POLITECNICO DI TORINO

Laurea Magistrale in Ingegneria Informatica

Tesi di Laurea

Identificazione e studio di misure di bias dei dati nei sistemi automatici di decisione

Relatore
Prof. Antonio Vetrò

Candidato
Ten. Cecilia Ruggiero
Matricola: 250155

Anno Accademico 2019 - 2020

Alla mia Principessa,
che mi ha insegnato a non mollare mai,
che mi ha asciugato le lacrime,
che mi ha insegnato il significato della parola amore;
sei il dono più bello che la vita potesse farmi.

Ai miei genitori,
fonte di ispirazione e supporto quotidiano.

Sommario

Ai giorni d'oggi è sempre più diffusa all'interno dei processi decisionali la presenza di sistemi di decisione automatica basati sui dati. Essi vengono utilizzati per prendere scelte anche in ambiti significativi come quello giudiziario o occupazionale. L'accrescere dell'utilizzo di questi sistemi pone una crescente preoccupazione per il loro potenziale impatto discriminatorio. In particolare, i sistemi di apprendimento automatico addestrati su dati sbilanciati, corrono il rischio di perpetuare tali stereotipi presenti all'interno della società.
Una delle sfide centrali è quella di determinare in un primo momento se i modelli utilizzati mostrano pregiudizi discriminatori ma in seconda istanza quello di cercare di evitare questo comportamento anomalo.
I problemi di equità e di discriminazione sorgono soprattutto a causa di serie di dati sproporzionati. Ciò è dovuto dal fatto che questi algoritmi cercano modelli comuni nei dati di input per adattarsi correttamente a nuovi dati mai visti in precedenza tratti dalla stessa distribuzione utilizzata per creare il modello.
I dati ti training quindi risultano essere uno degli anelli chiave per il giusto addestramento.
Se quindi insiemi di dati sproporzionati portano a risultati anomali l'obiettivo è quello di trovare delle metriche in grado di far porre l'attenzione sull'utilizzo di un training dataset equo, oppure di rilevare quando non lo è .
La ricerca di metriche già note in letteratura da applicare ai dataset si è sviluppata tra un indice di eterogeneità (indice di Gini), due indici di diversità (indice di Shannon e Simpson) e un indice di entropia (indice di Theil).

Come dataset applicativo per il caso d'uso la scelta è ricaduta su un dataset già oggetto di studi ovvero quello utilizzato nell'algoritmo Compas. I dati contengono variabili utilizzate per assegnare un punteggio di recidività agli imputati, insieme ai loro risultati entro 2 anni dalla decisione, per oltre 10.000 imputati penali nella Broward County, Florida.
Per riuscire a focalizzare l'attenzione su singoli elementi e vederne le variazioni di comportamento degli indici ho scelto di utilizzare altri due dataset creati ad hoc sulla falsa riga dell'originale variandone il numero di classi dell'attributo sensibile preso in esame e le sue frequenze così da limitare al minimo i cambiamenti e riuscire

ad individuare gli elementi che maggiormente influiscono gli indici utilizzati.
In questo ambito, riuscire ad agire in anticipo rispetto al manifestarsi di un problema, permette di evitare l'insorgere di comportamenti discriminatori che in quanto tali violano i diritti fondamentali dell'uomo.

Ringraziamenti

Nonostante mille peripezie sembra giunto al termine questo lungo percorso iniziato nell'agosto del 2014 tra le mura dell'Accademia Militare di Modena.
Ancora non ci credo e forse ci vorrà un po' di tempo per realizzare di avercela fatta a concludere questo percorso.
Questa esperienza mi ha visto andar via di casa quasi sei anni fa per intraprendere un lungo e faticoso percorso lontana dall'affetto dalla mia famiglia.
E' proprio a loro che vanno i primi ringraziamenti: ai miei genitori per aver creduto sempre in me, anche quando ero io la prima a pensare di non farcela, per avermi appoggiato nonostante tutto e per avermi sempre fatto sentire libera di tornare indietro.
Grazie a mia sorella Martina per aver accettato questa mia scelta e sopportato la distanza che ci ha separato: sei stata la mia forza durante tutto questo periodo. La mia assenza durante la tua crescita è l'unica cosa che realmente non mi perdono nonostante la soddisfazione di questo traguardo.
Grazie a tutta la mia famiglia, alle mie nonne, ai miei zii ed ai miei cugini per essermi sempre stati vicini e aver gioito con me per i piccoli traguardi intermedi.
...
Infine grazie anche un pò a me stessa per non aver mollato davanti alle difficoltà.
Ora è giunto il momento di iniziare un nuovo percorso!

Indice

Introduzione

Prima di affrontare l'argomento principale del lavoro di tesi è bene introdurre brevemente il contesto nel quale ci troviamo facendo chiarezza su concetti di base utili a comprendere a fondo il problema dal quale siamo partiti.
per questo motivo inizierò dando una definizione di intelligenza artificiale poiché sarà questo il contesto nel quale vogliamo applicare i dataset cercando di spiegare l'importanza di questa tecnologia e l'impatto che essa potrebbe avere ai giorni d'oggi.
Successivamente tratterò di bias per poi addentrarmi nel vivo del discorso analizzando delle metriche utilizzabili nei dataset al fine di misurarli e mitigarli.
Applicando successivamente queste metriche già note in letteratura ad un caso d'uso specifico analizzerò le peculiarità delle stesse soffermandomi sui possibili elementi che le caratterizzano e influiscono sui risultati.
Comprendere a fondo le metriche che si utilizzano permette di non arrivare a conclusioni totalmente o parzialmente errate solamente sulla base ad un indice che, se non contestualizzato e valutato nei giusti modi, poco ci dice sul nostro dataset.
L'importanza dell'individuazione e prevenzione di bias è fondamentale quando il sistema decisionale si colloca in contesti sociali delicati.
Il principale lo individuerei con l'articolo 2 della "Dichiarazione Universale dei Diritti Umani" approvata il 10 dicembre 1948 dall'Assemblea Generale delle Nazioni Unite la quale sancisce che "ad ogni individuo spettano tutti i diritti e tutte le libertà enunciate nella Dichiarazione, senza distinzione alcuna, per ragioni di razza, di colore, di sesso, di lingua, di religione, di opinione politica o di altro genere, di origine nazionale o sociale, di ricchezza, di nascita o di altra condizione" [1].

Capitolo 1

Intelligenza artificiale

L'intelligenza artificiale come disciplina accademica viene istituita nel 1956 con l'obiettivo di far sì che i computer svolgessero compiti considerati fino a quel momento solamente propri dell'essere umano.
In termini tecnici, l'Intelligenza Artificiale (in inglese Artificial Intelligence) è un ramo dell'informatica che permette la programmazione e progettazione di sistemi sia hardware che software che permettono di dotare le macchine di determinate caratteristiche che vengono considerate tipicamente umane quali, ad esempio, le percezioni visive, spazio-temporali e decisionali.
Si tratta cioè, non solo di intelligenza intesa come capacità di calcolo o di conoscenza di dati astratti, ma anche e soprattutto di tutte quelle differenti forme di intelligenza che sono riconosciute dalla teoria di Gardner, e che vanno dall'intelligenza spaziale a quella sociale, da quella cinestetica a quella introspettiva.
Un sistema intelligente, infatti, viene realizzato cercando di ricreare una o più di queste differenti forme di intelligenza che, anche se spesso definite come semplicemente umane, in realtà possono essere ricondotte a particolari comportamenti riproducibili da alcune macchine [2] .
Sfruttare l'intelligenza artificiale quindi significa far sì che un computer imiti in una certa maniera il comportamento umano con un l'intervento di quest'ultimo ridotto al minimo.
Intelligenza artificiale non è sinonimo di apprendimento automatico (più comunemente conosciuto col nome inglese Machine learning) anche se nel gergo comune vengono molte volte intercambiate erroneamente.
Il Machine learning è di fatto un sottoinsieme dell'intelligenza artificiale che consiste in tecniche che consentono ai computer di comprendere le cose dai dati e fornire applicazioni AI.
Il termine AI, infatti, non ci fornisce dettagli sul come questi problemi sono risolti in quanto esistono molte tecniche differenti per farlo. Il machine learning è una di queste tecniche che ha iniziato a diffondersi negli anni agli inizi degli anni '80 poiché alcuni problemi erano troppo difficili da risolvere per le prime tecniche usate

per l'intelligenza artificiale.
Con l'aumentare della difficoltà dei problemi da affrontare tramite tecnologie avanzate si arrivò al Deep learning che è a sua volta un sottoinsieme del ML capace, tramite reti neurali artificiali organizzate in diversi strati, , di risolvere problemi di complessità maggiore [3].
In figura 1.1. è mostrata graficamente la cronologia storica di queste tecniche.

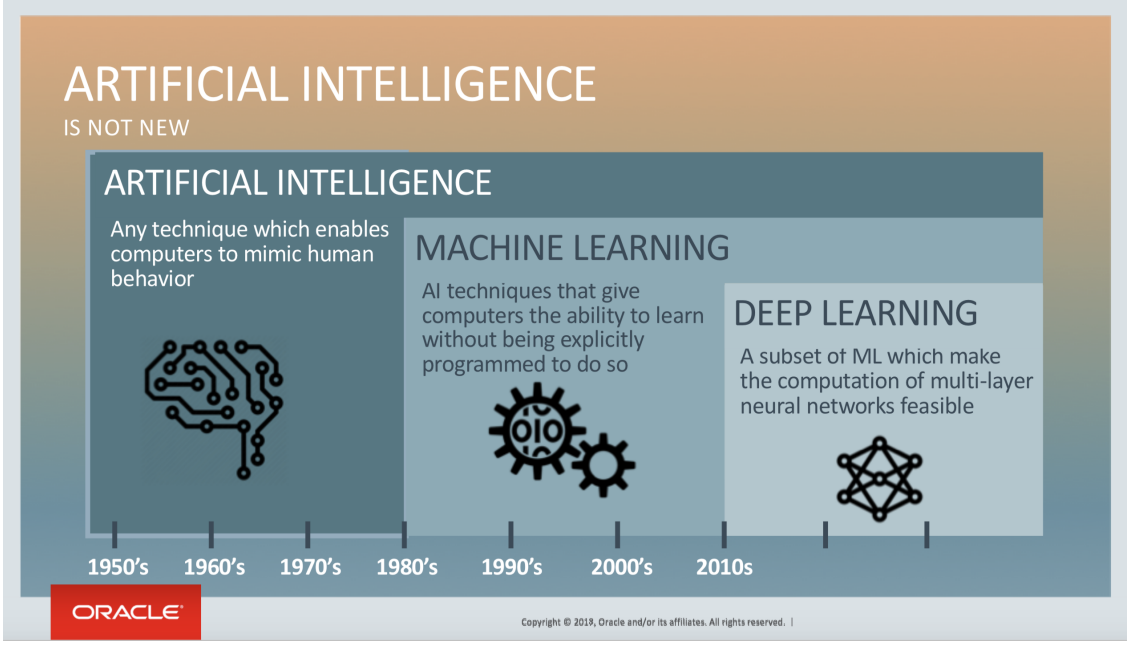

Figura 1.1. Relazione tra Intelligenza artificiale, Machine learning e Deep learning [3]

1.1 Machine learning

Le precedenti tecniche AI come gli algoritmi hard-coded (a codifica fissa ovvero con presente nel codice sorgente dei valori costanti che non possono essere cambiati senza ricompilazione) o i sistemi fissi basati su regole non hanno funzionato bene per determinate attività quali il riconoscimento di immagini o l'estrazione del significato dal testo. Per questo motivo si è ricercata la soluzione in un campo più profondo che non si limitasse ad imitare il comportamento umano ma nel cercare di simulare la maniera in cui gli esseri umani apprendono. L'esempio più semplice che si può fare per spiegare questo concetto è tramite una similitudine con l'essere umano durante le prime fasi di vita che si approccia ad imparare a leggere. Il bambino non aspetterà di imparare tutta la grammatica ma inizierà a leggere i libri più semplici per poi col tempo imparare e riuscire ad affrontare quelli più complicati. In tal modo il bambino imparerà le regole e le eccezioni dell'ortografia e della

grammatica semplicemente leggendo. In altre parole, è possibile imparare dall'elaborazione di dati. Ed è esattamente questa l'idea che sta dietro l'apprendimento automatico che studia tecniche e approcci differenti per insegnare ad una macchina l'abilità di apprendere in modo autonomo, migliorare le proprie performance o la propria conoscenza tramite l'esperienza. L'idea è di fornire una moltitudine di dati a un algoritmo (analogamente a quello che si fa con un cervello del bambino che legge i primi libri) e lasciare che esso comprenda dai dati. Ad esempio fornendo a un algoritmo molti dati sulle transazioni finanziarie e dicendogli quali sono quelle fraudolenti. Dopodiché, lasciare che esso elabori ciò che contraddistingue una frode, in modo che lo possa prevedere in futuro. Man mano che questi algoritmi si sviluppavano, diventavano in grado affrontare molti problemi di difficoltà differenti.

Il Machine Learning può essere utilizzato per ottemperare a svariati compiti tra cui la classificazione di esempi, la regressione, il calcolo delle probabilità fino al reperimento di nuova conoscenza.[4]
Esso è la combinazione dei seguenti elementi:

1. **Task:** la descrizione degli obbiettivi, ovvero quali problemi si intende risolvere (classificazione, regressione, clustering, etc...);
2. **Model:** la rappresentazione matematica delle soluzioni ai problemi sopra citati (classificatori lineari, alberi decisionali, etc...);
3. **Feature:** la rappresentazione degli oggetti coinvolti nei modelli(numeri, categorie, costruzione/selezione di feature, etc...).

1.2 I Task

Il machine learning può essere utilizzato in svariati contesti applicativi e permette di risolvere diverse classi di problemi. Nonostante la varietà di problematiche da risolvere, è possibile individuare delle macro-categorie e costruire una tassonomia dei possibili task suddividendoli in base a due criteri:

- Obbiettivi del modello
- Modalità di apprendimento

Il primo criterio di suddivisione, individua due possibili categorie:

1. **Task prendittivi:** il loro scopo è quello di prevedere una certa variabile (ad esempio una categoria di appartenenza, una valore numerico, un cluster, etc...);
2. **Task descrittivi:** il loro scopo è quello di identificare una struttura sottostante ai dati.

Il secondo criterio, invece, consiste del suddivide i task in due tipologie principali ovvero:

1. **Apprendimento supervisionato:** noto anche come supervised learning nel quale l'apprendimento viene diretto da esempi correlati dalla loro soluzione (ad esempio per i task di classificazione di immagini apprendono partendo da una serie di immagini alle quali sono già associate le etichette corrette);

2. **Apprendimento non supervisionato:** noto anche come unsupervised learning nel quale l'apprendimento viene diretto da esempi privi di ulteriori informazioni (ad esempio nelle Self-Organizing Map (SOM) le immagini di esempio sono prive di informazioni sul loro contenuto).

La tabella rappresentata la tassonomia generata e fornisce al suo interno esempi di task che ricadono in quella categoria.

	Modello predittivo	Modello descrittivo
Apprendimento supervisionato	classificazione, regressione	Subgroup discovery (tecnica di data mining che scopre associazioni interessanti tra diverse variabili rispetto a una proprietà di interesse)
Apprendimento NON supervisionato	clustering predittivo	clustering descrittivo

Figura 1.2. Tassonomia dei task.

