vuole creare un ecosistema europeo di attori pubblici e privati che creino sistemi di IA in linea con i valori dell'Unione e sbloccino il potenziale della **trasformazione digitale** in tutte le regioni dell'Unione (rif. Bussola digitale 2030)

Il problema della analfabetizzazione è uno dei principali ostacoli per la diffusione dei sistemi di AI.

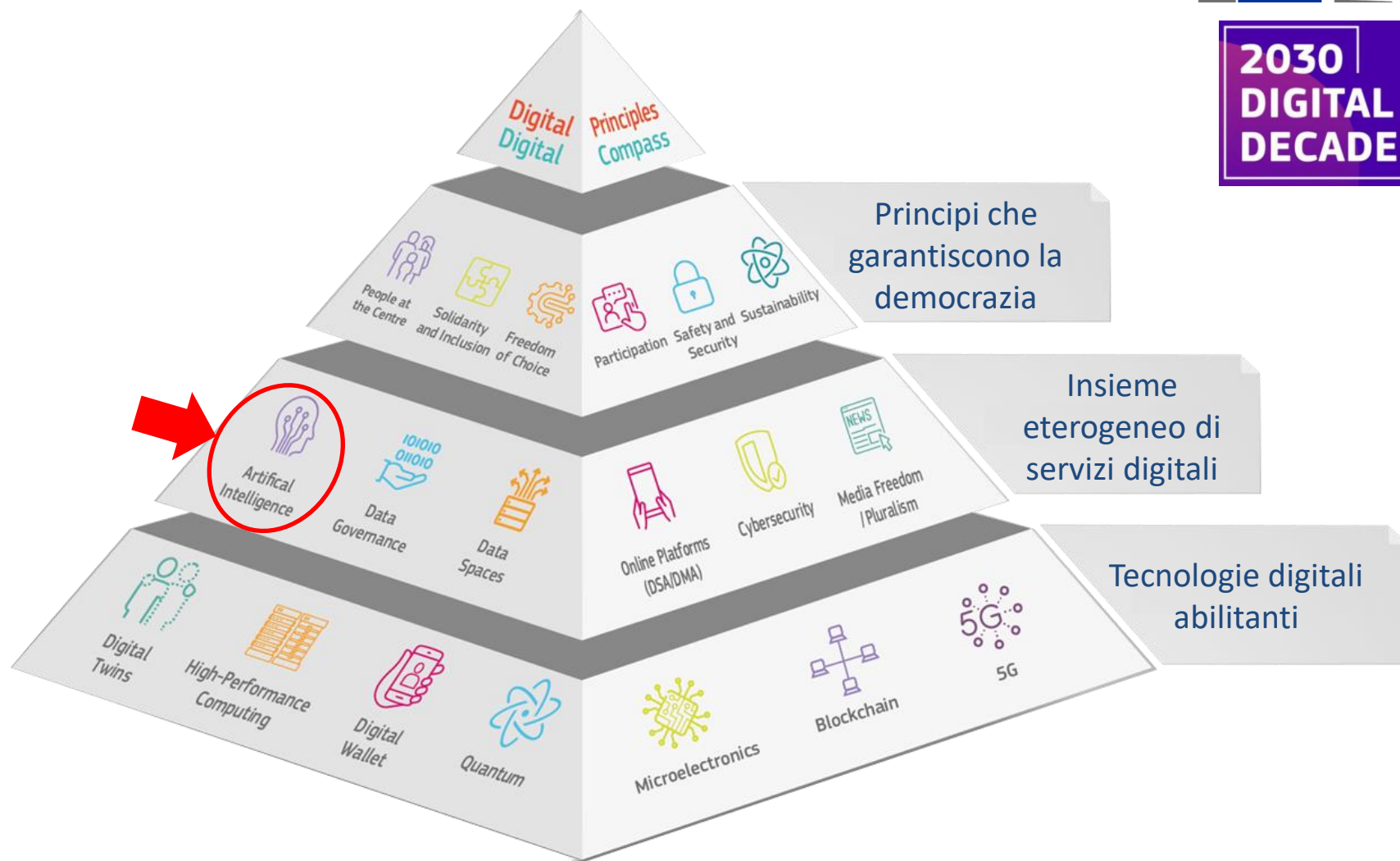
È necessaria la collaborazione di tutti i soggetti per superare la mancanza di conoscenza e anche all'adeguamento del regolamento da parte della Commissione.



Linee guida per lo sviluppo delle tecnologie digitali



2030
DIGITAL
DECADE



[https://wseas.com/journals/sac/2022/a345103-007\(2022\).pdf](https://wseas.com/journals/sac/2022/a345103-007(2022).pdf)

Aspettative nell'uso dell'AI

- La scelta è oggettiva senza influenze derivanti da preconconcetti sociali
- La risposta è immediata
- La valutazione considera un universo di informazioni quindi la migliore possibile
- La scelta può essere automatizzata e sostituire decisioni umane o fornire un supporto alle decisioni

Alcune di queste aspettative sono vere se e solo se i dati da cui sono stati costruiti i modelli di apprendimento sono di **Qualità**.

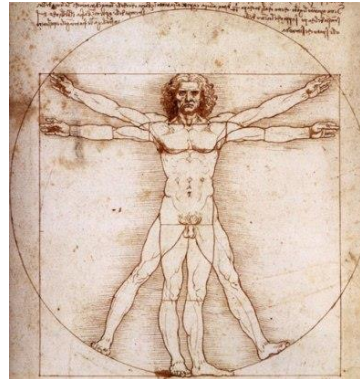


Etica nei sistemi AI



Lo scopo del regolamento AI ACT

- 1) Migliorare il funzionamento del mercato interno istituendo un **quadro giuridico uniforme** per quanto riguarda i sistemi di AI
- 2) promuovere la diffusione di un'intelligenza artificiale (IA) **antropocentrica** e affidabile (salute, sicurezza, il rispetto della Carta dei diritti fondamentali, lo Stato di Diritto, la democrazia, ambiente)



- 3) proteggere contro i possibili effetti nocivi dei sistemi di IA
- 4) promuovere l'innovazione impedendo la limitazione nello sviluppo, nella vendita e nell'utilizzo dei sistemi di IA
- 5) UE leader nell'adozione di un AI affidabile

Nel 2019 il gruppo AI HLEG ha elaborato **sette principi** etici non vincolanti per l'AI per avere garanzia di affidabilità ed etica per un sistema di AI.

Principi che dovrebbero essere **by-default** e **by-design**:

1. Intervento e sorveglianza umana



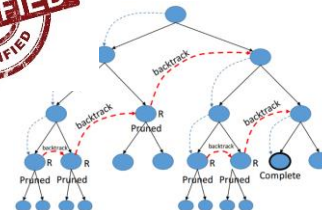
2. Robustezza tecnica e sicurezza (cybersecurity e resilienza)



3. Privacy e governance dei dati (**qualità dei dati**)



4. Trasparenza (tracciabilità e spiegazione)



5. Diversità, non discriminazione ed equità (**bias** o **pregiudizio**)



6. Benessere sociale e ambientale



7. Responsabilità (Accountability)



<https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai>

Analogamente al GDPR l'approccio adottato è basato sulla «**valutazione del rischio**»: la combinazione della probabilità del verificarsi di un danno e la gravità del danno stesso.



Il regolamento stabilisce sanzioni per il mancato rispetto delle norme, con multe che possono arrivare fino al maggiore tra **35 ml** euro e il **7% del fatturato globale** dell'azienda.

Classifica i sistemi di intelligenza artificiale in base al livello di rischio che rappresentano per la società e i diritti fondamentali secondo le quattro categorie.

Classificazione dei sistemi di AI in base al rischio



Un sistema di IA è considerato ad alto rischio se:

- è parte o rientra nei sistemi soggetti a valutazione ai fini di sicurezza previsti dall'**Allegato I** (direttiva macchina, giocattoli, imbarcazioni, ascensori, dispositivi medici,...)



