Università del Piemonte Orientale

Dipartimento di Scienze e Innovazione Tecnologica

Corso di Laurea in Intelligenza Artificiale e Innovazione Digitale

Relazione per la prova finale

Intelligenza Artificiale e trial clinici: applicazioni del Machine Learning nelle analisi di efficacia

Tutore interno:

Prof. Fabrizio Faggiano

Candidato:

Federico Rivella

Anno Accademico 2023/2024

Indice

1 Introduzione

1.1 Tabacco e sanità pubblica

Il consumo di tabacco rappresenta un fattore di rischio per le malattie cardiovascolari e polmonari, diversi tipi di tumori e alcune condizioni di salute debilitanti (Organization, nd). Secondo l'Organizzazione Mondiale della Sanità (OMS) ogni anno più di 8 milioni di persone muoiono a causa dell'uso di tabacco (World Health Organization, 2020). Anche l'esposizione passiva al fumo è un fattore di rischio per diverse condizioni patologiche ed è causa di circa 1.2 milioni di morti l'anno, di cui circa 65000 sono bambini (World Health Organization, 2020).

In Europa secondo i dati del rapporto ESPAD del 2019, in media il 41% degli studenti ha fumato almeno una volta nella vita, il 20% negli ultimi 30-giorni prima del questionario e il 18% ha fumato a un'età pari o inferiore a 13 anni (ESPAD Group, 2020). In Italia, invece, i risultati dell'indagine 2021/2022 della sorveglianza HBSC evidenziano un aumento della prevalenza dell'abitudine al fumo direttamente proporzionale all'aumentare dell'età in entrambi i sessi e che, a partire dai 13 anni, le ragazze riferiscono di fumare più dei ragazzi, come si può vedere in Fig. 1 :

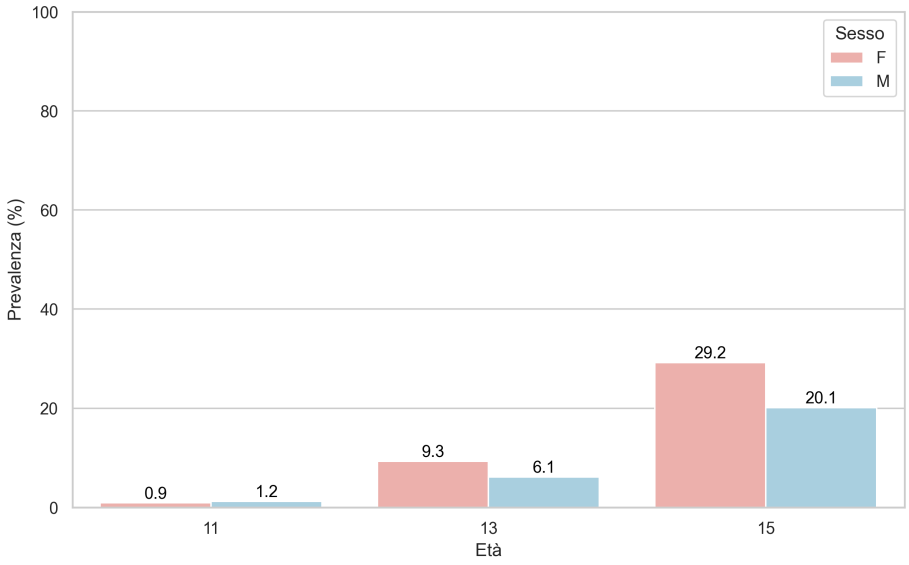

Figura 1: Abitudine al fumo negli ultimi 30gg per sesso ed età

Nella stessa indagine emerge che Lo status socio-economico della famiglia non sembra influenzare l'abitudine al fumo nei ragazzi di 11, 13 e 15 anni che presentano prevalenze sia di fumo di sigaretta molto simili nelle famiglie con FAS

(Family Affluence Scale, scala di agiatezza/ricchezza familiare) basso, medio e alto. A 17 anni, al contrario, risulta più alta la prevalenza di fumo di sigaretta fra i ragazzi appartenenti alle famiglie più benestanti, come si vede in Fig. 2