1.2.1 I Task predittivi

I task predittivi sono tutti quei task il cui obbiettivo è quello di eseguire delle inferenze corrette su dati non ancora osservati, in base all'attuale conoscenza del dominio [4].
Tale modello viene ottenuto dalla fase di apprendimento su un insieme di esempi chiamato training set e deve essere in grado di generalizzare su dati sconosciuti, che non appartengono al suddetto insieme. Alcuni dei principali esempi di questa categoria di task vengono riportati di seguito:

1. **Classificazione binaria e multiclasse:** lo scopo è classificare i dati in input in due (binary classification) o più (multiclass classification) categorie;

2. **Regressione:** l'associazione non viene fatta con un valore appartenente ad un insieme finito, bensì con un valore reale risultante da una funzione (il modello) in grado di rappresentare i punti nello spazio;

3. **Clustering predittivo:** l'obbiettivo è determinare il cluster di appartenenza di una determinata istanza basandosi su un criterio di vicinanza.

Il principale problema che affligge in particolare i task predittivi nel machine learning è l'overfitting (letteralmente sovradattamento) ovvero l'eccessivo adattamento. In termini pratici si traduce in un modello estremamente performante sul training set, il quale però si rivela inefficiente sul test set, o più in generale su qualsiasi istanza al di fuori di quelle usate in fase di addestramento.
Questo si rivela particolarmente problematico: un modello che soffre di questo problema si rivela inutile perché è inutilizzabile per fare inferenza corretta su dati che non appartengono al training set, cosa per il quale invece sarebbe destinato.
Proprio per valutare le capacità del modello di generalizzare correttamente si usano due insiemi diversi di istanze:

1. **training set:** l'insieme delle istanze utilizzate per la fase di addestramento;

2. **test set:** un insieme di istanze diverse da quelle contenute nel training set utilizzate per la verifica delle prestazioni e la capacità del modello di generalizzare adeguatamente.

Nella valutazione delle performance quindi non si deve puntare solamente a massimizzare quelle sul training set, si rischierebbe di cadere nella trappola dell'overfitting, bensì si deve cercare di minimizzare la distanza fra le performance del training set e le performance del test set. Questo vale a dire che il modello è stato in grado di generalizzare molto bene, riuscendo a classificare efficacemente anche tutte le istanze nuove, mai viste durante l'addestramento. Al contrario quando le performance sul test set si discostano maggiormente si ha un modello eccessivamente adattato al training set.
Nella Figura 1.3 si possono identificare le diverse parti che compongono il processo di apprendimento automatico e come sono fra loro relazionate per i task di tipo predittivo.

1.2.2 I Task descrittivi

L'obbiettivo dei task descrittivi è quello di esplicitare una struttura sottostante ai dati, ad esempio individuando legami funzionali fra gli attributi delle istanze, distribuzione delle istanze in alcune classi o notevoli generalizzazioni e astrazioni del campione di dati.

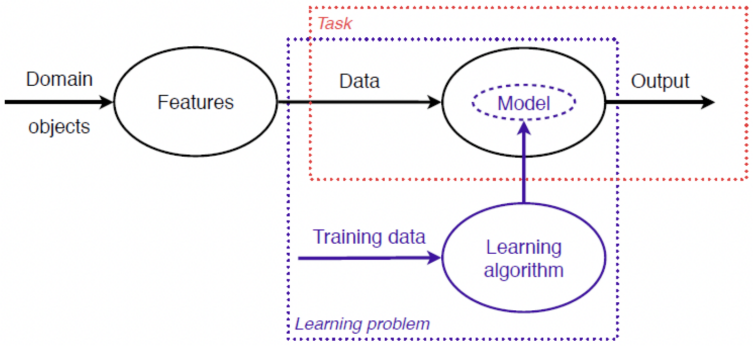

Figura 1.3. Schema riassuntivo del funzionamento del machine learning per task di tipo predittivo.

1. **subgroup discovery:** l'obbiettivo è la ricerca di un sottoinsieme di istanze nel dataset con caratteristiche peculiari, come ad esempio una particolare distribuzione delle classi rispetto all'insieme di partenza;

2. **descriptive clustering:** l'obbiettivo è la ricerca dei possibili cluster con i quali si possono classificare tutte le istanze nel dataset;

3. **association rule discovery:** l'obbiettivo è la ricerca di regole ricorrenti in diversi attributi delle istanze nel dataset.

Si noti che il clustering può quindi essere declinato in due versioni differenti: nei task predittivi conosciamo i cluster e si devono classificare al loro interno nuove istanze, nei task descrittivi dalle istanze del training set si devono scoprire i cluster in cui dividerle.
Semplificando, diciamo che i task predittivi generano modelli che dovrebbero valere anche su istanze non conosciute (test set) mentre i task descrittivi si occupano soltanto di ottenere le informazioni su quel dataset passato in input e non si occupano di generalizzare su istanze al di fuori di esso.

1.3 I Modelli

Dopo aver parlato dei diversi tipi di task che il machine learning può affrontare, si deve dare spazio ad un aspetto ugualmente importante: i modelli, ovvero il processo in cui nella fase di apprendimento vengono mappati i dati in input descritti sotto forma di feature agli output opportuni [4].
I modelli si possono classificare essenzialmente in tre categorie principali:

1. **modelli geometrici:** i dati vengono mappati in uno spazio geometrico e la soluzione è descrivibile anch'essa geometricamente (ad esempio una retta che divide un piano);

2. **modelli probabilistici:** il dominio viene descritto attraverso elementi probabilistici, associando ai dati una probabilità, la quale permette di fare inferenza probabilistica o calcolare distribuzioni probabilistiche per ridurre l'incertezza;

3. **modelli logici:** sono costruiti con formule logiche grazie alle quali è possibile fare inferenza.

Al fine di spiegare in maniera più chiara questo scenario presento un esempio di un classico dominio su cui applicare queste tecniche.

1.4 Assassin-Spam Filter

Prendiamo in considerazione uno dei più famosi esempi nell'ambito del machine learning, il filtro anti-spam. Supponiamo di voler costruire un filtro il quale determini in maniera automatica se una certa mail in arrivo sia ham o spam. Questo compito può essere risolto con svariate tecniche.
La più semplice, ma anche la meno efficiente, risulta quella dell'utilizzo di un semplice pattern matching, ricercando delle sequenze giudicate tipiche di una mail spam in modo tale che, nel caso siano riscontrate nelle mail ricevute, quest'ultime possono essere considerate spam.
Questo tipo di approccio soffre di due problemi fondamentali:

* risulta indipendente dal contesto

* scarsa manutenibilità

Per un algoritmo simile, risulta forzata l'inclusione nelle tecniche di machine learning, poiché la macchina applica uno schema di classificazione che non si modifica nel tempo e non sfrutta esperienze pregresse per migliore le proprie performance.
Una tecnica che permette un grado di flessibilità e adattabilità maggiore è implementare invece un modello geometrico con un classificatore lineare.
Un classificatore lineare valuta il risultato di una combinazione lineare di valori, per determinare la classificazione più corretta ad un determinato input.
L'idea è che ogni input può essere descritto prendendo in considerazione alcune sue proprietà, o attributi a cui sarà assegnato un certo valore.
Da un punto di vista geometrico, la classificazione lineare può essere rappresentata come una partizione dello spazio dei dati.
Tale concetto è espresso graficamente in figura 1.4, nella quale si può identificare

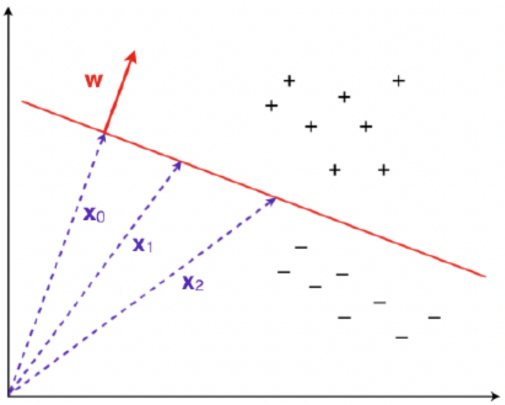

Figura 1.4. Un esempio grafico di classificatore lineare

il vettore dei pesi w evidenziato in rosso, perpendicolare alla soglia che divide gli esempi altrimenti nota come decision boundary o linear boundary.

Quanto espresso a parole e graficamente, può essere formalizzato come segue:

* $x_i = \{0, 1\}$ risultato binario dell'i-esimo test;

* $w = \{w_1, w_2, ..., w_n\}$ il vettore dei pesi;

Il modello lineare tramite il quale si classifica un input con una soglia di decisione t tramite il prodotto scalare w $\cdot$ x risulta essere:

$$\sum_{i=0}^{n} w_i \cdot x_i > t$$

Per fare un esempio numerico supponiamo che ad ogni e-mail siano valutati gli attributi x_1,x_2. Se i pesi w_1, w_2 associati sono w_1, $w_2 = 4$ e la soglia t = 5 allora il classificatore classificherà le istanze del dataset in esempio in questo modo:

Dalla tabella si può notare come algoritmo giudica spam se e solo se entrambi i test x_1 e x_2 hanno dato esito positivo.
Riassumendo si possono identificare i tre elementi introdotti in precedenza nel paragrafo 1.1:

1. **Task:** la classificazione delle e-mail in entrata;
2. **Model:** la funzione definita per calcolare il punteggio di ogni e-mail da valutare (nel nostro caso specifico abbiamo preso in considerazione un modello lineare);

Email	X1	X2	Smap?	4x1 + 4x2
1	1	1	1	8
2	0	0	0	0
3	1	0	0	4
4	0	1	0	4

3. **Feature:** i risultati dei due test x_1 e x_2,

Il processo di classificazione lineare può essere sintetizzato come nello schema in figura dove l'algoritmo apprende dei pesi tramite un training set.

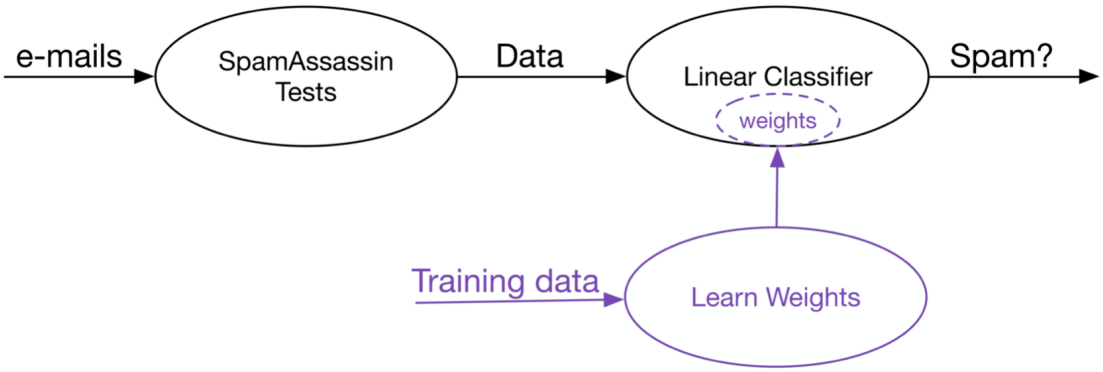

Figura 1.5. Schema generale di apprendimento per un problema di classificazione lineare

1.5 L'importanza dei training set

L'insieme dei dati forniti per eseguire l'addestramento viene detto training set; un training set è formato da coppie x, f(x) dove x è una possibile istanza descritta tramite i suoi attributi mentre f() rappresenta la reale funzione che assegna i giusti valori e che l'algoritmo tenta di approssimare.
Nel caso di Assassin Spam analizzato precedentemente devono essere forniti all'algoritmo dei buon esempi di email spam in modo che la macchina comprenda come etichettare le istanze quando dovrà lavorare su input mai osservati.
Risulta necessario cercare di individuare quali caratteristiche deve avere un training

set e gli esempi per poter essere etichettato come “buoni”.
Di fatto quando parliamo di buon training set intendiamo che esso debba essere in grado di generalizzare lo scenario in modo tale da fornire alla macchina gli strumenti per lavorare correttamente su dati nuovi che non sono mai stati osservati prima.
Affinché questo avvenga il training set deve avere la caratteristica principale di essere sufficientemente rappresentativo del problema in modo che l'algoritmo faccia le giuste inferenze.
Da ciò si può dedurre che uno dei problemi principali sorge quando il training set non risulta sufficientemente rappresentativo. Infatti, essendo che l'algoritmo utilizza il training set per costruire un proprio modello di comprensione del dominio, allora le sue scelte future saranno influenzate da ciò che l'algoritmo ha appreso dal training set.
Questo può portare all' apprendimento di un modello del problema parzialmente errato o anche totalmente errato.
Quando accade che l'algoritmo apprende un modello errato, dovuto al sovradattamento di quest'ultimo ad un insieme di esempi iniziali, allora si parla di overfitting.
L'overfitting è un problema importante, ma non è il solo a mettere a repentaglio le prestazioni e a decretare il successo o il fallimento di un training set.
Un altro problema si può presentare quando il training set non fornisce una descrizione esaustiva degli esempi e non sono presenti attributi essenziali per la comprensione del dominio (come ad esempio la mancanza di alcuni attributi in un gran numero di record del database o l'utilizzo di una scala numerica non descritta nel modo corretto).
Correlato a questo l'algoritmo deve anche essere in grado di comprendere quali sono dei buoni test, selezionando i giusti attributi da considerare per portare a termine il suo compito, ovvero deve sapere come interpretare i dati in modo da comportarsi nella maniera voluta.
Non è detto che il dominio su cui lavorare sia conosciuto a sufficienza da poter conoscere sia i buoni test sia i buoni esempi da fornire ma questo non implica necessariamente l'impossibilità di intraprendere una qualche comprensione dei dati.
Infatti molte applicazioni nel campo del machine learning sono rivolte alla costruzione di un modello di interpretazione dei dati, cercando di cogliere gli aspetti essenziali di un certo dominio.
Sulla base di quanto appena discusso circa l'overfitting e la semplificazione dei modelli, si potrebbe essere indotti a credere che questo principio debba guidare sempre la risoluzione dei problemi di apprendimento verso modelli tendenzialmente lineari, i più semplici che ci sono.
Questa è una conclusione inesatta, ed è il nocciolo del bias-variance dilemma.
Non basta infatti puntare al modello più semplice, perché questo potrebbe rivelarsi insufficiente per minimizzare le penalità ed ottenere un modello in grado di svolgere con accuratezza le predizioni.

Alla luce di quanto appena detto si può quindi analizzare l'errore che un modello commette e distinguerne due cause. I modelli troppo semplici sono maggiormente soggetti ad errori dovuti alla mancanza di complessità che impedisce loro di stimare con sufficiente precisione; i modelli troppo complessi che invece commettono errori perché si adattano troppo ai dati e quindi vengono penalizzati non appena le istanze si discostano anche poco.

1.6 Bias-Variance dilemma

In statistica e in ambito machine learning il dilemma bias-varianza è la proprietà di un insieme di modelli predittivi in cui i modelli con un minor bias nella stima dei parametri hanno una maggiore varianza delle stime dei parametri tra i campioni e viceversa.
Quando parliamo di errori in un sistema di previsione possono insorgere tre tre tipi di errore differenti : i bias, la varianza ed l'errore irriducibile.
Col termine bias si possono intendere concetti differenti in base al contesto in cui viene utilizzato.
In questa circostanza si intende la differenza tra la previsione attesa (o media) del modello e il valore reale che si cerca di prevedere. In altre parole il bias è all'accuratezza del modello (accuracy), che può essere influenzata dalle assunzioni errate. Naturalmente, se c'è un solo modello, parlare di valori di previsione attesi o medi potrebbe risultare errato.