Un sistema di IA è considerato ad alto rischio se è utilizzato in un contesto previsto dall'**Allegato III**:

1. biometria/emozioni (esclusa l'autenticazione) 
2. infrastrutture critiche (digitali, traffico, luce, gas,...) 
3. istruzione e formazione professionale 
4. occupazione, gestione dei lavoratori 
5. servizi privati essenziali (enti locali, assistenza sanitaria,...) 
6. attività di contrasto (law enforcement) 
7. migrazione, asilo e gestione del controllo delle frontiere 
8. amministrazione della giustizia e processi democratici (elezioni/referendum) 

Un sistema di AI che rientra nell'Allegato III non è ad alto rischio se c'è stata un'**analisi preventiva del rischio** che documenta che:

- 1) non presenta un rischio significativo di danno per la salute, la sicurezza o i diritti fondamentali delle persone fisiche,
- 2) non influenza materialmente il risultato del processo decisionale.

Si applica quando è soddisfatta almeno una delle seguenti condizioni:

- a. un coinvolgimento limitato nel processo,
- b. migliorare il risultato di un'attività umana già completata, l'AI è il supporto all'attività umana,
- c. non è finalizzato a sostituire o influenzare la valutazione umana
- d. ha solo il compito di effettuare un compito preparatorio.

1. In relazione ai sistemi di IA ad alto rischio è (1) istituito (si attiva nella realtà un sistema di gestione del rischio), (2) attuato, (3) documentato (accountability) e (4) mantenuto (manutenzione costante) un sistema di gestione dei rischi.
2. Il **sistema di gestione dei rischi** è un processo iterativo continuo pianificato ed eseguito nel corso dell'intero ciclo di vita di un sistema di IA ad alto rischio, che richiede un riesame e un aggiornamento costanti e sistematici. Le fasi sono:
 - a. **identificazione e analisi dei rischi** noti e ragionevolmente prevedibili per come il sistema è stato pensato;
 - b. **stima e valutazione dei rischi** che possono emergere quando il sistema di IA ad alto rischio è usato in condizioni di **uso improprio ragionevolmente prevedibile**;
 - c. **valutazione di altri eventuali rischi** derivanti dall'analisi dei dati raccolti dal sistema di monitoraggio successivo all'immissione sul mercato di cui all'articolo 72;
 - d. **adozione di misure di gestione dei rischi** opportune e mirate intese ad affrontare i rischi individuati ai sensi della lettera a).

Qualità dei dati: i set di dati di addestramento, dovrebbero essere pertinenti, sufficientemente rappresentativi e, nella misura del possibile, esenti da errori e **completi**. I risultati forniti dai sistemi di IA potrebbero essere influenzati da tali **distorsioni intrinseche** e quindi a **perpetuare e amplificare le discriminazioni esistenti**, in particolare nei confronti delle persone che appartengono a determinati gruppi vulnerabili, inclusi gruppi razziali o etnici.

..i dati di addestramento, convalida e prova devono essere: di **Qualità** e sufficientemente **rappresentativi** (secondo le proprietà statistiche appropriate) per quanto riguarda le persone o i gruppi di persone.

Come si misura la Qualità dei dati e il rischio che bias nei dati si possa perpetrare nelle previsioni?

Chi sviluppa le norme?



- Il **Joint Technical Committee 1** (JTC 1) è il comitato tecnico congiunto che si occupa degli aspetti di standardizzazione dell'IT per ISO ed IEC.



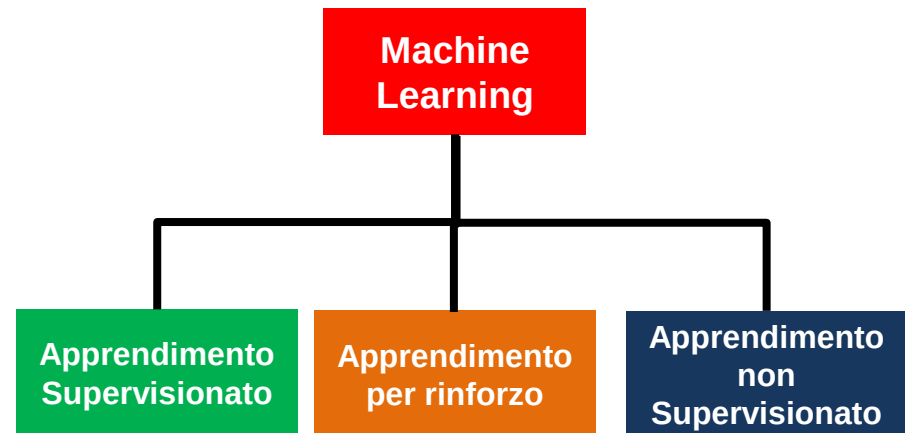
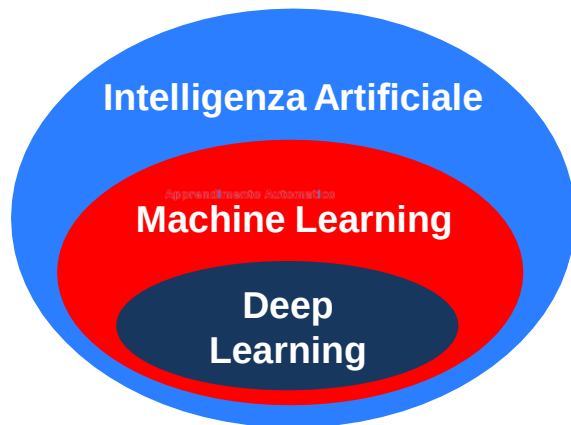
Panorama normativo per AI sviluppato in ambito internazionale

Life Cycle ISO/IEC/IEEE 12207:18 Sw life cycle process	Foundation ISO/IEC 2382:15 IT Vocabulary			ISO/IEC 31000:18 Risk Mgmt Guidelines	ISO 9001:15 Quality Mgmt Systems — Requir.	Software ISO/IEC 25000:14 System & SW engineering — SQuaRE	ISO/IEC 29119:24 SW testing (11 parts))
Life Cycle ISO/IEC TR 5338:23 AI Life Cycle Process	Foundation ISO/IEC 22989:22 IA concepts & term.			ISO/IEC 23053:22 Framework AI Systems Using ML			
AI Governance ISO/IEC 38507:22 AI Governance implic.	Risk Mgmt ISO/IEC 23894:23 AI Risk Mgmt	Quality ISO/IEC CD TR 42106:xx Overview benchmark vs AI quality characteristic		Data Quality (DQ) for Analytics & ML ISO/IEC 24668:22 Process Mgmt Framework for Big Data Analytics		Trustworthiness ISO/IEC TR 24028:20 AI Trustworthiness	Testing ISO/IEC TR 29119-11:24 Guidelines AI Testing
CEN/CLC TR 18115:24 AI Data governance & quality	Functional Safety ISO/IEC TR 5469:24 AI Functional Safety	ISO/IEC 25059:23 AI Quality Model		ISO/IEC 5259-1:24 DQ for Analytics and ML – Overview & examples		ISO/IEC TR 24027:21 AI Bias	ISO/IEC 42119:xx AI Testing
AI Mgmt System ISO/IEC 42001:23 AI – Mgmt System		Assessment ISO/IEC TR 24029:21 Robustness of Neural Network		Application ISO/IEC TR 24030:24 AI use cases		ISO/IES DTS 6254:xx Objectives for AI explainability and interpretability	Big Data ISO/IEC 20546:19 Overview & Vocabulary
ISO/IEC 42005:xx AI impact assessment		ISO/IEC TS 4213:22 ML performance		ISO/IEC 5339:24 AI Application Guidance		ISO/IEC TS 8200:24 AI controllability	
ISO/IEC 42006:xx Req. for audit & cert.		ISO/IEC TR 24372:21 Overview of computational approaches		ISO/IEC 5259-4:24 DQ for Analytics and ML – Process Framework		ISO/IEC TR 24368:22 AI Overview of ethical concerns	
				ISO/IEC 5259-5:25 DQ for Analytics and ML – Governance Framework			ISO/IEC 20547:20 Series (5 parts)
		ISO/IEC 5392:24 Ref. Architecture of knowledge eng.		ISO/IEC CD TR 5259-6:xx DQ for Analytics and ML – Visualization Framework			



Ing. Alessandro Simonetta

Bias nei sistemi di Intelligenza Artificiale



Apprendimento supervisionato: si costruisce un modello a partire da dati di addestramento già etichettati, in questo modo è possibile effettuare delle predizioni.

Apprendimento per rinforzo: si costruisce un sistema che migliora le proprie prestazioni sulla base di interazioni con l'ambiente, attraverso un segnale di ricompensa (reward). Si tratta di un caso particolare dell'apprendimento supervisionato.

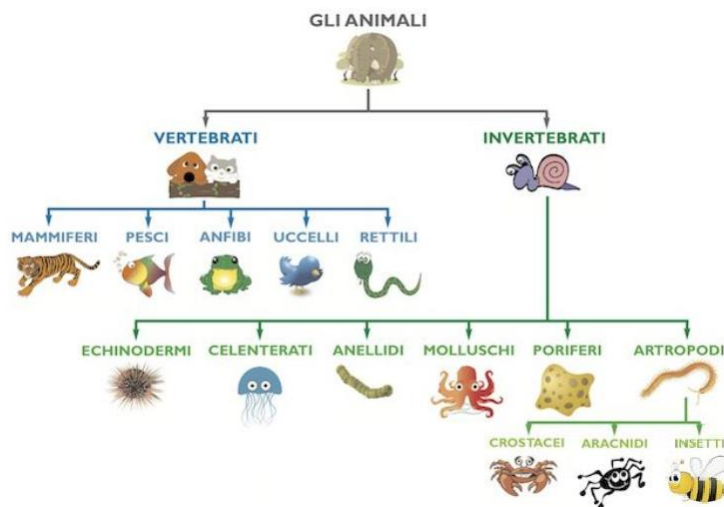
Apprendimento non supervisionato: il sistema è in grado di estrarre informazioni utili dai dati senza la guida di una etichetta risultato o di una funzione di ricompensa. È in grado di identificare cluster nei dati.

Etica nei sistemi AI

L'apprendimento supervisionato

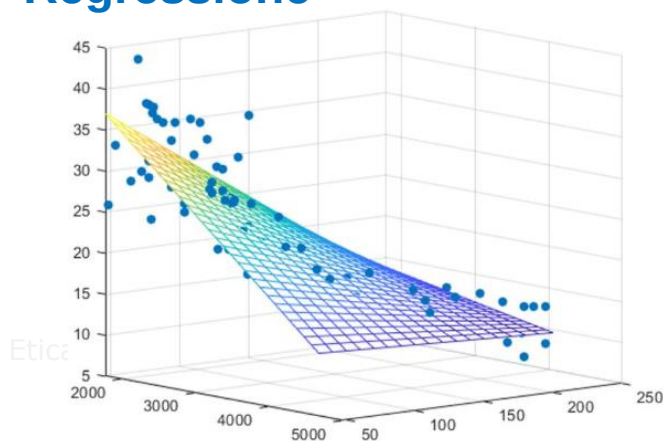
L'apprendimento supervisionato si divide tipicamente in due categorie a seconda del dominio dei valori delle variabili target: discreto o continuo.

Classificazione



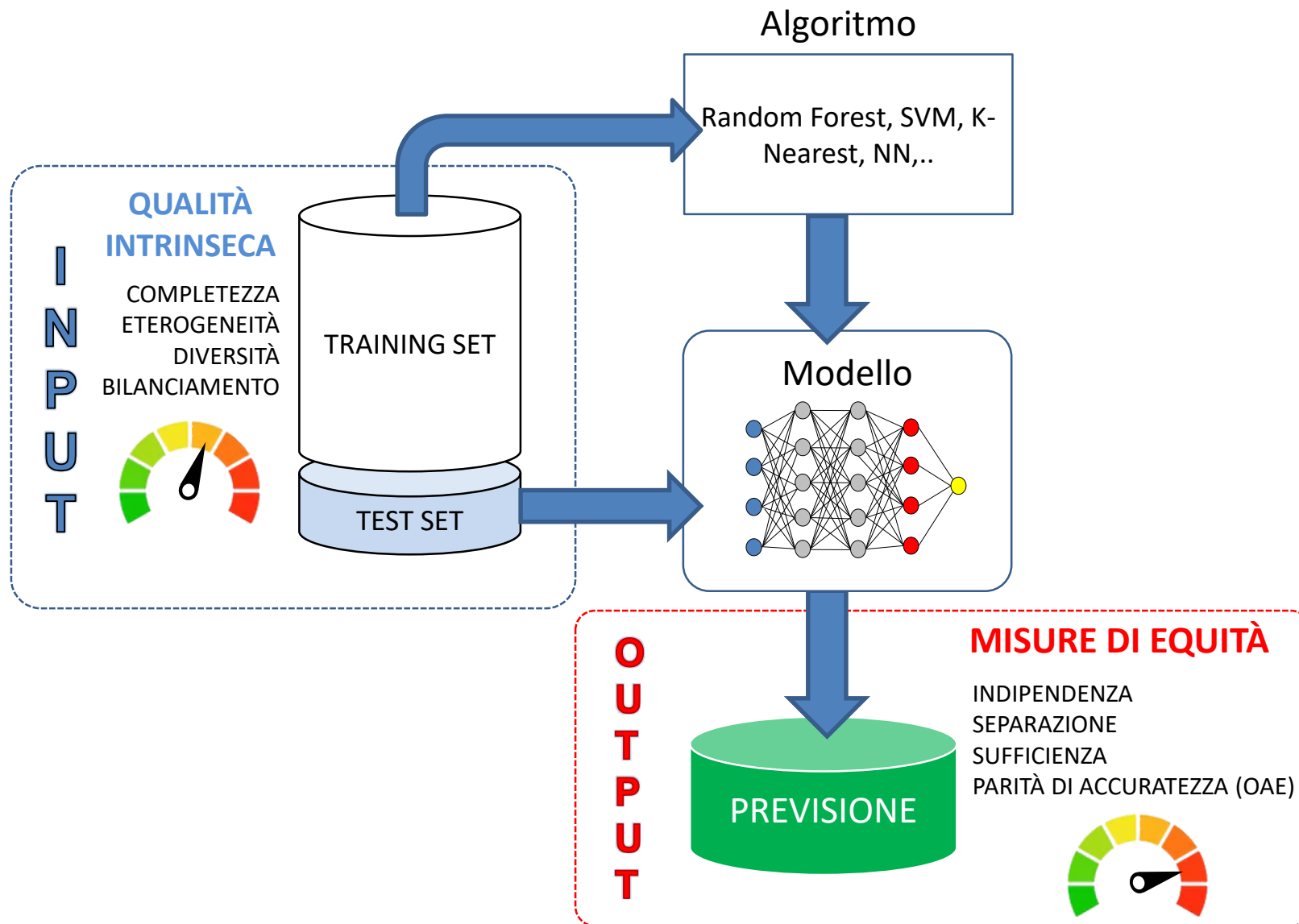
prevede variabili target categoriche, che rappresentano classi o etichette **discrete**.
Per esempio, classificare le e-mail come spam o non spam, o prevedere la necessità di sostituzione di un pezzo meccanico in base all'usura. Gli algoritmi di classificazione sono in grado di corrispondere alle caratteristiche in ingresso un valore discreto nell'ambito di classi predefinite.

Regressione



prevede variabili numeriche target **continue**.
Ad esempio, la previsione del premio assicurativo in base alla tipologia di rischio, o la probabilità di contrarre una malattia in base alle analisi. Gli algoritmi di regressione sono in grado di corrispondere alle caratteristiche in ingresso un valore numerico reale.

Dove e cosa misurare?