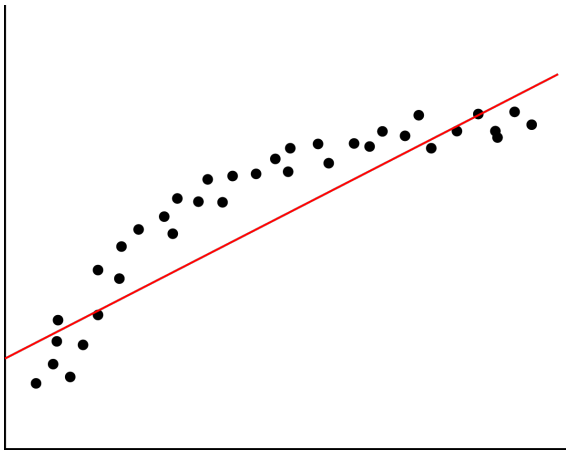

Figura 1.6. Illustrazione grafica del fenomeno dell'underfitting.

Tuttavia possibile ripetere il processo di costruzione del modello più di una volta ed ogni volta raccogliere i dati eseguendo una nuova analisi e creando un nuovo modello. Per la casualità dei set di dati che lo compongono, i modelli ottenuti

avranno una serie di previsioni.
La distorsione (il bias) misura quanto sono distanti, in generale, le previsioni di questi modelli dal valore corretto.
Immaginando di applicare una regressione lineare ad un set di dati che hanno un modello non lineare: indipendentemente da quante altre osservazioni vengano raccolte, una regressione lineare non sarà in grado di modellare le curve in quei dati.
Con un high bias (scarsa accuratezza) si sarà un'alta tendenza al sottoadattamento, perché l'attenzione posta ai dati di training è minima.
Questo fenomeno è noto come underfitting ed è rappresenatto in figura 1.6.

In statistica, invece, quando si parla di varianza si fa riferimento alla sensibilità (sensitivity) del modello alle piccole variazioni nel dataset di training.
Anche in questo caso si può immaginare la possibilità di ripetere più volte l'intero processo di costruzione del modello. La varianza avviene quanto variano le previsioni per un dato punto tra le diverse realizzazioni del modello. Per esempio, esiste un algoritmo che si adatta ad un modello completamente libero e flessibile al set di dati.
In pratica il modello non vincolato memorizza il set di addestramento, incluso tutto il rumore. Questo fenomeno è noto come overfitting ed è rappresentato in figura 1.7.

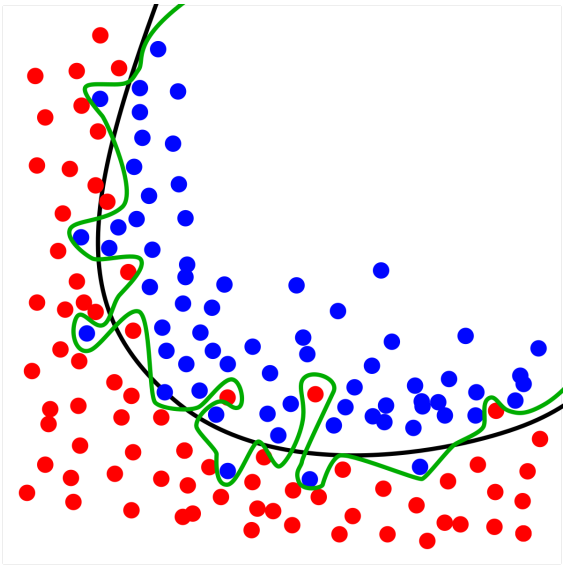

Figura 1.7. Illustrazione grafica del fenomeno dell'overfitting.

L'errore irriducibile è la terza tipologia di errore che concorrere nella determinazione dell'errore di generalizzazione di un modello. Esso è il più difficilmente

riducibile (come sufferisce il nome stesso) in quanto legato al rumore dei dati; infatti deriva tipicamente dalla casualità intrinseca o da un insieme incompleto di caratteristiche.

L'unico modo per attenuarne l'effetto, riducendo questa parte dell'errore, è operare:

* rimuovendo le anomalie

* controllando le sorgenti dei dati (es: sensori mal funzionanti)

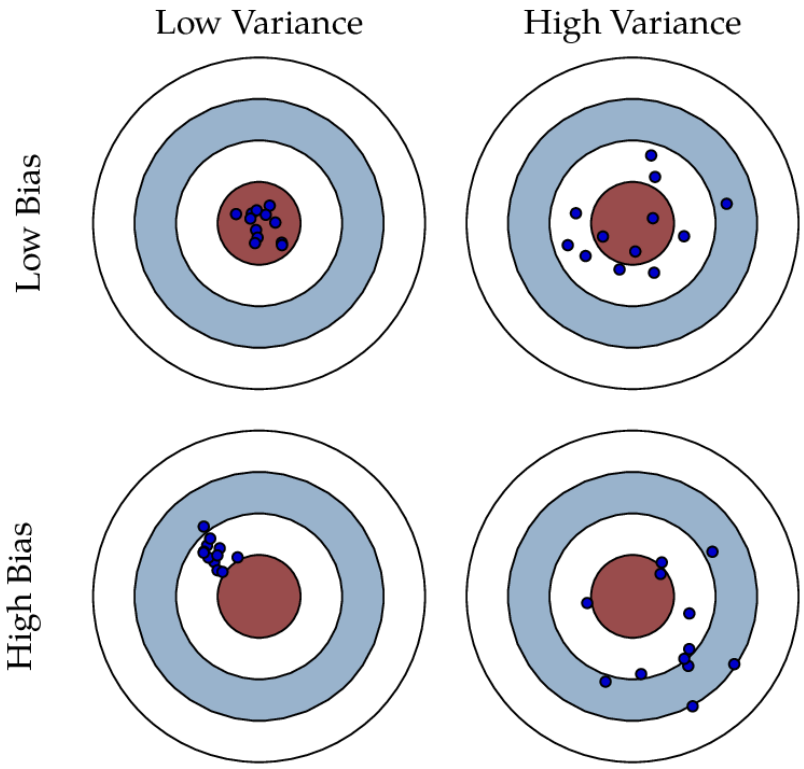

Figura 1.8. Illustrazione grafica del compromesso bias-varianza [6].

Si può quindi concludere che modelli molto complessi soffrono della varianza (variance) dei dati nel dataset ma sono molto precisi (fino al raggiungimento dell'overfitting); al contrario modelli molto semplici non patiscono la varianza, ma al contrario introducono un bias che non può essere nemmeno risolto da grandi quantità di dati in input. Volendolo tradurre in formula matematica [6]:

$$\frac{\Delta bias}{\Delta complessità} = -\frac{\Delta varianza}{\Delta complessità}$$

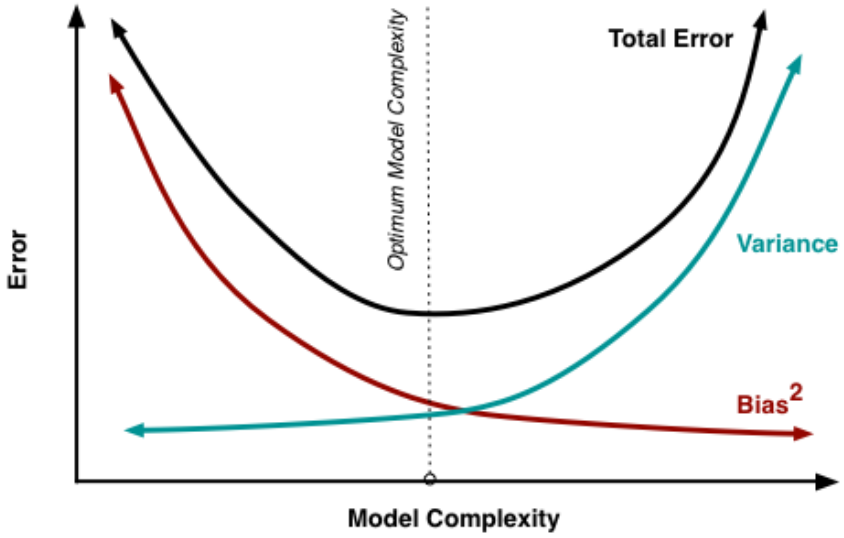

Figura 1.9. Relazione tra bias e varianza [5].

L'errore totale risulterà quindi essere:

$$ErroreTotale = Bias^2 + Varianza + ErroreIrriducibile$$

Capitolo 2

Metriche

Ai giorni d'oggi i sistemi di decisione automatica basati sui dati svolgono un ruolo importante nel processo decisionale in svariati settori e vengono utilizzati per prendere decisioni anche in ambiti significativi come quello giudiziario.
L'accrescere dell'utilizzo di questi sistemi pone una crescente preoccupazione per il loro potenziale impatto discriminatorio. In particolare, i sistemi di apprendimento automatico addestrati su dati di parte, hanno il rischio di imparare in maniera scorretta e di conseguenza perpetuare tali pregiudizi.
Per questo motivo una delle sfide centrali è quella di determinare se i loro modelli mostrano pregiudizi discriminatori.
Come abbiamo affrontato nel primo capitolo, i dati svolgono un ruolo cardine in quanto potrebbero pregiudicare il corretto esito di un sistema decisionale istruendo in maniera errata il modello. Una volta compreso che l'origine dei bias potrebbero non essere intrinsecamente solo nel modello possiamo comprendere l'origine del problema: la selezione di una metrica di equità appropriata per un determinato dataset risulta identificarsi con il paradosso dell'uovo e della gallina perché i tipi di distorsioni presenti nei dataset ci indirizzando nella determinazione della metrica (metodo di misurazione di caratteristiche) maggiormente appropriata ma per poter determinare quali tipi di distorsioni sono presenti è necessario un modo per misurare le stesse.
È noto che i problemi di equità e di discriminazione sorgono soprattutto a causa di serie di dati sproporzionati. Ciò è dovuto dal fatto che questi algoritmi cercano modelli comuni nei dati di input per adattarsi correttamente a nuovi dati non visti in precedenza, tratti dalla stessa distribuzione utilizzata per creare il modello. Insiemi di dati sproporzionati portano a risultati sproporzionati.
Il semplice campionamento casuale (che è il metodo più utilizzato nelle indagini statistiche) richiede che la probabilità di estrazione del campione sia nota e non pari a zero, e che non solo ogni elemento ma anche ogni combinazione di elementi (di pari numero) abbia la stessa probabilità di essere estratto. Un campione parziale porta a stime parziali.

Per questo motivo, il campionamento statistico è un passo fondamentale. Molti dei set di dati utilizzati tutt'oggi non sono stati generati utilizzando il campionamento probabilistico, ma sono piuttosto selezionati attraverso metodi non probabilistici.
La natura dei dati può essere varia ma nella maggior parte dei casi essi provengono da uno storico di ciò si vuole analizzare. Da questo riusciamo a capire che gli stessi dati possono contenere stereotipi di cui la nostra società è affetta.
Potrebbe essere utile individuare una misura quantitativa che rifletta quanti tipi diversi (come i gruppi protetti) sono rappresentati.
Per questo motivo la mia analisi cerca di individuare tra delle metriche note in letteratura quelle che potrebbero darci delle indicazioni di sproporzione dei dataset analizzandone la struttura.
La scelta è stata di comprendere e applicare questi indici con particolare attenzione agli attribuiti sensibili che potenzialmente potrebbero generare decisioni impari.

2.1 Classificazione dei caratteri

In statistica le informazioni rilevanti che si individuano sono dette caratteri.
La maniera in cui in carattere si manifesta è detto modalità.
E' fondamentale che le modalità individuate rappresentino tutti i possibili stati che il carattere assume e che ad ogni unità si possa associare una ed una sola modalità.
In base alle loro caratteristiche possiamo individuare diversai tipi di caratteri:

1. **Caratteri qualitativi:** Esprimono una qualità, ovvero le modalità sono dei valori non numerici (ad esempio: il genere o il credo religioso).
   In base ai valori i caratteri qualitativi si distinguono in:

   (a) **Sconnessi (o nominali):** hanno per modalità caratteristiche qualitative tra le quali non esiste (e non è possibile stabilire oggettivamente) un ordinamento (sesso, religione, caratteristiche fisiche, nazionalità ecc). Per comprendere meglio in concetto si può far l'esempio della religione: non è possibile affermare che la religione cristiana abbia una precedenza (di alcun tipo) su quella musulmana. infatti l'unico confronto tra le due modalità e l'uguaglianza o la diversità. In altri termini si dice che si tratta di carattere nominare se per le sue modalità è possibile affermare solo se sono uguali o diverse.

   (b) **Ordinati:** hanno per modalità caratteristiche qualitative tra le quali esiste un ordinamento naturale (anno di nascita, anno di conseguimento del titolo di laurea etc). Questo tipo di caratteri costituisce una scala ordinale in quanto è possibile dare un ordine alle modalità in modo da affermare quale modalità precede l'altra. Tra i caratteri qualitativi ordinati va individuato un sottogruppo di caratteri (detti caratteri ordinati

ciclici) per i quali è necessario individuare un elemento primo ed un ultimo per poter essere confrontati. questo è il caso per esempio dei giorni della settimana che devono ricorrere ad una convenzione per identificare quale viene prima dell'altro.

2. **Caratteri quantitativi:** Esprimono una qualità, ovvero le modalità sono dei valori numerici che esprimono una misura. Quando si opera con caratteri quantitativi, date due modalità è possibile non solo stabilire quale sia maggiore rispetto all'altro ma è possibile effettuare operazioni matematiche quali il rapporto o la sottrazione confrontandone le quantità.
   Nel caso in cui sia possibile calcolare solo la differenza si parla di scala a intervalli; quando invece è possibile calcolarne anche il rapporto tra modalità si parla di scala di rapporti.
   In base ai valori i caratteri quantitativi si distinguono in:

   (a) **Discreti:** le modalità rappresentano dei conteggi perché numerano la quantità di un dato insieme.
   Sono rappresentati con numeri interi (numero di presenti, numero di allievi iscritti etc).

   (b) **Continui:**le modalità sono rappresentate (in linea teorica) da tutti i numeri reali compresi in un determinato intervallo (una qualsiasi misura di peso, tempo etc).

   I caratteri quantitativi possono inoltre essere distinti tra "trasferibili" e "non trasferibili" in base alla possibilità di poter o meno cedere parte del proprio carattere ad un'altra unità. Per esempio il reddito risulta essere un carattere trasferibile diversamente dal peso che è non trasferibile.

2.2 Indice di eterogeneità

Nel caso di variabili qualitative nominali la varianza e gli altri indici derivati non possono essere calcolati in quanto, come detto nel paragrafo precedente, non sono calcolabili la media n'è la mediana n'è altri valori numerici di riferimento dai quali calcolare le distanze.
Nonostante questa limitazione potrebbe comunque risultare necessario calcolare un indice che misuri la dispersione della distribuzione delle frequenze ovvero l'eterogeneità.

2.2.1 Indice di Gini

L'indice di eterogeneità di Gini è un indice di misurazione applicabile alle variabili qualitative nominali ed utilizzato in diversi ambiti: da quello economico a quello psicologico passando per quello ecologico ed in ogni ambito assume una denominazione differente.

Definiamo:

* s le categorie (le variabili qualitative);

* $\{x_1, x_2, ..., x_n\}$ come il campione preso in esame in cui occorrono i distinti valori $v_1, v_2, ..., v_s$;

* f_i come la frequenza relativa dell'elemento v_i per $i = 1, .., s$;

La quantità I definita come segue è detta indice di eterogeneità di Gini assoluto.

$$I = 1 - \sum_{i=1}^{s} (f_i)^2$$

Si noti che il valore di I è compreso ta 0 ed $\frac{S-1}{S}$ dove un indice maggiore rappresenta la massima eterogeneità e viceversa un numero più vicino allo 0 evidenzia una maggiore concentrazione di frequenze in poche categorie, quindi minore eterogeneità.