Formalizzazione della classificazione

<https://fairmlbook.org/classification.html>

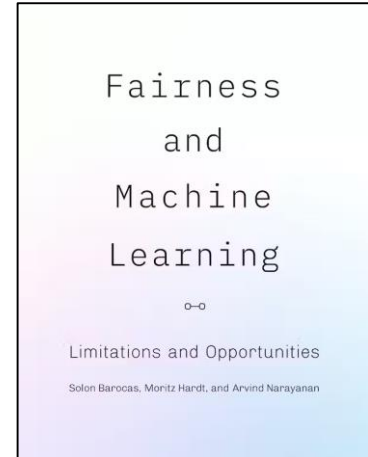
L'obiettivo della classificazione è determinare un valore plausibile per una variabile target Y sconosciuta a partire dalle covariate osservate X :

$$\hat{Y} = f(X)$$

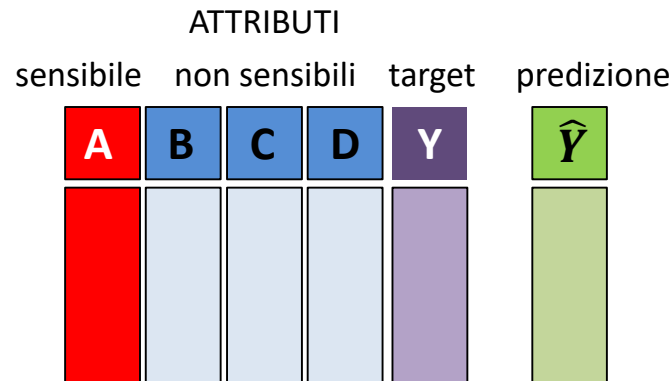
La funzione f è chiamato *classificatore* o *predittore*.

L'output del classificatore è chiamato *etichetta* o *predizione*.

Le covariate X e target Y sono variabili casuali distribuite congiuntamente. Ciò significa che esiste una distribuzione di probabilità su coppie di valori (x,y) che le variabili casuali (X,Y) possono assumere. Questa distribuzione di probabilità modella una popolazione di istanze del problema di classificazione.



Solon Barocas, Moritz Hardt,
Arvind Narayanan

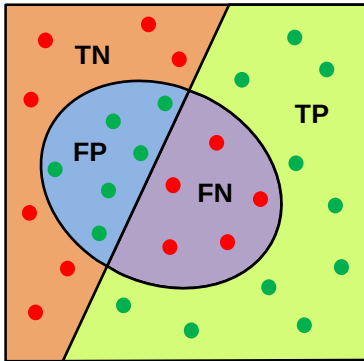


$$\hat{Y} = \begin{cases} 1 & R \geq t \\ 0 & R < t \end{cases}$$

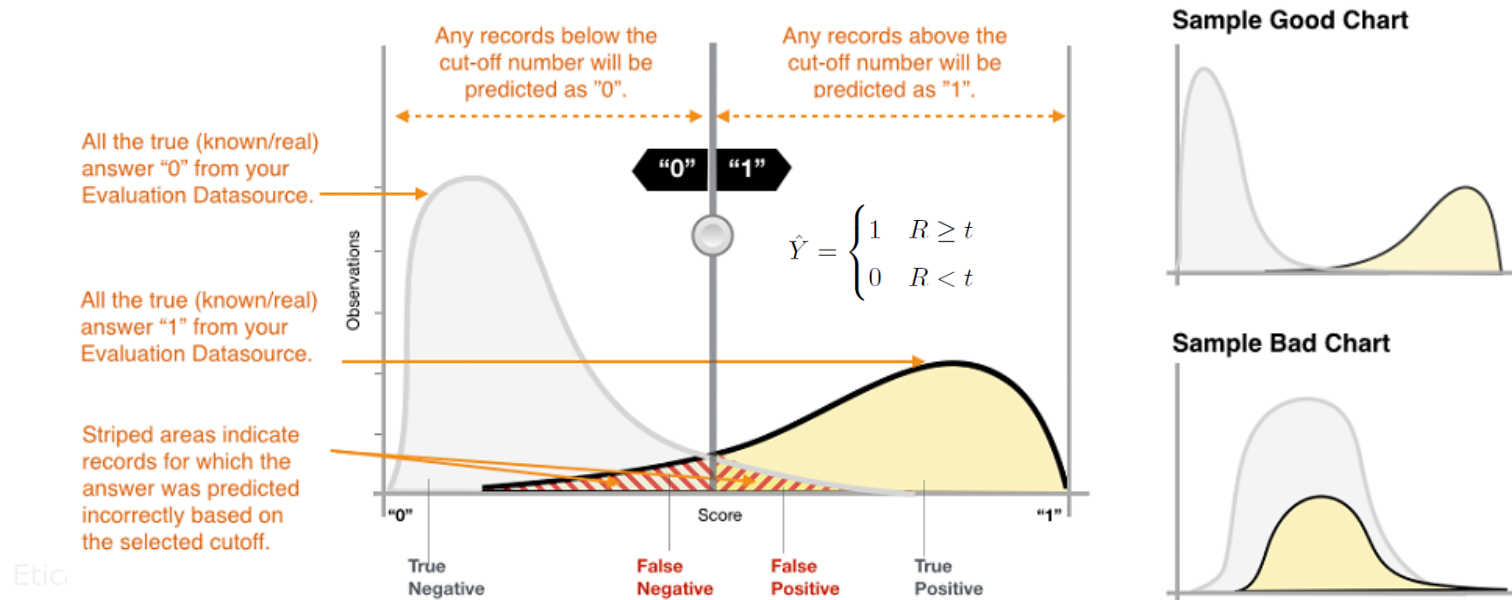


Formalizzazione della classificazione

<https://fairmlbook.org/classification.html>



Common classification criteria		
Event	Condition	Resulting notion ($\mathbb{P}\{\text{event} \mid \text{condition}\}$)
$\hat{Y} = 1$	$Y = 1$	True positive rate, recall
$\hat{Y} = 0$	$Y = 1$	False negative rate
$\hat{Y} = 1$	$Y = 0$	False positive rate
$\hat{Y} = 0$	$Y = 0$	True negative rate



<https://developers.google.com/machine-learning/crash-course/classification?hl=it>

Ing. Alessandro Simonetta

Bias nei sistemi di Intelligenza Artificiale



Metodi di valutazione di equità statistici

L'accuratezza e la precisione sono due delle 15 caratteristiche di qualità riconosciute nello dello standard **ISO/IEC 25012** e del **5259-2** (Data quality measures)

Matrice di confusione

	$\hat{Y} = 1$	$\hat{Y} = 0$
$Y=1$	TP	FN
$Y=0$	FP	TN

$$PPV = \frac{TP}{TP + FP} = 1 - FDR$$

Positive Predictive Value o **Precision**

$$FOR = \frac{FN}{FN + TN} = 1 - NPV$$

False Omission Rate

$$FDR = \frac{FP}{TP + FP} = 1 - PPV$$

False Discovery Rate

$$NPV = \frac{TN}{FN + TN} = 1 - FOR$$

Negative Predictive Value

Separazione

$TPR = \frac{TP}{TP + FN} = 1 - FNR$ True Positivve Rate o Recall o Sensitivity	$FNR = \frac{FN}{TP + FN} = 1 - TPR$ False Negative Rate o False Rejection Rate (FRR) (*)
$FPR = \frac{FP}{FP + TN} = 1 - TNR$ False Positive Rate o True Acceptance Rate (TAR) (*)	$TNR = \frac{TN}{FP + TN} = 1 - FPR$ True Negative Rate o Specificity

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Prevalence = \frac{TP + FN}{TP + FP + TN + FN}$$

Sufficienza

(*) Advances in Computer Vision and Pattern Recognition
Handbook of BiometricAnti-Spoofing
Trusted Biometrics under Spoofing Attacks
Sébastien Marcel, Mark S. Nixon, Stan Z. Li



Metodi di valutazione di equità statistici

<https://fairmlbook.org/classification.html>

$$\mathbb{P}\{\hat{Y} = 1|A = a_i\} = \mathbb{P}\{\hat{Y} = 1|A = a_j\} \forall_{i,j}$$

IND Indipendenza

$$\mathbb{P}\{\hat{Y} = Y|A = a_i\} = \mathbb{P}\{\hat{Y} = Y|A = a_j\} \forall_{i,j}$$

Overall
Accuracy
Equality

OAE

Parità di
accuratezza
complessiva

$$\mathbb{P}\{\hat{Y} = 1|Y = 1, A = a_i\} = \mathbb{P}\{\hat{Y} = 1|Y = 1, A = a_j\} \forall_{i,j}$$

TPR

$$\mathbb{P}\{\hat{Y} = 1|Y = 0, A = a_i\} = \mathbb{P}\{\hat{Y} = 1|Y = 0, A = a_j\} \forall_{i,j}$$

FPR

Separazione

$$\mathbb{P}\{Y = 1|\hat{Y} = 1, A = a_i\} = \mathbb{P}\{Y = 1|\hat{Y} = 1, A = a_j\} \forall_{i,j}$$

PPV

$$\mathbb{P}\{Y = 1|\hat{Y} = 0, A = a_i\} = \mathbb{P}\{Y = 1|\hat{Y} = 0, A = a_j\} \forall_{i,j}$$

NPR

Sufficienza

Etica nei sistemi AI



La misura di indipendenza

Senza perdita di generalità considereremo la misura di indipendenza che è quella più intuitiva. L'indipendenza è definita tra due gruppi di attributi sensibili come distanza tra le due probabilità:

$$Ind(a_i, a_j) = |\mathbb{P}\{\hat{Y} = 1 | A = a_i\} - \mathbb{P}\{\hat{Y} = 1 | A = a_j\}|$$



Obiettivi

- 1** trovare un modo per sintetizzare in un'unica dimensionalità la nozione di indipendenza, essendo definita tra coppie di valori di un attributo sensibile (misura sull'OUTPUT)
- 2** individuare un indice che permetta di evidenziare la scarsa qualità nei training set (misura sull'INPUT)
- 3** studiare la relazione che sussiste tra scarsa qualità nei training set e risultato iniquo e, se è possibile, anticipare la probabilità che un bias nei dati si possa perpetrare anche nelle previsioni (INPUT vs OUTPUT).
- 4** annullare gli effetti nocivi del bias nei dati quando non si hanno a disposizione ulteriori dati di apprendimento



Case study: COMPAS dataset

PRO PUBLICA

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals.
And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Two Drug Possession Arrests

DYLAN FUGETT	BERNARD PARKER
RISK: 3	RISK: 10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Il sistema prevedeva erroneamente un maggior grado di recidività dei detenuti di razza afro-americana, mentre nella realtà accadeva che i detenuti bianchi avevano una maggiore propensione a reiterare i reati.



Calcolo dell'indipendenza tra le diverse etnie

Ricordando la nozione di indipendenza tra due gruppi di attributi sensibili:

$$Ind(a_i, a_j) = \left| \mathbb{P}\{\hat{Y} = 1 | A = a_i\} - \mathbb{P}\{\hat{Y} = 1 | A = a_j\} \right|$$

$A = a_i$	$P(\hat{Y}=1 A = a_i)$
Caucasian	0.33
Hispanic	0.28
Other	0.20
Asian	0.23
African-American	0.58
Native-American	0.73

	Asian	Caucasian	Hispanic	Native American	Other
African-American	35,03%	24,51%	29,90%	15,12%	37,20%
Asian		10,51%	5,12%	50,15%	2,17%
Caucasian			5,39%	39,63%	12,69%
Hispanic				45,03%	7,29%
Native American					52,32%

Etica nei sistemi AI





1 Identificazione di un unico indicatore (misura sull'OUTPUT)

La situazione ideale di equità perfetta per un attributo sensibile A si verifica quando tutte le probabilità $\mathbb{P}\{\hat{Y} = 1 | A = a_i\}$ hanno il medesimo valore.

Un indice di indipendenza rappresentativo dell'insieme delle misure di indipendenza tra coppie di valori dell'attributo sensibile si può calcolare:

- A** considerando la media dei valori di indipendenza tra tutte le coppie di valori dell'attributo sensibile
- B** In relazione alla distanza tra gruppi di probabilità affini per trattamento
- C** In base alla massima sperequazione tra valori di probabilità $\mathbb{P}\{\hat{Y} = 1 | A = a_i\}$
- D** Utilizzando la nozione di mutua informazione e di entropia

Etica nei sistemi AI

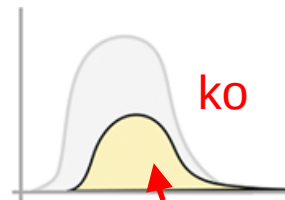
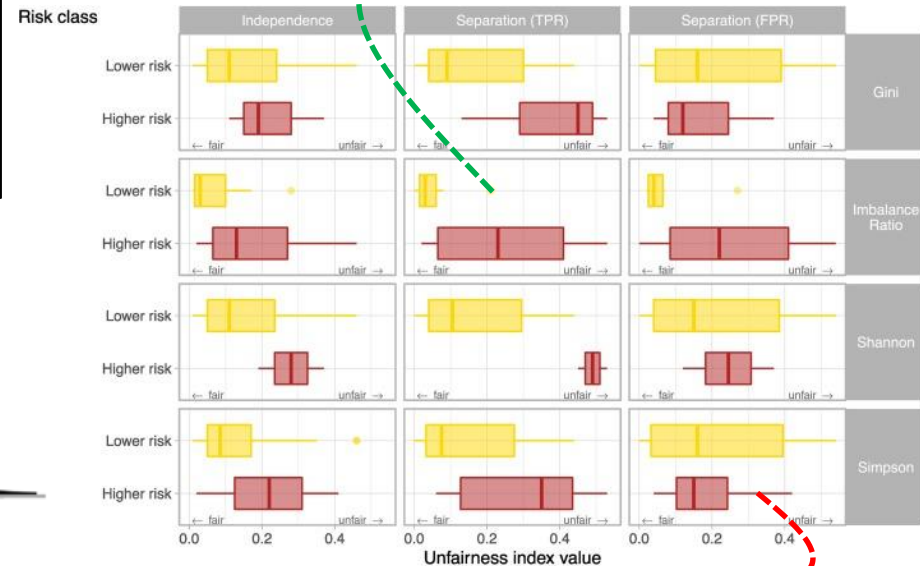


Utilizzo della media delle distanze

Alcuni autori hanno proposto la media delle misure di indipendenza:

$$\mathfrak{U}(a_1, \dots, a_m) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^m |P(R=1|A=a_i) - P(R=1|A=a_j)|$$

$$Ind(A) = \frac{2}{m \cdot (m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^m Ind(a_i, a_j) = 24,8 \%$$



Etica nei sistemi AI



Evidenza dei differenti gruppi di trattamento

La conferma dell'esistenza di più gruppi di trattamento è provato considerando il concetto di equità come un **vettore multidimensionale** le cui componenti sono le singole misure di equità (indipendenza, TPR, FPR, PPV, NPV e OAE).

Le relazioni tra i vettori, calcolate attraverso l'indice di Pearson, permettono di raggruppare le etnie in base alle analogie di trattamento.

Race	Fairness Index					
	<i>Ind.</i>	<i>SepTPR</i>	<i>SepFPR</i>	<i>SufPPV</i>	<i>SufNPV</i>	<i>OAE</i>
Caucasian	33,10%	50,36%	22,01%	59,48%	29,00%	67,19%
Hispanic	27,70%	41,80%	19,38%	56,03%	29,89%	66,21%
Other	20,41%	33,87%	12,79%	60,00%	30,04%	67,93%
Asian	22,58%	62,50%	8,70%	71,43%	12,50%	83,87%
African American.	57,61%	71,52%	42,34%	64,95%	35,14%	64,91%
Native American.	72,73%	100%	50%	62,50%	0,00%	72,73%

Race	Race					
	<i>African-A.</i>	<i>Native A.</i>	<i>Caucasian</i>	<i>Hispanic</i>	<i>Other</i>	<i>Asian</i>
African-American	1	0,901	0,801	0,680	0,562	0,848
Native American	0,901	1	0,515	0,364	0,210	0,596
Caucasian	0,801	0,515	1	0,983	0,941	0,991
Hispanic	0,680	0,364	0,983	1	0,984	0,956
Other	0,562	0,210	0,941	0,984	1	0,900
Hispanic	0,848	0,596	0,991	0,956	0,900	1

Groups	Races			
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
G0	African-American	Native American		
G1	Caucasian	Hispanic	Other	Asian

Etica nei sistemi A

B

La nozione di indipendenza tra gruppi

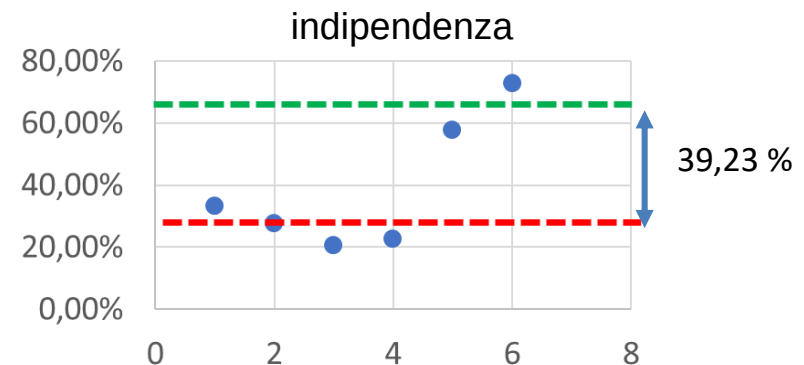
Utilizzando le tecniche di clustering (come l'algoritmo DBSCAN) è possibile raggruppare i dati per aree omogenee di trattamento, che avranno anche valori simili di probabilità $\mathbb{P}\{\hat{Y} = 1 | A = a_i\}$.