In caso di eterogeneità minima (o massima omogeneità), tutti gli elementi del campione assumono lo stesso valore, dunque esiste un solo j per cui $f_j = 1$ e per ogni $i \neq j$ si ha $f_i = 0$, pertanto I = 1 - 1 = 0.

In caso di eterogeneità massima tutte le osservazioni hanno invece la medesima frequenza $f_i = \frac{1}{s}$ e quindi $I = 1 - \frac{1}{s} = \frac{(s-1)}{s}$.

Nella nostra applicazione successiva ci risulterà più utile una misura relativa del grado di eterogeneità. Per questo motivo definiamo come I_N l'indice di Gini normalizzato come segue:

$$I_N = \frac{I}{\frac{s-1}{s}}$$

Così facendo I_N varierà tra 0 e 1, dove 1 significa che tutte le classi hanno la stessa frequenza (potremmo ipotizzarlo come condizione di equità), e 0 significa che c'è solo una o poche classi con frequenza significativamente più alta delle altre (enorme disparità).

Per comprendere meglio l'applicazione si può fare un esempio numerico.
Si consideri la distribuzione dei titoli di studio conseguiti nella scuola secondaria di secondo grado dagli studenti che desiderano immatricolarsi presso l'università riportati in tabella 2.1.

Tipologia di diploma	Frequenza assoluta
Istituto commerciale	35
Liceo classico	25
Liceo linguistico	6
Liceo scientifico	102
Istituto tecnico	30
Altro	2
TOTALI	**200 studenti**

Figura 2.1. Distribuzione dei titoli di studio

Il numero di osservazioni nel campione si ottiene sommando i valori nella seconda colonna della tabella e ottenendo 200. Pertanto l'indice di Gini per i dati riportati è pari a:

$$I = 1 - \frac{1}{200^2}(35^2 + 25^2 + 6^2 + 102^2 + 30^2 + 2^2) = \approx 0.67$$

Il numero di diverse osservazioni nel campione è uguale al numero di righe della tabella, e cioè 6, così che l'indice di Gini normalizzato assume il valore:

$$I_N = \frac{6}{5} I \approx 0.80$$

Avendo parlato dell'indice di eterogeneità Gini non si può non menzionare l'indice di concentrazione di Gini che però, essendo applicabile a variabili quantitative trasferibili, nel nostro caso di attributi sensibili principalmente qualitativi nominali non trova applicazione.

2.3 Entropia

La teoria dell'informazione è un sotto campo della matematica che si occupa di quantificare le informazioni per la comunicazione.
Più precisamente esso si occupa di compressione dei dati e i suoi limiti quando si parla di elaborazione dei segnali.
La quantificazione della quantità di informazioni richiede l'uso di probabilità; per questo motivo vi è la relazione tra la teoria dell'informazione e la probabilità come supporto per l'apprendimento automatico.
Alla base vi è l'idea di misurare quanta sorpresa c'è in un evento. Gli eventi rari in quanto sorprendenti hanno una bassa probabilità di accadere e quindi contengono più informazioni degli eventi non rari ovvero che hanno un'alta probabilità di avvenire[9] .
Il calcolo delle informazioni per un evento viene espresso come h() e calcolato come segue:

$$H(X) = -\log(p(x))$$

Il segno negativo assicura che il risultato sia sempre positivo o nullo.

L'informazione sarà zero quando la probabilità di un evento è 1,0 o una certezza, ad esempio non c'è sorpresa.

Per fare un esempio consideriamo il lancio di una singola moneta.
La probabilità che esca testa è del 50% (p=0.5), esattamente la stessa che possa uscire la croce. Possiamo calcolarne il valore applicando la formula e risulterà che sarà necessario un solo bit per poter rappresentare questo caso.
Differentemente se la moneta venisse lanciata n volte le informazioni necessarie per questa sequenza di lanci risulterebbero essere pari ad n bit (mantenendo l'ipotesi che la moneta abbia uguali probabilità di uscire testa o croce).
Contrariamente, se la moneta avesse una maggiore probabilità di far uscire croce (es: p=0.1), l'evento sarebbe più raro e di conseguenza richiederebbe più di 3 bit di informazioni.

Oltre al calcolo delle informazioni per un evento è possibile quantificare in numero di informazioni presenti in una variabile casuale la quale prende il nome di entropia. L'entropia fornisce una misura della quantità media di informazioni necessarie per rappresentare un evento, ricavata da una distribuzione di probabilità per una variabile casuale.
Quest'ultima è calcolata come il negativo della somma della probabilità di ogni evento moltiplicato per il logaritmo della probabilità di ogni evento:

$$H(X) = -\sum_{i=1}^{K}(p(k) * \log(p(k)))$$

con X che rappresenta la variabile causale e K gli stati discreti.

L'entropia più grande per una variabile casuale sarà se tutti gli eventi sono ugualmente probabili e viceversa un'entropia più bassa quando quando una variabile causale ha probabilità che avvenga del 100%.

2.3.1 Indice di Theil

L'indice di Theil viene utilizzato in statistica principalmente in ambito socio-economico per misurare la disuguaglianza di reddito di una comunità e fenomeni economici.
Esso deriva dall'indice di entropia generalizzata.
In contesti economici l'indice Theil misura una "distanza" entropica che rappresenta la lontananza dallo stato egualitario (immaginato come ideale) nel quale tutti posseggono lo stesso reddito.
Più in generale invece è una misura di disuguaglianza basata sull'entropia, cioè calcola quanto si è lontani dalla situazione in cui tutte le unità presentano la stessa modalità del carattere.
Se viene pensata come entropia però potrebbe generare confusione in quanto l'entropia risulta avere un numero elevato per il "disordine" contrariamente all'indice di Theil che produce un valore alto per l' "ordine". Per questo motivo risulta più chiaro parlare in queso caso di ridondanza.
La formulazione dell'indice per rappresentare l'entropia negativa quindi viene calcolata la disuguaglianza piuttosto che di uguaglianza.
Definiamo:

* N come il numero di possibilità di una variabile di interesse presente nell'insieme preso in esame;

* un insieme $y_i = 1,2..,N$ dove j_i è la caratteristica iesima;

* μ come la media della variabile di interesse in tutte le regioni.

La quantità T definita come segue è detta indice di Theil.

$$T = \frac{1}{N} \sum_{i=1}^{N} \frac{y_j}{\mu} \ln \frac{y_j}{\mu}$$

Si noti che il valore di T è compreso ta 0 ed 1 con zero che rappresenta una ripartizione ben distribuita e valori più alti che rappresentano un livello di disuguaglianza più elevato.

2.4 Indice di diversità

Un indice di diversità è una misura matematica che ci permette di studiare la diversità delle specie in una comunità (definite prima più genericamente come le differenti modalità di un carattere).
In ambito biologico gli indici di diversità forniscono maggiori informazioni sulla composizione della comunità rispetto alla semplice ricchezza delle specie (cioè il numero di specie presenti); essi tengono anche conto delle abbondanze relative delle diverse specie.
Per fare un esempio si considerino due insieme di 100 individui composta da 10 specie diverse. Ipotizziamo un caso A nel quale sian presenti 10 individui di ciascuna specie e un caso B in qui 91 individui siano di una specie e un individuo per ogni specie restante. Quale comunità risulterà più diversificata?
Chiaramente l'esempio è ha una maggior diversificazione ma entrambe le comunità hanno la stessa ricchezza di specie.
Tenendo conto delle abbondanze relative, un indice di diversità dipende non solo dalla ricchezza delle specie, ma anche dall'uniformità, o equità, con cui gli individui sono distribuiti tra le diverse specie.
Per questo motivo l'importanza di questa analisi risiede nell'osservare importanti informazioni sulla rarità e la somiglianza delle specie in una comunità.
Tra i vari indici presenti in letteratura per misurare la diversità troviamo l'indice di Simpson, l'indice Shannon, l'indice Berger-Parker , il valore N1 dell'indice di diversità Hill e le statistiche Q.
Sia l'indice di Simpson che quello di Shannon sono comunemente utilizzati per caratterizzare la diversità delle specie in una comunità riuscendo a rappresentare sia l'abbondanza che l'uniformità delle specie presenti.

2.4.1 Indice di Shannon

Definiamo:

* S come il numero totali di possibili modalità in un determinato insieme;

* p_i la proporzione di S costituita dalla i-esima modalità;

Nell'indice di Shannon la proporzione della specie i rispetto al numero totale di specie p_i viene calcolata e poi moltiplicata per il logaritmo naturale di questa proporzione $\ln(p_i)$.
Il prodotto risultante viene sommato tra le specie e moltiplicato per -1. [7]
La quantità H definita come segue è detta indice di Shannon.

$$H = -\sum_{i=1}^{s}(p_i)\ln(p_i)$$

L'equità di Shannon (o uniformità) E_H può essere calcolata dividendo H per H_{max} (nel nostro caso $H_{max} = \ln S$).

$$E_H = \frac{H}{H_{MAX}} = \frac{H}{\ln(S)}$$

L'equità E_H assume un valore compreso tra 0 e 1 con 1 come completa uniformità.

Per comprendere meglio l'applicazione si può fare un esempio numerico.
Prendiamo in considerazione i due esempi precedenti dove prendiamo in considerazione due insieme di 1000 individui.
Supponiamo che siano presenti 5 caratteri diversi.
Ipotizziamo un caso A_{EQ} nel quale vi sia una distribuzione uniforme di caratteri (quindi con 20 individui per ciascun carattere) e un caso A_{UN} in qui il 90% degli individui siano appartenenti ad un unico carattere e gli altri individui sono distribuiti uniformemente tra le restanti caratteri.
Nel caso A_{EQ} il valore di p_i per qualunque valore di i compreso tra 1 e s varrà 0.2. Differentemente, nel caso A_{UN} il valore di p per il carattere dominante varrà 0.9 e per gli altri caratteri 0.02. Replichiamo lo stesso ragionamento per i casi B_{EQ} , B_{UN} , C_{EQ} , C_{UN} , D_{EQ} , D_{UN} con rispettivamente 10 differenti caratteri per il caso B, 20 per il caso C e 50 per il caso D.
Il numerico dei differenti è sintetizzato per chiarezza n figura 2.2.

	CARATTERE DOMINANTE	CARATTERE NON DOMINANTE
A_EQ	200	200 (per ognuno dei 4 caratteri rimanenti)
A_UN	900	25 (per ognuno dei 4 caratteri rimanenti)
B_EQ	100	100 (per ognuno dei 9 caratteri rimanenti)
B_UN	901	11 (per ognuno dei 9 caratteri rimanenti)
C_EQ	50	50 (per ognuno dei 19 caratteri rimanenti)
C_UN	905	5 (per ognuno dei 19 caratteri rimanenti)
D_EQ	20	20 (per ognuno dei 49 caratteri rimanenti)
D_UN	902	2 (per ognuno dei 49 caratteri rimanenti)
TOTALI	**1000 INDIVIDUI PER OGNI DISTINTO CASO**	

Figura 2.2. Distribuzione dei caratteri nelle comunità prese in esame

Andando a calcolare il valore di H e successivamente quello di E_H possiamo vedere nel grafico l'andamento dei valori e dell'impatto della non uniformità ben visibile nell'indice di Shannon.
Nel grafico rappresentato in figura 2.3 sono stati rappresentate sull'asse delle x il numeri di differenti caratteri presenti all'interno della comunità (5,10,20 e 50), sull'asse delle y il rispettivo valore di E_H.
In verde sono stati rappresentati i quattro casi di distribuzioni uniformi di caratteri (eq) invece in blu i casi di non uniformità (un).

2.4.2 Indice di Simpson

Definiamo:

* S come il numero totali di possibili modalità in un determinato insieme;

* p_i la proporzione di S costituita dalla i-esima modalità;

Nell'indice di Simpson la proporzione delle modalità del carattere i rispetto al numero totale p_i viene calcolata sommando le proporzioni al quadrato di tutte e

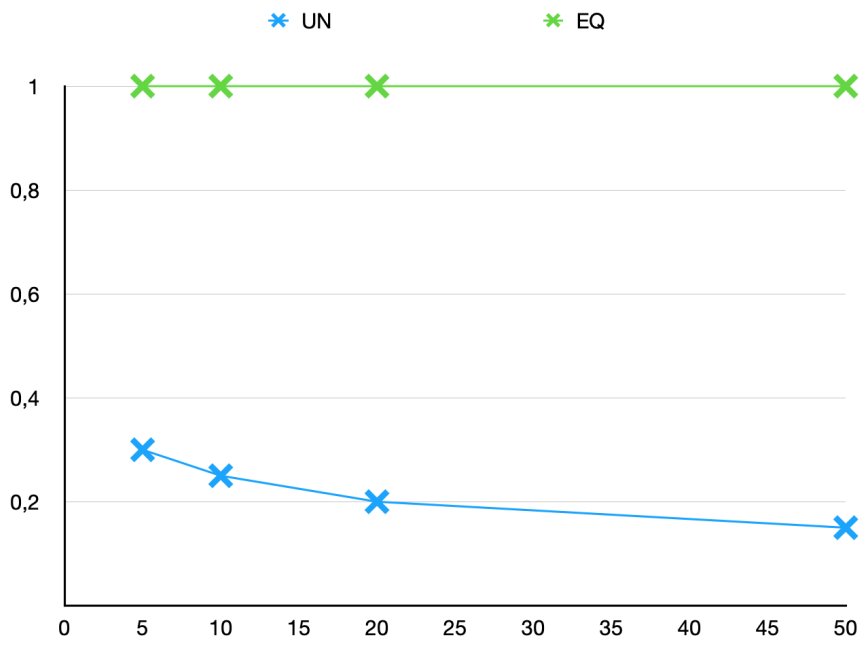

Figura 2.3. Grafico rappresentate i differenti valori di equità di Shannon

prendendone il reciproco.
[8] La quantità D definita come segue è detta indice di Simpson.

$$D = \frac{1}{\sum_{i=1}^{s}(p_i)^2}$$

Dato un determinato valore di S , D aumenta all'aumentare dell'equità e una data equità D aumenta all'aumentare delle modalità.
L'equità E_D può essere calcolata prendendo l'indice di Simpson D ed esprimendolo come proporzione del valore massimo che D potrebbe assumere se gli individui nella comunità fossero completamente distribuiti in modo uniforme (ovvero $D_{MAX} = S$).

$$E_D = \frac{D}{D_{MAX}}$$

L'equità assume un valore compreso tra 0 e 1, con 1 come completa uniformità.
In ambito biologico per la misurazione delle biodiversità le teorie di Shannon e Simpson non sono sufficientemente adeguate poiché pongono le specie su un medesimo livello quando in realtà alcune hanno valori ecologici più elevanti svolgendo

un' attività di controllo permettendo la coesistenza di un numero elevato di specie. Nonostante questa considerazione questi indici paiono non riscontrare questo limite essendo utilizzabili per calcolare l'eterogeneità dei nostri dataset dovendo analizzare attributi sensibili tutti con un medesimo peso.

Per comprendere meglio l'applicazione si può fare un esempio numerico. Riprendiamo l'esempio precedente prendendo in considerazione solamente i casi non uniformi (unequal). Su di essi rappresentiamo graficamente i seguenti valori:

* H ovvero l'indice di Shannon;
* E_H l'equita dell'indice di Shannon;
* D ovvero l'indice di Simpson;
* D_H l'equita dell'indice di Simpson;

ponendo (come precedentemente) sull'asse delle x il numero di differenti caratteri presenti all'interno della comunità (5,10,20 e 50) e sull'asse delle y i rispettivi valori calcolati.

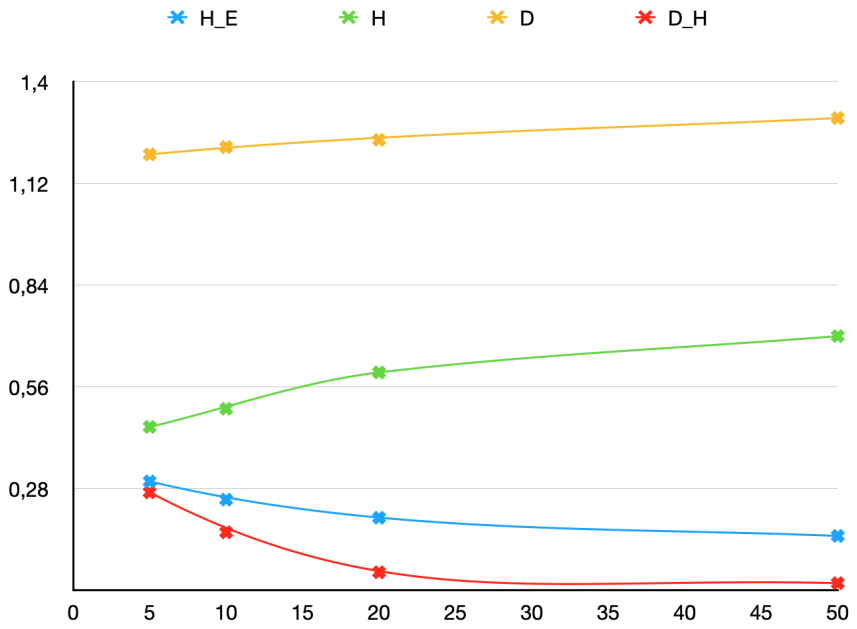

Figura 2.4. Grafico rappresentate i differenti indici di diversità

Come si può notare dai valori numerici rappresentati in figura 2.4 l'indice di Shannon, a causa della presenza del logaritmo, dà relativamente più peso, rispetto all'indice di Simpson, alle modalità più rare.

2.5 Correlazione e collinearità

La correlazione misura la relazione tra due variabili.
Quando due variabili sono altamente correlate al punto dal poter prevedere una conoscendo l'altra abbiamo la collinearità tra le variabili.
La multicollinearità si verifica quando le variabili indipendenti in un modello di regressione sono correlate [12].
La multicollinearità causa principalmente due problemi: il primo riguarda le stime dei coefficienti mentre il secondo la precisione dei coefficienti di stima.
I coefficienti di regressione sono delle stime dei parametri che descrivono la relazione tra una variabile predittrice e la rispettiva risposta.
Per esempio, nella regressione lineare, i coefficienti sono i valori che moltiplicano i valori del predittore.
Supponendo di avere la seguente equazione di regressione:

$$Y = +5X + 10$$

il valore +5 risulta essere il coefficiente, X il predittore e +10 la costante.
Il segno dei coefficienti indicano la direzione della relazione tra una variabile predittrice e la variabile di risposta.
Un valore positivo indica che all'aumentare della variabile predittrice aumenta anche la variabile di risposta, viceversa per un valore negativo.
Le stime dei coefficienti possono oscillare in modo incontrollato in base a quali altre variabili indipendenti sono presenti nel modello. In questo modo i coefficienti diventano molto sensibili ai piccoli cambiamenti del modello.
Inoltre la multicollinearità riduce la precisione dei coefficienti di stima indebolendo la potenza statistica del modello di regressione.

2.5.1 Statistica inferenziale

La statistica inferenziale consiste in metodi statistici utilizzati per testare se vi sono relazioni tra variabili.
Ad esempio nel caso di studenti iscritti ad una facoltà di ingegneria una possibile relazione potrebbe riscontrarsi tra gli studenti provenienti da precedenti studi scientifici.
Per illustrare il rapporto tra queste variabili è necessario utilizzare le statistiche

inferenziali per dimostrare in modo più rigoroso se esiste o meno una relazione tra queste due variabili [14].
Tra queste troviamo il coefficiente di correlazione di Pearson (noto anche come R di Pearson) e il test del chi-quadrato.

2.5.2 Coefficiente di Pearson

Il coefficiente di correlazione di Pearson ha lo scopo di determinare se esiste una relazione significativa (cioè una correlazione) tra due variabili.
Questo coefficiente può avere un valore compreso tra -1 e +1 dove 0 rappresente la totale disgiunzione delle variabili (nessuna relazione).
Un coefficiente di correlazione di +1 significa che c'è una perfetta correlazione positiva ovvero che all'aumentare di una variabile, la seconda variabile aumenta proporzionalmente.
Viceversa con coefficiente di correlazione -1 vi è una perfetta correlazione negativa nella quale la seconda variabile diminuisce con la stessa proporzione.

Definiamo:

* $\sum xy$ come la somma del prodotto di tutte le nostre coppie di dati;

* $\bar{x}, \bar{y}$ come la media di entrambe le variabili;

* $\sum x^2, \sum y^2$ come la somma dei valori al quadrato di entrambe le variabili;

* N come numeri di casi.

La quantità R è definita come coefficiente di Pearson.

$$R = \frac{\sum xy - N\bar{x}\bar{y}}{\sqrt{(\sum x^2 - N\bar{x^2})(\sum y^2 - N\bar{y^2})}}$$

Dopo aver calcolato il coefficiente è necessario calcolare il valore α per determinare se questa correlazione è statisticamente significativa o meno (confrontando i valori tramite una tabella standard rappresentata in figura 2.5) ovvero per vedere quanto è probabile che una R si sia verificata per caso.
Per consultare la tabella è necessario calcolare i gradi di libertà tramite la formula:

$$df = N - 2$$

df \ α	0.2	0.1	0.05	0.02	0.01	0.001
1	0.951057	0.987688	0.996917	0.999507	0.999877	0.999999
2	0.800000	0.900000	0.950000	0.980000	0.990000	0.999000
3	0.687049	0.805384	0.878339	0.934333	0.958735	0.991139
4	0.608400	0.729299	0.811401	0.882194	0.917200	0.974068
5	0.550863	0.669439	0.754492	0.832874	0.874526	0.950883
6	0.506727	0.621489	0.706734	0.788720	0.834342	0.924904
7	0.471589	0.582206	0.666384	0.749776	0.797681	0.898260
8	0.442796	0.549357	0.631897	0.715459	0.764592	0.872115
9	0.418662	0.521404	0.602069	0.685095	0.734786	0.847047
10	0.398062	0.497265	0.575983	0.658070	0.707888	0.823305
11	0.380216	0.476156	0.552943	0.633863	0.683528	0.800962
12	0.364562	0.457500	0.532413	0.612047	0.661376	0.779998

Figura 2.5. Tabella dei valori critici di Pearson

Per comprendere meglio l'applicazione si può fare un esempio numerico.
Si considerino i dati rappresentati in tabella 2.6 che rappresentano 10 soggetti differenti e i rispettivi risultati ottenuti ad un test per rilevare la propensione alla facoltà di ingegneria e la media dei voti in aerea scientifica a fine anno.

Studente	Test	Media dei voti
A	8	6.5
B	7	7.5
C	6	6
D	5	7
E	10	8
F	3	6
G	4	7
H	12	8.5
I	11	8
L	**15**	**9.5**

Figura 2.6. Valori per rilevare la propensione alla facoltà di ingegneria

I numeri ci casi (N) presi nell'esempio sono 10. Applicando la formula è possibile calcolare il valore di R:

$$\sum xy = (8x6.5)+(7x7.5)+(6x6)+(5x7)+(10x8)+(3x6)+(4x7)+(12x8.5)+(11x8)+(15x9.5)$$

$$\bar{x} = \frac{8+7+6+5+10+3+4+12+11+15}{N}$$

$$\bar{y} = \frac{6.5+7.5+6+7+8+6+7+8.5+8+9.5}{N}$$

il quale risulta essere pari a 0.89.
Calcolando successivamente i gradi di libertà identificandoli in 8 e consultando la tabella dei valori critici di Pearson è conferma la relazione positiva tra le medie dei voti e il test sostenuto.

2.5.3 Test Chi-quadrato dell'indipendenza

Il test del Chi-Square è un test di indipendenza determina se esiste un'associazione tra variabili ovvero se le variabili sono indipendenti o correlate.
Esso valuta l'esistenza di una relazione tra due variabili non specificando il tipo di relazione. Il test consiste nell'utilizzo di una tabella di crosstabulazione per analizzare i dati nella quale questi ultimi sono classificati secondo due variabili categoriche [13].
Tutte le possibili modalità di una prima variabile verranno poste sulle colonne e per la seconda variabile sulle righe.
Per poter trovare applicazione test i dati devono soddisfare diversi requisiti tra i quali:

- Ogni variabile deve avere due o più modalità;
- Le variabili non devono essere associate in alcun modo (ad esempio osservazioni pre-test/post-test);
- Le dimensioni del campione relativamente grandi;
- Le frequenze previste per ogni cella sono almeno 1 e almeno 5 per la maggioranza delle celle.

Definiamo:

* o_i come la frequenza osservata;
* e_i come la frequenza prevista;
* i il numero della cella.

La quantità X^2 è definita come la statistica chi-quadrato.

$$X^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{(e_i)}$$

dove:

$$\frac{o_i - e_i}{\sqrt{e_i}}$$

viene indicato come il residuo della cella.
Il residuo serve a confrontare i conteggi osservati con quelli previsti.
Il segno (positivo o negativo) indica se la frequenza osservata nella cella i è superiore o inferiore al valore misurato nel modello.
Il valore calcolato di X^2 viene poi confrontato con il valore critico della tabella di distribuzione X^2 rappresentata in figura 2.7.

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22

Figura 2.7. Valori per rilevare la propensione alla facoltà di ingegneria

Per consultare la tabella è necessario calcolare i gradi di libertà tramite la formula:

$$df = N - 1$$

Il valore X^2 calcolato non deve essere maggiore del valore di soglia critico.

Per comprendere meglio l'applicazione si può fare un esempio numerico. Si considerino i dati rappresentati in tabella 2.8 dove vengono presi in esame 100 studenti ripartiti in 6 classi.

$$X^2 = \frac{(21-16)^2}{16} + \frac{(12-16)^2}{16} + \frac{(16-16)^2}{16} + \frac{(19-16)^2}{16} + \frac{(14-16)^2}{16} + \frac{(18-16)^2}{16}$$

$$X^2 \approx 3.6$$

Per poter stabilire se il valore 3.6 è significativo o meno è necessario calcolare i gradi di libertà per poi poter consultare la tabella con i valori corretti.

Classe	Frequenza
A	21
B	12
C	16
D	19
E	14
F	18

Figura 2.8. Frequenza di cento studenti ripartiti in 6 differenti classi

In questo caso il grado di libertà risulta essere 5.
Consultando la tabella 2.7 il valora risulta essere inferiore al p= 0.75 ma superiore al p=0.5 quindi non vi è una forte dipendenza tra le variabili.

2.5.4 Coefficiente di Spearman

L'indice di correlazione R per ranghi di Spearman è una misura statistica non parametrica di correlazione.
Grazie ad essa è possibile misurare il grado di relazione tra due variabili.
L'unica ipotesi sui valori delle variabili è che essi siano ordinabili, e, se possibile, continui.
Diversamente dal coefficiente di correlazione lineare di Pearson, il coefficiente di Spearman non misura una relazione lineare anche qualora vengano usate misure poste ad intervalli. Infatti esso permette di stabilire quanto bene una relazione tra due variabili può essere descritta usando una funzione monotona [15] .
A livello pratico il coefficiente è rappresentato come p_r ed è un caso particolare del coefficiente di correlazione di Pearson dove i valori vengono convertiti in ranghi prima di calcolare il coefficiente.
Il coefficiente assume i valori compresi tra – 1 e + 1 indicando nel segno e nel valore il tipo e la forza della correlazione.
Il segno positivo indica una correlazione direttamente proporzionale ed il segno negativo indica una correlazione inversamente proporzionale; valori di p_r vicini ad 1 indicano una forte correlazione positiva mentre valori vicini allo 0 una correlazione nulla.
[16]

Capitolo 3

Caso di studio

ProPublica è un'organizzazione no-profit americana che mira a produrre giornalismo investigativo di interesse pubblico.
Nel 2016 pubblicò un articolo scritto da Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin nel quale evidenziava un problema di carattere razziale all'interno di un software utilizzato nei tribunali statunitensi per aiutare i giudici a prendere una decisione giuridica in base ad una probabilità di reiterazione del reato. [10]
Come la stessa ProPublica afferma nel suo lavoro di analisi, precedentemente erano stati svolti dei test per valutare l'equità di questo algoritmo ma le indagini erano state completate per la maggior parte dei casi dalle stesse persone che avevano sviluppato il software.
Il software in questione si chiama Compas (acronimo di Correctional Offender Management Profiling for Alternative Sanctions).
Sulla base di uno studio di follow up di 2 anni (cioè chi ha effettivamente commesso crimini o crimini violenti dopo 2 anni) venne dimostrato che l'algoritmo non era equo e sottostimava la pricolosità delle persone di pelle bianca rispetto a quelle di pelle scura.
Su una nota piattaforma di dati opensource di nome Kaggle è disponibile il dataset oggetto di analisi.
Pro publica nel suo studio analizzava come gli imputati neri avessero più probabilità di risultare con un indice di recidività maggiore rispetto alle persone di pelle bianca valutandone i dati.
Nei capitoli precedenti abbiamo capito che gli algoritmi di machine learning sono fortemente influenzati dai dati utilizzati. Per questo motivo inizialmente analizziamo in maniera generale i dati per poi applicare le metriche analizzate nel capitolo 2.

3.1 Librerie per analisi dei dati

Ho deciso di utilizzare come linguaggio di programmazione Python che fornisce diverse librerie specifiche per l'analisi dei dati. [11]
La libraria utilizzata principalmente si chiama Pandas. Essa fornisce strutture dati di alto livello e funzioni progettate per rendere il lavoro con dati strutturati o tabellari veloce, facile ed espressivo.
Questa libreria è basata su due oggetti fondamentali: Serie e DataFrame.
La Serie è un oggetto monodimensionale simile ad un array contenente una sequenza di valori (di tipo simile ai tipi NumPy) e un array associato di etichette di dati, chiamato indice.
Il DataFrame è una struttura dati tabulare, orientata alle colonne, con etichette sia a riga che a colonna.
Rappresenta una tabella rettangolare di dati e contiene una raccolta ordinata di colonne, ognuna delle quali può essere di un diverso tipo di valore (numerico, stringa, booleano, ecc.).
E' caratterizzato sia un indice di riga che da uno di colonna.
Nonostante sia bidimensionale ci permette di utilizzarlo in caso di maggiori dimensioni utilizzando una struttura gerarchica.