A questo punto è possibile eleggere i centri di massa dei cluster come rappresentativi dei gruppi e applicare la media delle distanze ().

A

Nel dataset COMPAS le probabilità congiunte si addensano su due centroidi rappresentativi dei due gruppi di trattamento.

$A = a_i$	$P(\hat{Y}=1 A = a_i)$	Centroid
Caucasian	0.33	0.26
Hispanic	0.28	
Other	0.20	
Asian	0.23	0.65
African-American	0.58	
Native-American	0.73	



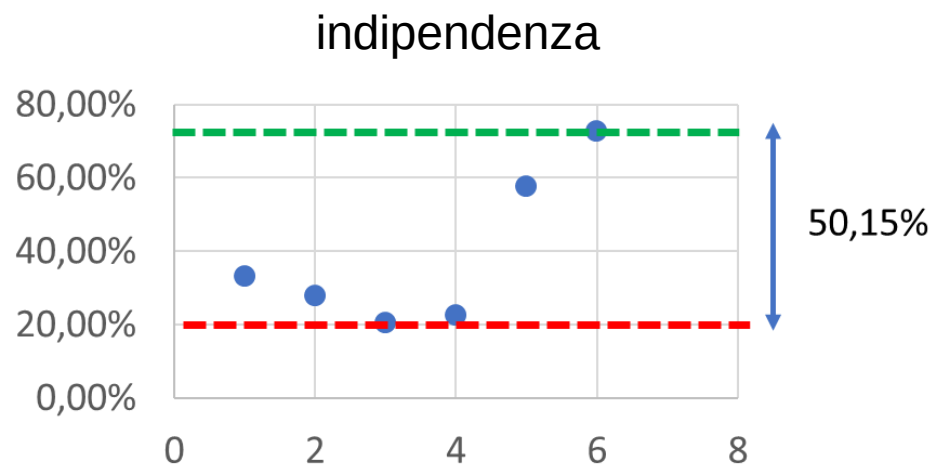
Ciò significa una equità di indipendenza all'interno del Gruppo e, quindi una iniquità tra gruppi.

Etica nei sistemi AI



Massima disparità di trattamento (algoritmo MaxMin)

Questo metodo si basa sull'idea che una misura dell'entità di ingiustizia si può calcolare considerando il caso peggiore, cioè il valore più alto di indipendenza.





Mutua Informazione ed Entropia

Date due variabili casuali, esse sono indipendenti se la loro informazione reciproca è zero:

$$I(\hat{Y}, A) = 0$$

Possiamo calcolare l'informazione reciproca utilizzando la nozione di entropia:

$$I(\hat{Y}, A) = H(\hat{Y}) + H(A) - H(\hat{Y}, A)$$

Da cui nello specifico:

$$\sum_{i=1}^n P(r_i) \log(P(r_i)) + \sum_{i=1}^n P(a_i) \log(P(a_i)) - \sum_{i=1, j=1}^{n, m} P(r_i \cap a_j) \log(P(r_i \cap a_j))$$

Utilizzando questo approccio è possibile calcolare tutti gli indici di equità analizzati in precedenza (indipendenza, sufficienza, separazione e parità di accuratezza complessiva).

Etica nei sistemi AI

2 Indici di valutazione della qualità del dataset (misure sull'INPUT)

I principali indici statistici utilizzati in letteratura sono:

- **Indice di Gini**, misura la disuguaglianza nella distribuzione (eterogeneità) di una variabile in un dataset
- **Indice di Shannon**, misura la diversità di specie in un ecosistema
- **Imbalance ratio**, rappresenta il grado in cui una classe di dati è più frequente di un'altra classe
- **Indice di Simpson**, misura la diversità di un gruppo di individui, specificamente la probabilità che due individui scelti a caso appartengano alla stessa specie

Indice di Gini (eterogeneità)	$G_n = \frac{m}{m-1} \cdot \left(1 - \sum_{i=1}^m f_i^2\right)$	Indice di Shannon (diversità)	$H' = -\frac{1}{\ln m} \sum_{i=1}^m f_i \ln f_i$
Rapporto di squilibrio	$I_n = \frac{\min f_i}{\max f_i}$	Indice di Simpson (diversità)	$D_n = \frac{1}{m-1} \left(\frac{1}{\sum_{i=1}^m f_i^2} - 1 \right)$



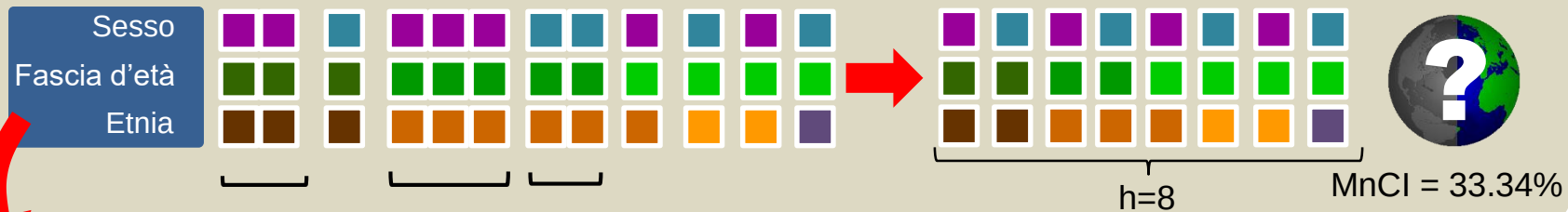
2 La nozione di completezza

La nozione di completezza è presente nella ISO/IEC 25012 ed è stata riportata anche nella ISO/IEC 5259-2 (Artificial intelligence — Data quality for analytics and machine learning (ML) - Part 2: Data quality measures).

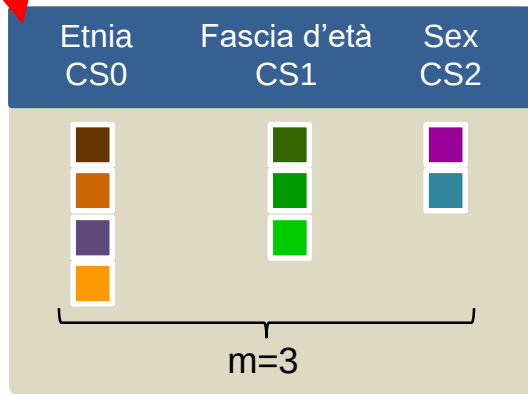
Completezza Minima	Misura il rapporto tra il numero distinto di tuple con attributi sensibili e il caso ideale in cui il dataset contiene tutte le possibili combinazioni distinte di attributi sensibili.
Completezza Massima	Misura il rapporto tra il numero di tuple con attributi sensibili e il caso ideale in cui il dataset contiene il numero di combinazioni distinte di attributi sensibili ripetuti esattamente il numero di volte della combinazione predominante