3.2 Caso COMPAS

I dati contengono variabili utilizzate dall'algoritmo COMPAS per assegnare un punteggio agli imputati, insieme ai loro risultati entro 2 anni dalla decisione, per oltre 10.000 imputati penali nella Broward County, Florida.
Sulla piattaforma Kaggle vengono forniti tre sottoinsiemi di dati differenti che comprendono informazioni distinte come, ad esempio, la sola recidività violenta (in contrapposizione all'essere incarcerato nuovamente per reati non violenti come il vagabondaggio o lo spaccio di sostanze stupefacenti).
Per l'esplorazione dei dati prenderò in esame in un primo moneto due differenti esempi.
Il primo lo identificherò come caso A e utilizzerò uno dei tre sottoinsiemi opensource della piattaforma Kaggle (quello ritenuto più conforme all' analisi).
Il secondo caso lo identificherò come caso B ed utilizzerò lo stesso database ma modificato nelle frequenze di alcuni attributi ritenuti importanti mantenendo invariati tutti gli altri elementi.
Inizio esplorando la dimensionalità dei dati che abbiamo a disposizione, i nomi e le loro caratteristiche (equivalenti per i due casi).

```
df = pd.read_csv('compas-scores-raw.csv')
```

```
df = df.drop_duplicates(subset="Person_ID")
print(df.shape)
```

con questi comandi estrapolo le informazioni di dimensionalità del dataset:

```
(18610, 28)
```

ovvero ci sono a dispisizione 18610 record composti da 28 colonne. Le colonne rappresentano le diverse informazioni per ogni record.

```
print(df.columns)
```

estrapolo le informazioni sulle colonne (per andare ad individuare le caratteristiche delle informazioni che disponiamo per ogni record):

```
Index(['Person_ID', 'AssessmentID', 'Case_ID', 'Agency_Text', 'LastName',
       'FirstName', 'MiddleName', 'Sex_Code_Text', 'Ethnic_Code_Text',
       'DateOfBirth', 'ScaleSet_ID', 'ScaleSet', 'AssessmentReason',
       'Language', 'LegalStatus', 'CustodyStatus', 'MaritalStatus',
       'Screening_Date', 'RecSupervisionLevel', 'RecSupervisionLevelText',
       'Scale_ID', 'DisplayText', 'RawScore', 'DecileScore', 'ScoreText',
       'AssessmentType', 'IsCompleted', 'IsDeleted'],
      dtype='object')
```

Di queste le più significative per l' indagine sulle discriminazioni razziali (ovvero i nostri attributi sensibili) risultano essere:

* Sex_Code_Text

* Ethnic_Code_Text

* Marital Status

Oltre a queste tre informazioni prendo in considerazione anche:

* ScoreText

in quanto è un attributo che può assumere tre differenti valori: Low, Medium e High in base rispettivamente ad una bassa, media o alta probabilità di recidiva.

3.2.1 Attributi sensibili

Prima di iniziare con l'analisi ho pulito i dati andando, per esempio, ad eliminare i duplicati.

```
df = df.drop_duplicates(subset="Person_ID")
```

Per quanto riguarda l'attributo 'Ethnic_Code_Text' ho aggregato i dati che ci interessano per l'analisi successiva uniformando i valori presenti:

```
df["Ethnic_Code_Text"]=df["Ethnic_Code_Text"].
replace("African-Am","African-American")
df["Ethnic_Code_Text"]=df["Ethnic_Code_Text"].
replace("Asian","Oriental")
```

I record sono così ripartiti per il caso A:

```
African-American    8125
Caucasian           6742
Hispanic            2728
Other                808
Oriental             117
Native American       65
Arabic                25
```

La frequenza per le differenti etnie è maggiormente visibile in figura 3.1.

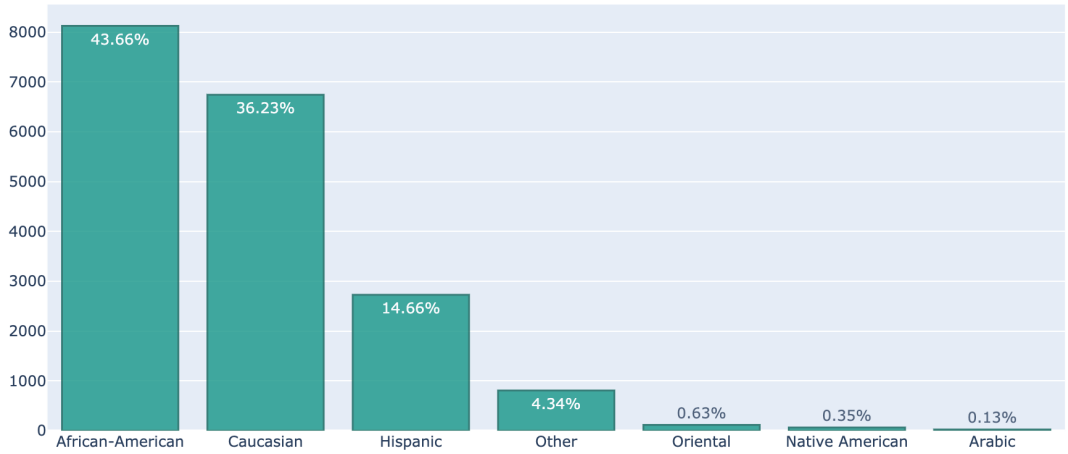

Figura 3.1. Caso A: frequenza per l'attributo che identifica l'etnia

Per il caso B nella creazione dei dati ho distribuito in modo uniforme le frequenze dell'attributo 'Ethnic_Code_Text' .
I record sono così ripartiti per il caso B:

```
African-American    3722
Caucasian           3722
Other               3722
Hispanic            3722
Oriental            3722
```

La frequenza per le differenti etnie è maggiormente visibile in figura 3.2.

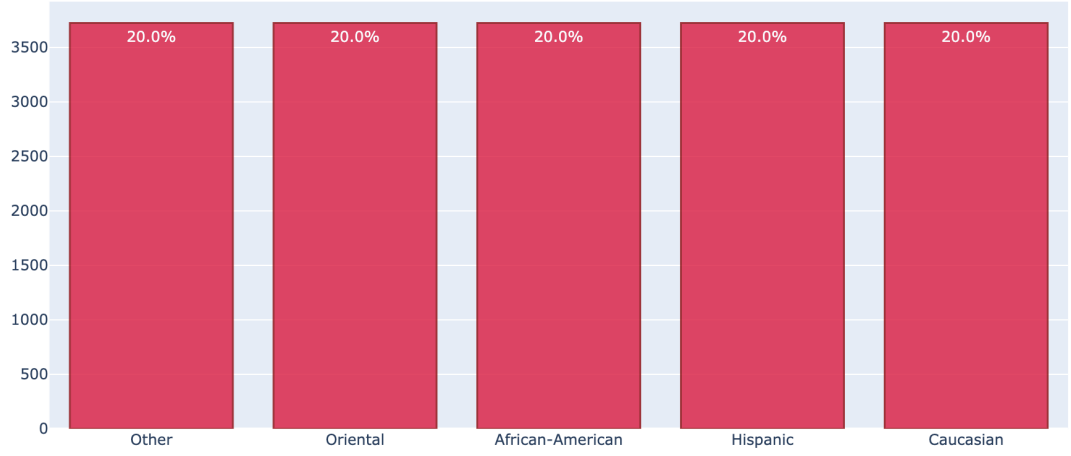

Figura 3.2. Caso B: frequenza per l'attributo che identifica l'etnia

Applico un'aggregazione per suddividere ulteriormente i dati in base all'attributo Sex_Code_Text evidenziando questo tipo di ripartizione per il caso A:

```
Ethnic_Code_Text  Sex_Code_Text
African-American  Female          1724
                  Male            6401
Arabic            Female             1
                  Male              24
Caucasian         Female          1706
                  Male            5036
Hispanic          Female           549
                  Male            2179
Native American   Female             19
```

	Male	46
Oriental	Female	22
	Male	95
Other	Female	146
	Male	662

La frequenza per le differenti etnie aggregate per Sex_Code_Text è maggiormente visibile in figura 3.3.

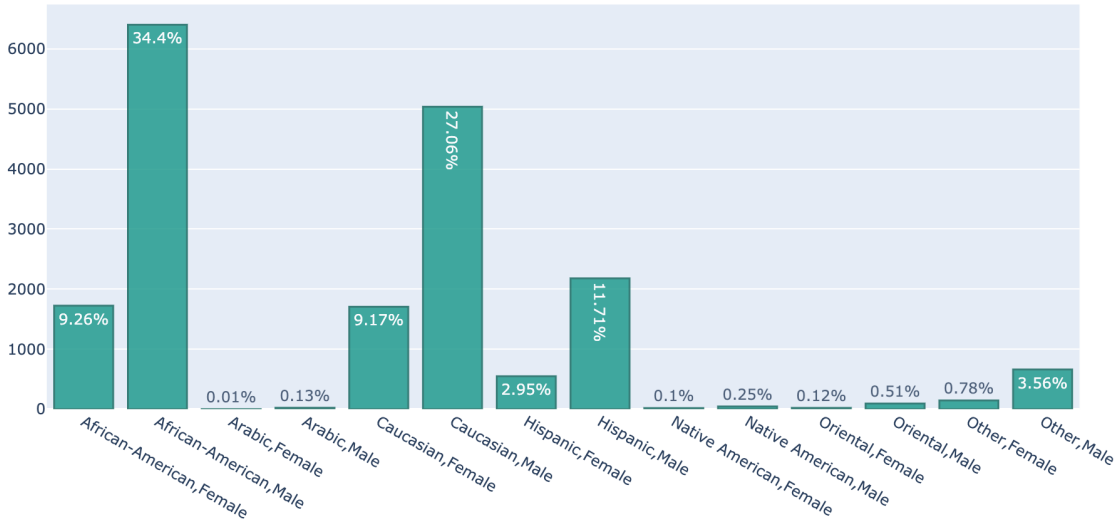

Figura 3.3. Caso A: frequenza per l'attributo che identifica l'etnia suddivisa per l'attributo che identifica il sesso

Applico la stessa aggregazione per il caso B:

African-American	Female	1014
	Male	2708
Caucasian	Female	1123
	Male	2599
Hispanic	Female	749
	Male	2973
Oriental	Female	832
	Male	2890
Other	Female	910
	Male	2812

Anch'esso visibile graficamente in figura 3.4 .

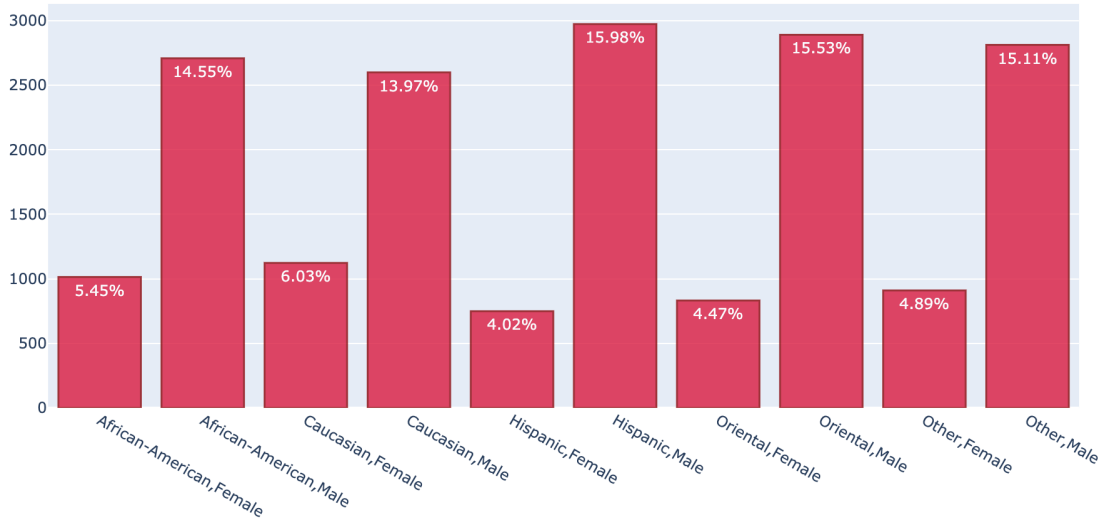

Figura 3.4. Caso B: frequenza per l'attributo che identifica l'etnia suddivisa per l'attributo che identifica il sesso

Applico per il caso A le metriche analizzate nel capitolo precedente rappresentando graficamente col colore verde le frequenze rappresentate in figura 3.1 e col colore verde petrolio le frequenze rappresentate in figura 3.3.
Una diversa aggregazione porta alla formazione di sottoclassi che variano le frequenze e di conseguenza i valori delle metriche applicate.
Li rappresento tutti insieme sullo stesso grafico per vederne le differenze in figura 3.6.

Nello stesso modo applico al caso B le metriche rappresentando graficamente col colore rosso le frequenze rappresentate in figura 3.2 e col colore arancione le frequenze rappresentate in figura 3.4.
Li rappresento tutti insieme sullo stesso grafico per vederne le differenze in figura 3.5.

Andando ad analizzare i valori degli indici per il caso B nel caso di aggregazione dei dati solamente per l'attributo 'Ethnic_Code_Text':

* Indice di Gini = 1

* Indice di Shannon = 1

* Indice di Simpson = 1

* Indice di Theil = 0

Figura 3.5. Caso B: calcolo dell' indice di Gini, indice di Shannon, indice di Simpson ed indice di Theil

Questi valori rappresentano in caso di una perfetta eterogenità e completa uniformità tra le possibili variabili qualitative.

Andando invece ad analizzare i valori degli indici nel caso di aggregazione dei dati anche per l'attributo Sex_Code_Text creati appositamente con delle classi predominanti (con frequenza maggiore del 14%) notiamo come l'indice di Simpson decresce maggiormente verso valori bassi rispetto all'indice di Shannon in quanto, senza il logaritmo all'interno della sommatoria, viene dato meno peso alle modalità più rare (in questo caso quelle con meno del 2%).

Analizzando i valori degli indici per il caso A (figura 3.6) nel caso di aggregazione dei dati per l'attributo Ethnic_Code_Text congiuntamente a Sex_Code_Text (in caso più rilevante in quanto la discriminazione riscontrata era di tipo razziale per quanto riguarda il valore di possibile recidività):

* Indice di Gini = 0.84
* Indice di Shannon = 0.64
* Indice di Simpson = 0.32
* Indice di Theil = 0.60

Figura 3.6. Caso A: calcolo dell' indice di Gini, indice di Shannon, indice di Simpson ed indice di Theil

Questi valori ci confermano che vi è uno sbilanciamento dei dati come già dimostrato negli studi precedenti [10]

Le analisi mostrano dati molto sbilanciati considerando le classi protette. Circa il 36% delle osservazioni del dataset si riferiscono a persone bianche, mentre circa il 44% si riferiscono a persone di colore, indicando che potrebbe esserci una sovrastima dell'attributo razziale che contribuirebbe alla stima della recidiva.
Coerentemente a questo la frequenza di persone di colore che ricevono un punteggio elevato è nettamente maggiore e la distribuzione è concentrata su poche classi. Questo ci viene confermato dai valori degli indici calcolati.
Sembra stonare l'indice di Shannon ma come sappiamo dalla formula esso da meno peso alle modalità più rare e quindi è meno sensibile.