2 Calcolo della completezza minima

Dataset di esempio da analizzare



Attributi sensibili

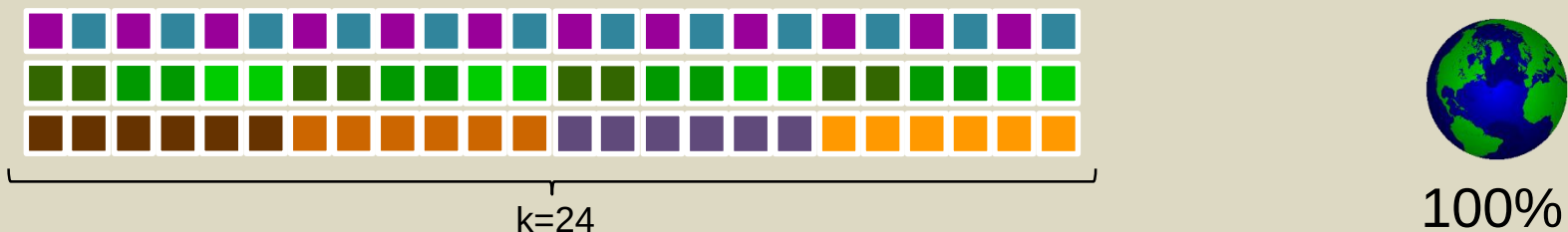


```
k= ( df['CS0'].unique().size *  
      df['CS1'].unique().size *  
      df['CS2'].unique().size )  
h=len (df.drop_duplicates())
```

Completezza minima

$$MnCI = h / k$$

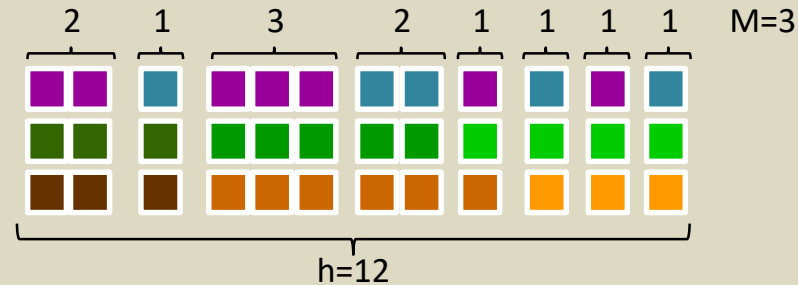
Dataset ideale





2 Calcolo della completezza massima

Dataset di esempio da analizzare



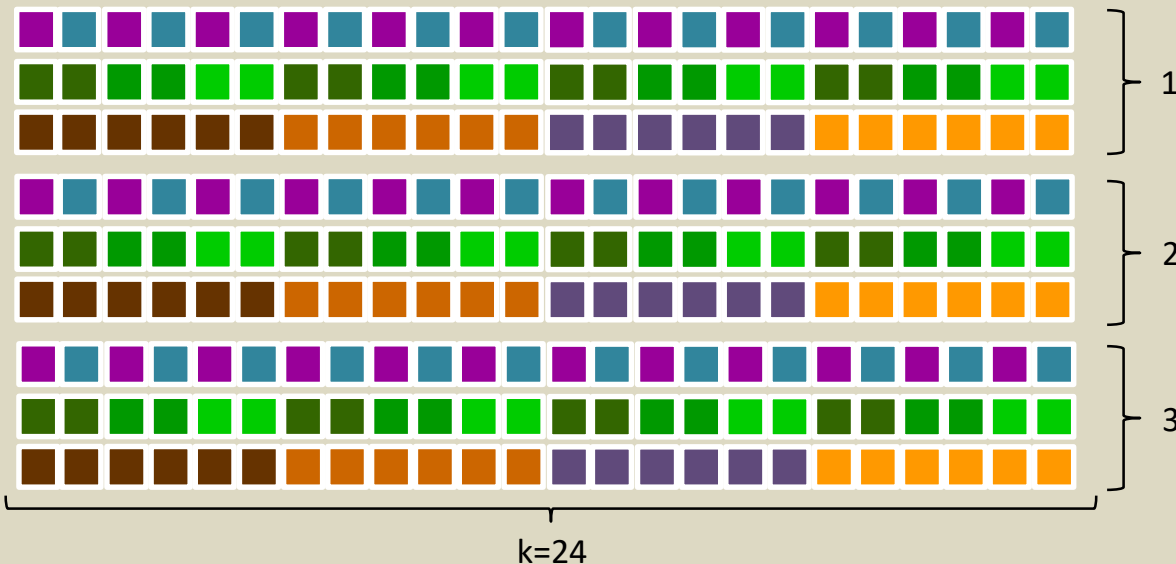
MxCI = 16.67%

```
M=df.groupby(['CS0','CS1','CS2']).size()  
.reset_index(name='counts').counts.max()  
h=len (df)
```

Completezza Massima

$$MxCI = h / (M*k)$$

Dataset ideale



100%



3 Relazione tra scarsa qualità nel training set e risultato iniquo

La sfida è nella capacità di anticipare le disparità di trattamento nei risultati di un sistema di AI valutando la qualità dei dati nel training set.

L'idea è trovare dei marcatori predittivi per confinare il rischio che un difetto nei dati si possa propagare all'interno nel sistema di apprendimento, perpetrando, o persino amplificando, pregiudizi della società rispetto a minoranze etniche, genere,...