Ho ripetuto le stesse analisi svolte per l'attributo rappresentante l'etnia aggregando per l'attributo MaritalStatus (ovvero stato civile) e successivamente aggregandoli per Sex_Code_Text e Marital Status. per gli altri due attributi sensibili individuati nel database (Sex_Code_Text e Marital Status rispettivamente sesso e stato civile).
I record sono così ripartiti per etnia e stato civile evidenziando due classi dominanti che corrisponde a African-American e Caucasian single:

African-American	Divorced	192
	Married	784
	Separated	177
	Significant Other	204
	Single	6702
	Unknown	31
	Widowed	35
Arabic	Married	6
	Separated	1
	Significant Other	1
	Single	17
Caucasian	Divorced	702
	Married	950
	Separated	180
	Significant Other	120
	Single	4694
	Unknown	23
	Widowed	73
Hispanic	Divorced	256
	Married	569
	Separated	134
	Significant Other	38
	Single	1706
	Unknown	17
	Widowed	8
Native American	Divorced	3
	Married	4
	Separated	8
	Significant Other	1
	Single	47
	Widowed	2
Oriental	Divorced	12
	Married	38
	Separated	5
	Single	62
Other	Divorced	31
	Married	218
	Separated	40
	Significant Other	37
	Single	480
	Widowed	2

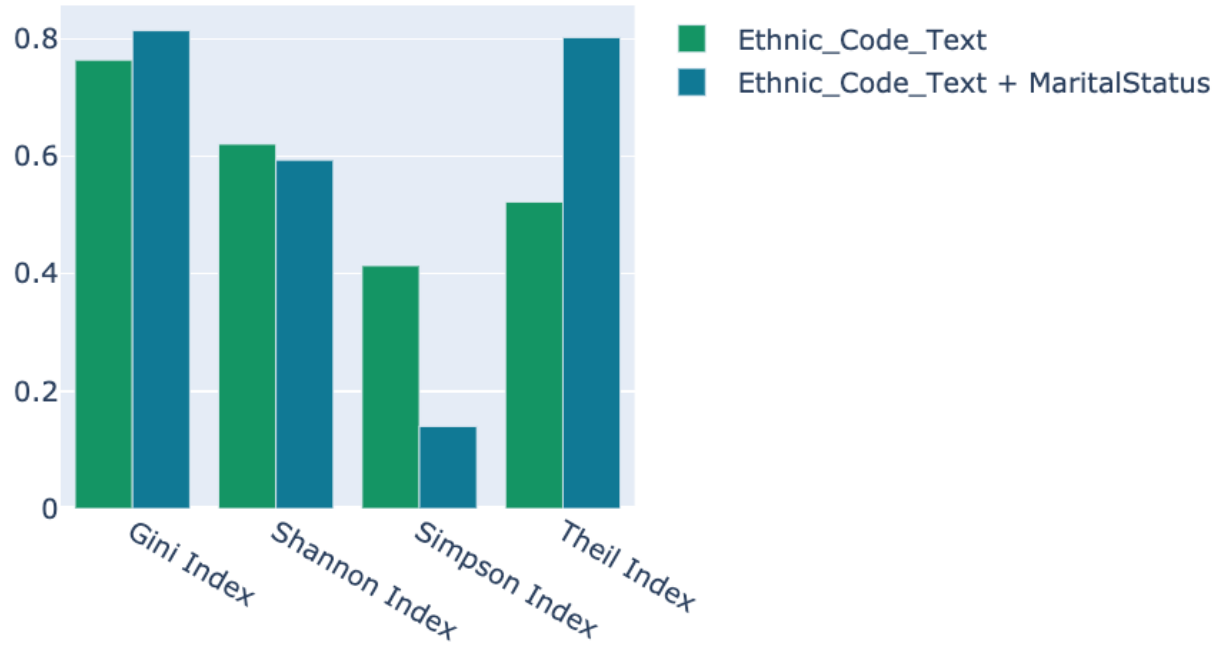

Figura 3.7. Caso A: attributo calcolo dell' indice di Gini, indice di Shannon, indice di Simpson ed indice di Theil sull'attributo sensible Ethnic_Code_Text e MaritalStatus

e per il sesso e lo stato civile:

Female	Divorced	351
	Married	507
	Separated	160
	Significant Other	80
	Single	3005
	Unknown	16
	Widowed	48
Male	Divorced	845
	Married	2062
	Separated	385
	Significant Other	321
	Single	10703
	Unknown	55
	Widowed	72

I valori degli indici sono riassunti nelle figure 3.7 e 3.8. Essi confermano le stesse considerazioni fatte nel caso dell'etnia.

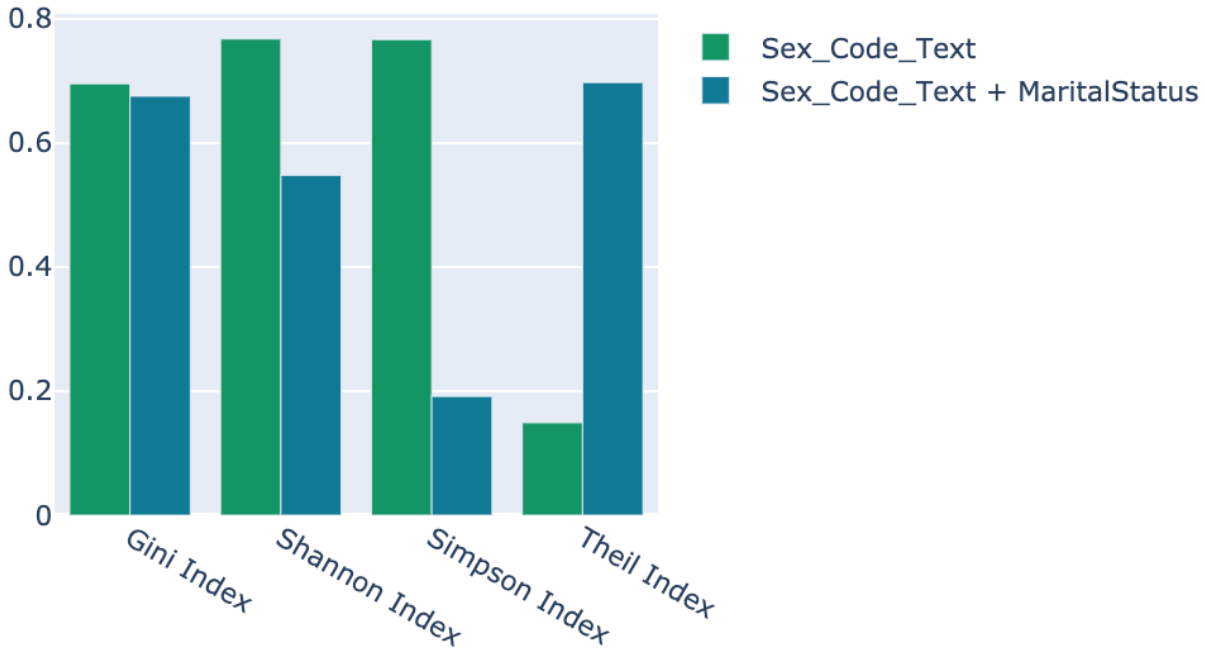

Figura 3.8. Caso A: attributo calcolo dell' indice di Gini, indice di Shannon, indice di Simpson ed indice di Theil sull'attributo sensibile Sex_Code_Text e MaritalStatus

3.2.2 Numero di classi

Riprendendo in esame il caso A (il database originale opensource tratto della piattaforma Kaggle) circa il 44% dei record si riferiscono a persone di colore contro solamente il 33% sono persone con pelle chiara.
Nonostante questa grossa sproporzionalità gli indici di Gini e di Shannon appaiono più alti di quanto ci saremmo aspettati.
Per capire la causa di questo ho preso in esame un caso C creando un database sulla falsa riga di quello originale ma variando il numero di categorie dell'attributo sensibile riguardante l'etnia.
La scelta dell'attributo sensibile è casuale in quanto questi dati (come nel caso B nel quale ho variato le frequenze) non sono fondati su alcuna base reale ma creati per analizzare meglio il comportamento delle metriche.

Confrontiamo il caso B nel quale avevamo 5 differenti etnie che aggregate con il sesso era composto da 10 classi differenti e il caso C con 10 differenti etnie e quindi 20 classi differenti.
Manteniamo le proporzioni così da far variare solamente il numero di classi aumentandole di numero.

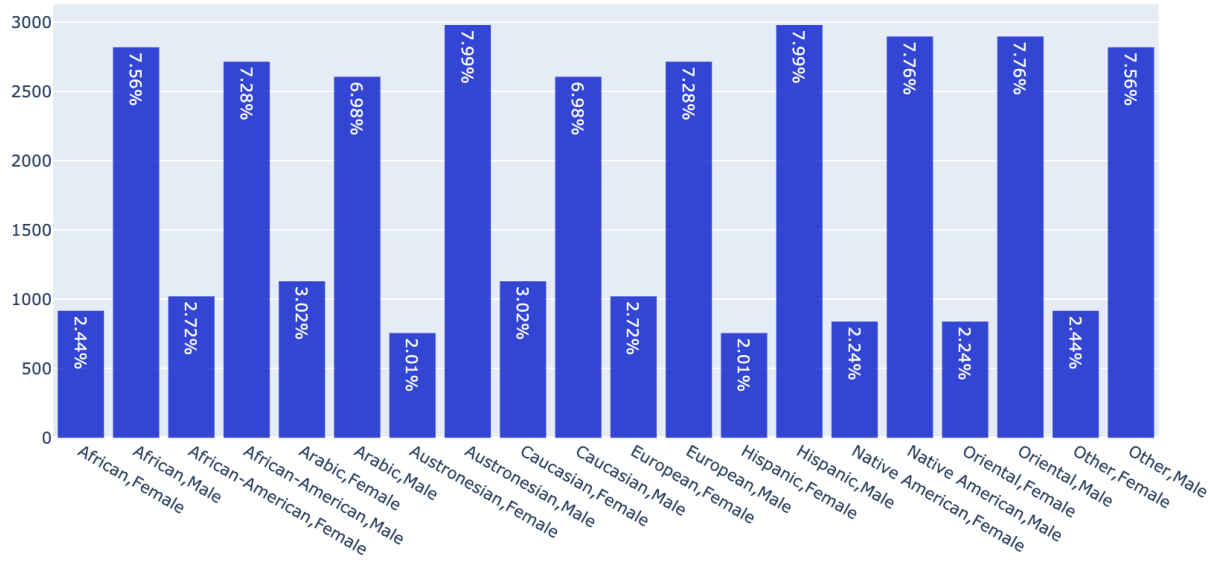

Figura 3.9. Caso C: frequenza per l'attributo che identifica l'etnia suddivisa per le tre categorie che rappresentano la probabilità di recidiva

In figura 3.4 avevamo già raffigurato le frequenze delle differenti classi del caso B. Quelle del caso C sono raffigurate in figura 3.9.

Come si può evincere dal grafico in figura 3.10, mettendo a confronto le frequenze si può notare come siano state mantenute per entrambe i casi le proporzioni nonostante l'aumento delle classi mantenendo $\frac{1}{2}$ delle classi con più di 2700 record 3 $\frac{1}{2}$ con numero di record inferiore a 800.

Premettendo queste condizioni mi aspetterei uno stesso valore degli indici ma così non è.
Infatti per il caso B:

* Indice di Gini = 0.94
* Indice di Shannon = 0.86
* Indice di Simpson = 0.57
* Indice di Theil = 0.31

Invece per il caso C:

* Indice di Gini = 0.98

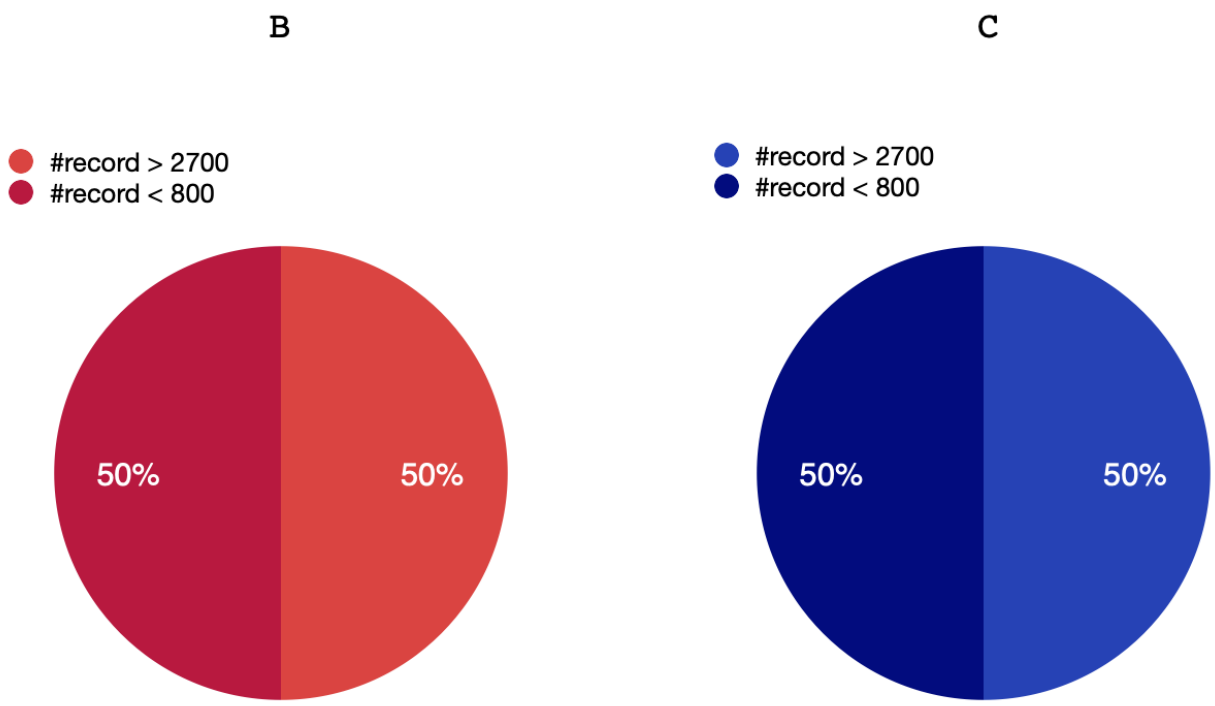

Figura 3.10. Caso B e C: proporzioni delle classi in base al numero di record

* Indice di Shannon = 0.89

* Indice di Simpson = 0.57

* Indice di Theil = 0.31

In figura 3.11 sono stati messi a confronto gli indici calcolati per il caso B in rosso e per il caso C in blu per poterne vedere meglio le differenze.

Dato che l'unico elemento differente tra di due dataset utilizzati per i due casi era in numero di classi è possibile identificare in questo la causa dei differenti valori degli indici calcolati.
Difatti l'indice di Gini e l'indice di Shannon risultano più altri nel caso C ovvero quando in numero di classi è superiore.
Per questo motivo in alcuni casi questi due indici potrebbero non riflettere bene lo squilibrio degli attributi, assumendo valori più alti di quanto ci si potrebbe aspettare nonostante la grande disproporzioni date dalle frequenze.
Possiamo altresì notare come l'indice di Simpson e l'indice di Theil non sembrano

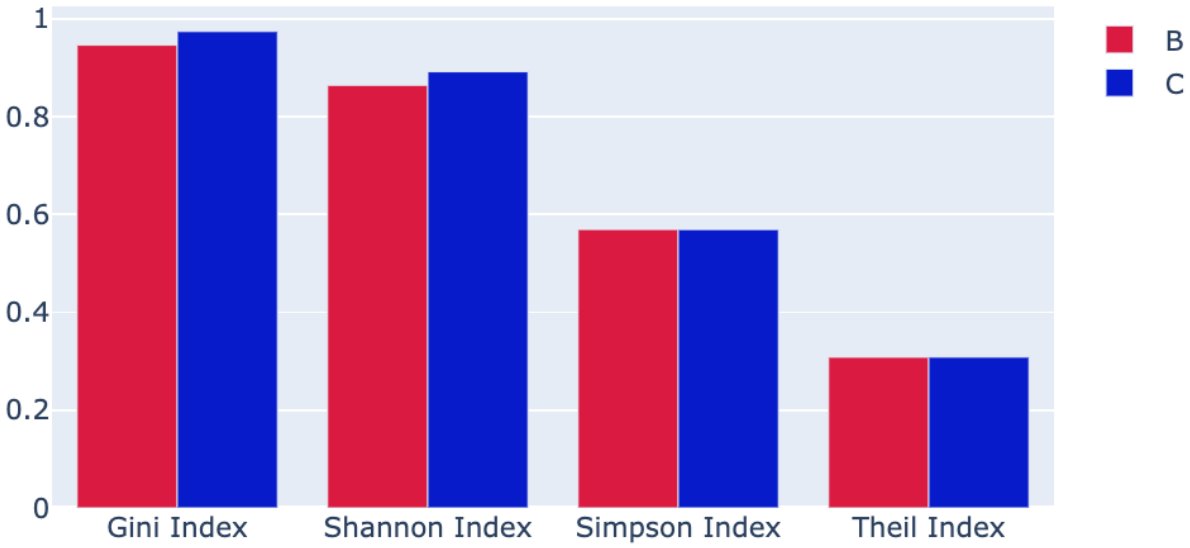

Figura 3.11. Caso B e C: confronto tra indice di Gini, indice di Shannon, indice di Simpson ed indice di Theil

essere influenzati dal numero di classi.

3.2.3 Correlazione tra variabili

Ritornando al caso A ovvero ai dati originali di Compas può risultare utile esplorare i dati alla ricerca di eventuali relazioni tra classi.
Per questo motivo applico entrambe le correlazioni analizzate nel capitolo due.
Ho deciso di applicarle entrambe sullo stesso set di dati in quanto forniscono informazioni differenti. Infatti, il coefficiente di Spearman è calcolato su gradi e quindi descrive relazioni monotoniche mentre quello di Pearson è su valori reali e quindi descrive relazioni lineari.
Anche confrontare tra di loro i valori potrebbe risultare interessante in quanto se si avesse, per esempio, un coefficiente di Spearman maggiore di quello di Pearson vorrebbe dire che vi è una correlazione monotona ma non lineare.
Dato che nei cadi presi ad esame non tutte le variabili sono espresse con valori numerici prima di applicare la funzione di correlazione è necessario modificare il database assegnando valori numerici alle variabili espresse con una sigla differenziando i valori.
Ad esempio per le informazioni relative all'etnia ho operato così:

```
dummy['Ethnic_Code'] = dummy['Ethnic_Code_Text'].apply(m.tran_math)
```

In questo modo è stata creata una nuova colonna colonna speculare alla precedente nella quale la funzione tran_math effettuava un'assegnazione numerica in base al valore di Ethnic_Code_Text:

```
def tran_math(x):
##Ethnic_Code_Text
    if x == 'African-American':
        return 1
    if x == 'Hispanic':
        return 2
    if x == 'Caucasian':
        return 3
    if x == 'Native American':
        return 4
    if x == 'Arabic':
        return 5
    if x == 'Oriental':
        return 6
    if x == 'Other':
        return 7
```

In figura 3.12 ho rappresentato la matrice di correlazione di Pearson mentre in figura 3.13 quella di Spearman

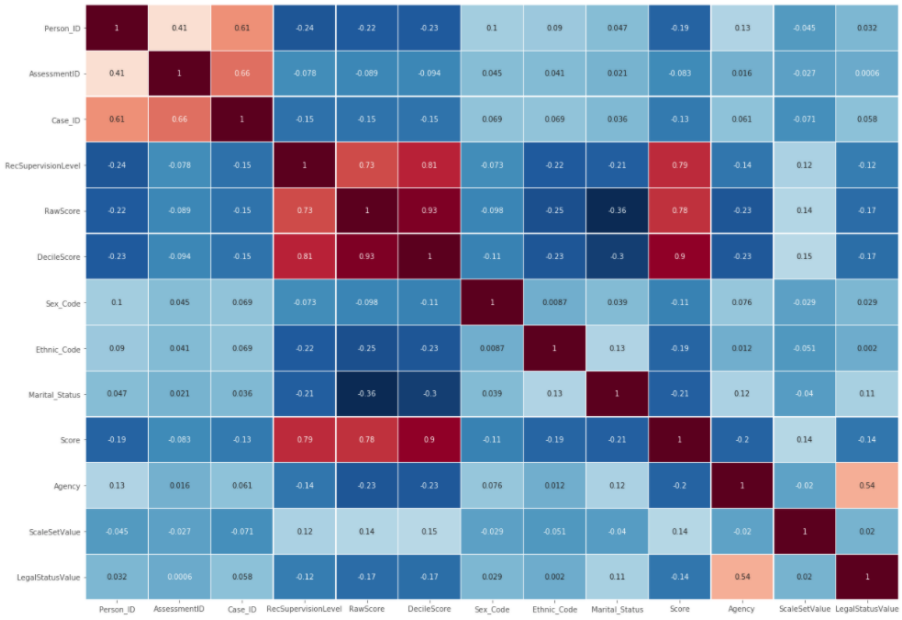

Figura 3.12. Caso A: Matrice di correlazione di Pearson.

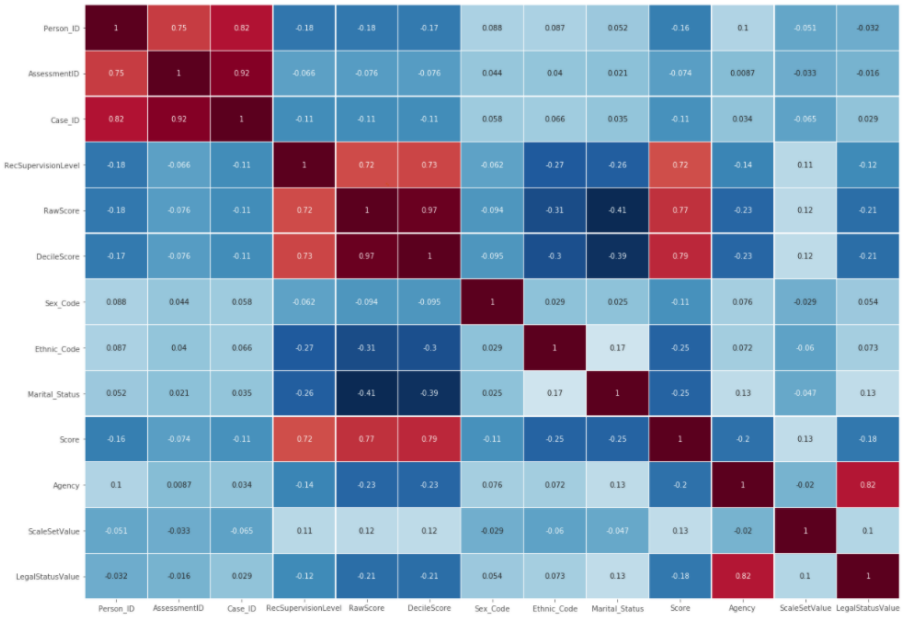

Figura 3.13. Caso A: Matrice di correlazione di Spearman.

Dal confronto delle figure 3.12 e 3.13 notiamo come esista una correlazione (seppur lieve) tra lo score di recidività ed attributi ritenuti sensibili come lo stato civile e l'etnia.
I valori di correlazione sono da interpretare in quanto per alcuni valori presenti sotto forma ti text sono stati convertiti in valori arbitrari solamente per differenziare i casi.

Infine ho applicato il test del chi-quadrato alle variabili Ethnic_Code_Text e ScoreText rilevando i seguenti valori di residui di Pearson sintetizzati in figura 3.14 alla ricerca di correlazioni tra i valori di queste due variabili.
La scala cromatica utilizzata vede i valori negativi rappresentati con colori più scuri (partendo da nero) ed i colori positivi invece di colori più chiari fino al bianco.

Grazie ad una funzione chi2_contingency() è possibile calcolare la statistica chi-quadrata e il valore p.
Il numero di gradi di libertà è espresso utilizzando funzioni e attributi numpy.

I valori dei residui sono standardizzati in quanto si divide per la radice del valore atteso.
Come introdotto nel capitolo 2 nella sezione sul test dell'indipendenzaI il residuo serve a confrontare i conteggi osservati con quelli previsti.
Il segno (positivo o negativo) indica se la frequenza osservata nella cella i è superiore o inferiore al valore misurato nel modello.

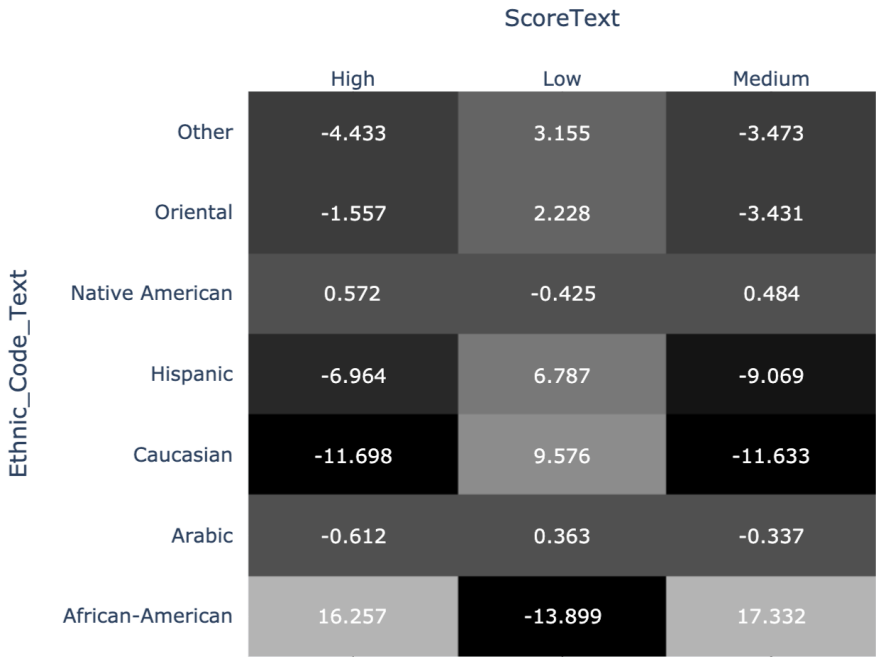

Figura 3.14. Caso A: Test chi-quadrato raffigurante i residui standardizzati di Pearson.

Andando ad osservare i valori i dei residui possiamo identificare i gruppi maggiormente responsabili della dipendenza.
Nel nostro caso pare evidente la classe rappresentante le persone di colore sia maggiormente collegata ad una recidiva high e medium rispetto alle altre classi.
Questo risulta essere coerente con gli studi precedenti [10].
Al fine di poter verificare la correttezza delle conclusioni fatte sul caso A ho applicato gli stessi ragionamenti sul caso B (nel quale avevo modificato nelle frequenze di alcuni attributi tra i quali Ethnic_Code_Text).
Riporto i valori dei residui di Pearson sintetizzandoli in figura 3.15

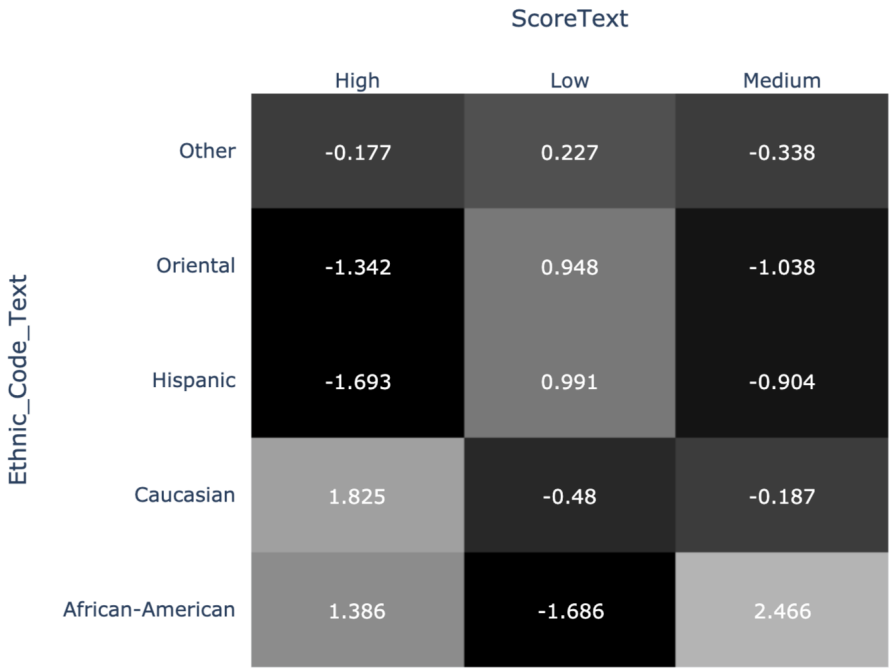

Figura 3.15. Caso B: Test chi-quadrato raffigurante i residui standardizzati di Pearson.

Capitolo 4

Conclusioni

Gli indici sono stati scelti in base a quelli già presenti in letteratura che potevano essere applicati a questo tipo di caratteri.
L'utilizzo di un database già noto ha permesso di focalizzare l'attenzione su alcuni elementi come quello degli attributi sensibili.
A questi attributi è necessario porre una particolare attenzione in quanto potenzialmente discriminatori se applicati in algoritmi decisionali.
La scelta di applicare stessi indici in dataset simili nel quale sono stati bilanciati i cambiamenti per evitare di giungere a conclusioni errate mi ha permesso di notare come alcuni indici fossero influenzati dal numero di classi degli attributi presi in considerazione.
Infatti se l'indice di Gini e l'indice di Shannon sembrerebbero non essere influenzati dal differente numero di classi, l'indice di Simpson e Theil sembrano comportarsi in maniera differente dei due casi nel quale erano stato variato solo quella caratteristiche.
Un'analisi accurata dei training dataset consiste non solo nella pre elaborazione dei dati e nell'applicazione di indici già noti. La corretta comprensione di quest' ultimi permette di prendere in considerazione tutti i fattori e non giungere a conclusioni errate solamente in base a dei valori di indici.
In questo ambito, delicato per via dei suoi campi applicativi ed estremamente attuale, riuscire ad agire in anticipo rispetto al manifestarsi di un problema, permette di evitare l'insorgere di comportamenti discriminatori che, in quanto tali, violano i diritti fondamentali dell'uomo.
Questa analisi apre scenari interessanti per possibili sviluppi futuri legati non solo all'individuazione di sproporzioni ma soprattutto alla possibile mitigazione e soluzione per risolvere questo problema.

Bibliografia

[1] Universal Declaration of Human Rights.
https://www.un.org/en/universal-declaration-human-rights

[2] Artificial Intelligence
http://www.intelligenzaartificiale.it

[3] What's the Difference Between AI and Machine Learning?
https://blogs.oracle.com

[4] Università degli Studi di Torino - Corso di apprendimento automatico. Appunti del corso
Professoressa Rosa Meo e Professor Roberto Esposito. a.a. 2017/2018

[5] Bias-variance dilemma.
https://towardsdatascience.com/bias-variance-dilemma-74e5f1f52b12

[6] Scott Fortmann-Roe - Understanding the Bias-Variance Tradeoff.
http://scott.fortmann-roe.com/docs/BiasVariance.html

[7] Shannon index
http://www.tiem.utk.edu/ gross/bioed/bealsmodules/shannonDI.html

[8] Simpson index
http://www.tiem.utk.edu/ gross/bioed/bealsmodules/simpsonDI.html

[9] A Gentle Introduction to Information Entropy
https://machinelearningmastery.com/what-is-information-entropy/

[10] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. 2016. Machine Bias - ProPublica.
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[11] Python for Data Analysis, Data Wrangling with Pandas, NumPy, and IPython
Wes McKinney - O'REILLY, 2nd Edicion

[12] Jim Frost, Multicollinearity in Regression Analysis

https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

[13] Kent State University, chi-square test of independence
https://libguides.library.kent.edu/SPSS/ChiSquare

[14] Practical statistics
https://www.sagepub.com/sites/default/files/upm-assets/33663_book_item_33663.pdf

[15] Coefficiente di Spearman
https://it.wikipedia.org/wiki/Coefficiente_di_correlazione_per_ranghi_di_Spearman

[16] Coefficiente di Spearman
https://laboratoriostatistica.files.wordpress.com/2014/09/spearman.pdf