Dataset utilizzati nella ricerca

- COMPAS Recidivism Dataset

<https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis>

- Recidivism in juvenile justice

<https://www.ojjdp.gov/ojstatbb/compendium/>

- UCI Statelog German Credit

<https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data>

- default of credit card clients Data Set

<https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients>

- Adult Data Set

<https://archive.ics.uci.edu/dataset/2/adult>

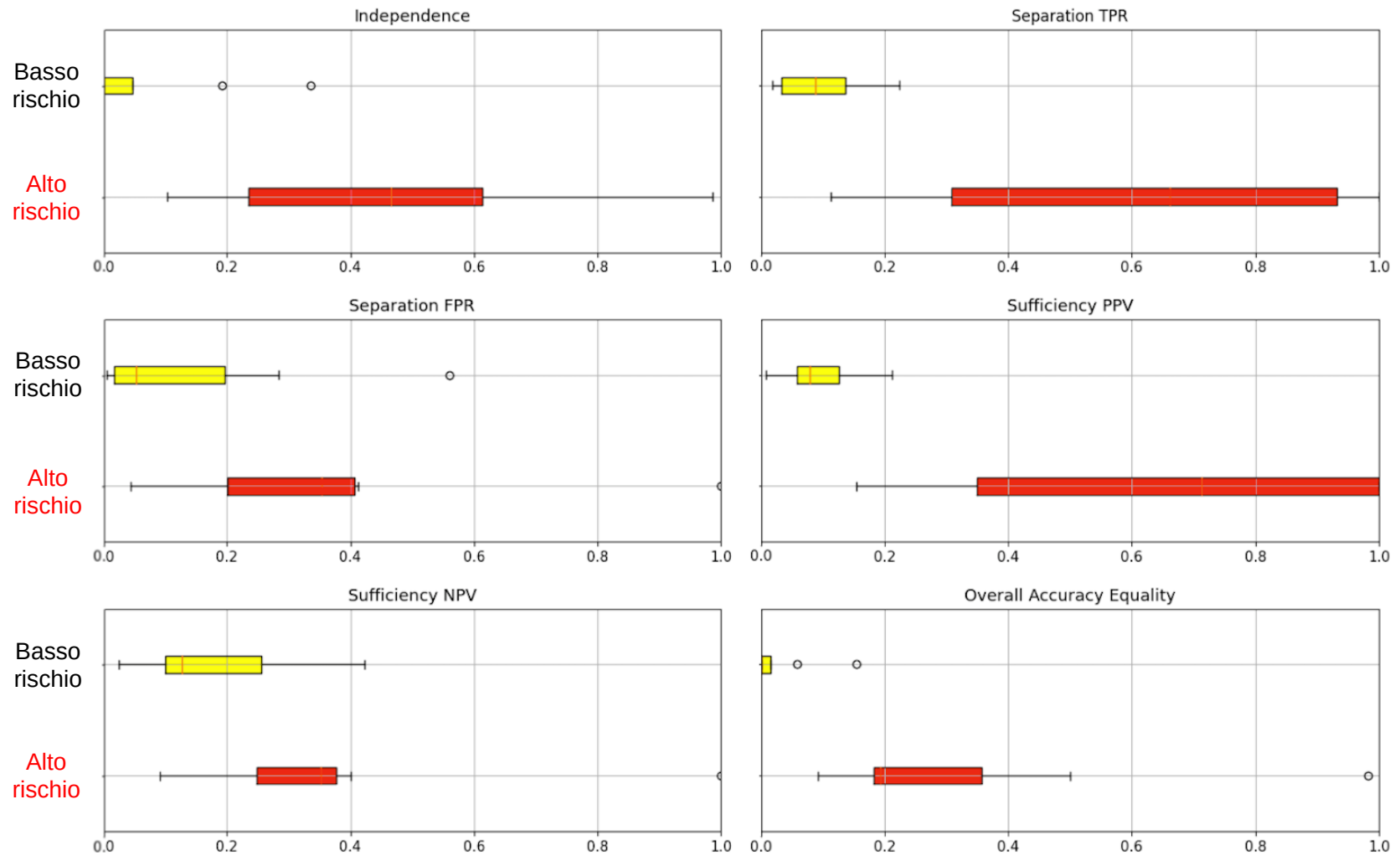
- Student Performance Data Set

<https://archive.ics.uci.edu/dataset/320/student+performance>

Etica nei sistemi AI



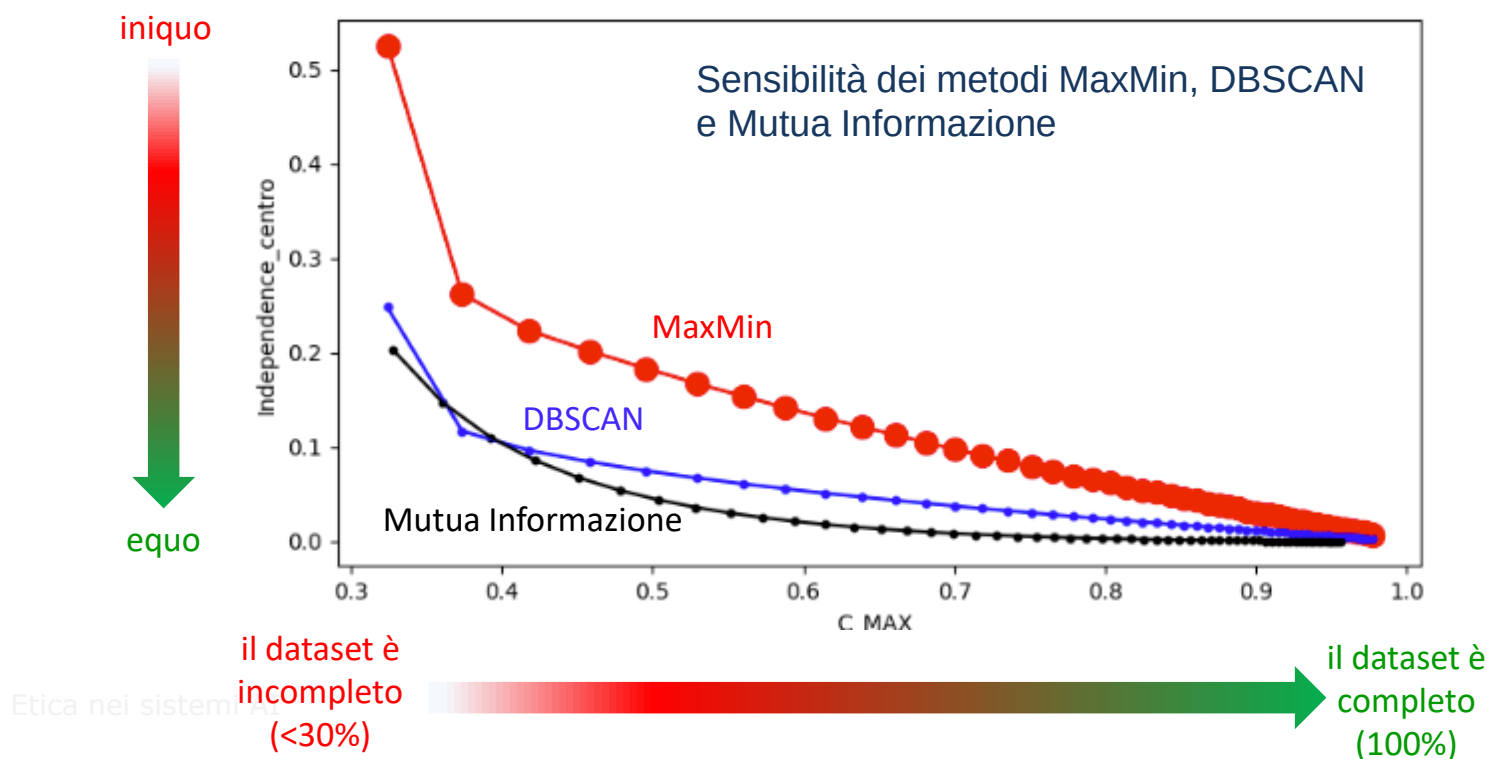
Relazione tra Cmax e misure di fairness calcolate con MaxMin



4 Annullare gli effetti nocivi del bias

Quando non è possibile arricchire il dataset con nuovi elementi si possono utilizzare le tecniche di arricchimento dei dati statistico.

Tra queste un esempio è il bootstrapping che utilizza un metodo di ricampionamento con sostituzione per generare nuovi set di dati da un campione originale.





Conclusioni

L'uso dei sistemi di IA nei processi decisionali presenta il rischio di perpetrare, o persino amplificare, i pregiudizi presenti nei dati.

La presenza di dati incompleti o sbilanciati può portare a risultati distorti.

Possiamo ridurre il rischio e rispondere all'AI ACT attraverso la misura di qualità presenti nelle norme internazionali come la nozione completezza dei dati.

Etica nei sistemi AI



Bibliografia

- 1) *"ISO/IEC standards and design of an Artificial Intelligence system"*, <https://ceur-ws.org/Vol-3916/> PP.39-43, Simonetta A., Paoletti M.C.
- 2) *"ISO/IEC quality standards for AI engineering"*, Oviedo J., Rodriguez M., Trenta A., Cannas D., Natale D., Piattini M., Computer Science Review, Volume 54, November 2024
- 3) *"The SQuaRE series as a guarantee of ethics in the results of AI systems"*, https://ceur-ws.org/Vol-3612, Simonetta A., Paoletti M.C., Nakaijma T.
- 4) *"Etica nei sistemi di Intelligenza Artificiale"*, <https://www.inail.it/cs/internet/comunicazione/pubblicazioni/catalogo-generale/pubbl-atti-sem-aggiorn-prof-ctss-csa-cit-salute-sicurezza.html>, Paoletti M.C., Simonetta A., Natale D.
- 5) *"Fairness Metrics and Maximum Completeness for the prediction of discrimination"*, <https://ceur-ws.org/Vol-3356/> PP.13-20, Simonetta A., Nakaijma T., Paoletti M.C., Venticinque A.
- 6) *"The use of Maximum Completeness to estimate bias in AI based recommendation systems"*, https://ceur-ws.org/Vol-3360, Simonetta A., Paoletti M.C., Venticinque A.
- 7) *"Using the SQuaRE series as a guarantee for GDPR compliance"*, <http://ceur-ws.org/Vol-3114>, Simonetta A., Paoletti M.C., Venticinque A.
- 8) *"Integrating SQuaRE data quality model with ISO 31000 risk management to measure and mitigate software bias"*, <http://ceur-ws.org/Vol-3114>, Simonetta A., Vetrò A., Paoletti M.C., Torchiano M.
- 9) *"Metrics for identifying bias in datasets"*, <https://ceur-ws.org/Vol-3118/>, Simonetta A., Trenta A., Paoletti M.C., Vetrò A.
- 10) *"A data quality approach to the identification of discrimination risk in automated decision making systems"*, Government Information Quarterly, Volume 38, Issue 4 October 2021, Vetrò A., Torchiano M., Mecati M.
- 11) *"Detecting Discrimination Risk in Automated Decision-Making Systems with Balance Measures on Input Data"*, Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021, Mecati M., Vetrò A., Torchiano M.,

Etica nei sistemi AI



ing. Alessandro Simonetta
alessandro.simonetta@gmail.com