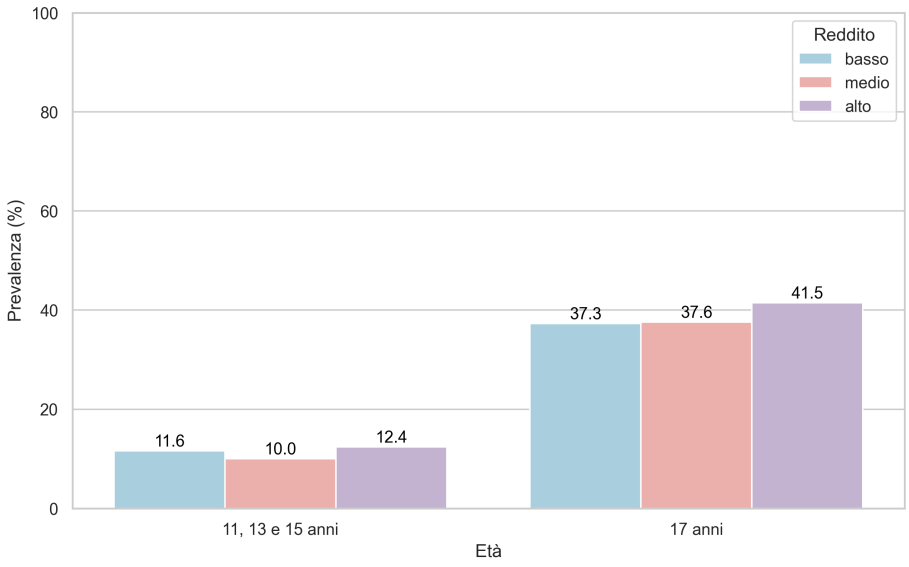

Figura 2: Abitudine al fumo negli ultimi 30gg stratificata per FAS

Le analisi sul trend dal 2002 al 2015 hanno mostrato che in Italia la prevalenza del fumo di tabacco tra gli studenti di età compresa tra 11 e 13 anni è leggermente diminuita, mentre è rimasta invariata se non aumentata quella negli studenti di 15-16 anni (Gorini et al., 2019). Nella stessa analisi si osserva un aumento degli alunni che riportano che i fumatori siano più attraenti e con più amici (Gorini et al., 2019). Contemporaneamente, nell'indagine Global Youth Tobacco Survey (GYTS) del 2014 si è osservato che meno della metà dei partecipanti considera il fumo passivo dannoso (Gorini et al., 2019).

Monitorare i comportamenti relativi al fumo in questa fascia di età è fondamentale per diverse ragioni: infatti, diverse analisi mostrano che almeno l'80% degli adulti fumatori ha iniziato prima dei 20 anni (Gorini et al., 2019; Hu et al., 2020). Inoltre, fumare o anche solo provare durante l'adolescenza si associa a un maggior rischio di continuare anche nell'età adulta. Infine, sembra esserci una relazione inversa tra l'età della prima esposizione al fumo e lo sviluppo di una dipendenza che si protrae anche nell'età adulta (Buchmann et al., 2013; Hu et al., 2020).

1.2 WHO Framework Convention on Tobacco Control (FCTC)

Il 27 febbraio 2005, sotto l'egida dell'OMS, è ufficialmente entrato in vigore nei paesi firmatari il primo trattato internazionale per contrastarne la crescita e la diffusione dell'epidemia di consumo di tabacco: la "Framework Convention on Tobacco Control" (FCTC) (Roemer et al., 2005).

L'idea nacque dall'incontro tra Ruth Roemer e Allyn L. Taylor nel 1993 (Roemer et al., 2005). Quest'ultima, in un precedente articolo, spiegò come l'OMS avesse l'autorità per istituzionalizzare gli sforzi volti al miglioramento della salute pubblica globale. Il contributo di Roemer fu quello di applicare questa idea per mettere in atto un meccanismo regolatorio internazionale per il controllo del tabacco (Roemer et al., 2005). Nel 1995 presentarono diverse proposte all'OMS e l'anno successivo l'Assemblea Mondiale della Sanità (AMS) votò per procedere allo sviluppo del piano. Durante l'AMS nel maggio 2003 venne adottata per consenso degli stati membri il FCTC.

Le misure adottate nella convenzione si possono raggruppare secondo due scopi: ridurre la domanda e ridurre la fornitura di tabacco (World Health Organization, 2003).

Le azioni suggerite per ridurre la richiesta di tabacco si trovano negli articoli 6-14 e prevedono che i paesi (World Health Organization, 2003):

- implementino politiche fiscali e di prezzo sui prodotti del tabacco e proibiscano (o limitino) le vendite e le importazioni di prodotti del tabacco esenti da tasse e dazi da parte di viaggiatori internazionali;
- implementino misure per la protezione dal fumo di tabacco in posti di lavoro al chiuso, trasporti pubblici e altri luoghi pubblici;
- emanino linee guida per testare e regolare i contenuti e le emissioni dei prodotti di tabacco;
- adottino misure tali che gli imballaggi dei prodotti non contengano messaggi promozionali e presentino avvertimenti riguardo i rischi per la salute;
- promuovano campagne educative per informare i cittadini dei rischi legati all'esposizione e al consumo di tabacco e dei benefici della loro cessazione;
- promuovano la cessazione dell'uso di tabacco e il trattamento della dipendenza attraverso linee guida e programmi integrati che coinvolgono istituzioni educative, strutture sanitarie e ambienti di lavoro

Le misure finalizzate a ridurre la fornitura di tabacco, invece, si trovano negli articoli 15-17 e prevedono che i paesi (World Health Organization, 2003):

- adottino misure di controllo e tracciamento dei prodotti per eliminarne il traffico illegale;
- vietino la vendita di prodotti ai minori;

- impediscano la vendita di pezzi unitari che renderebbero i prodotti più accessibili economicamente;
- incentivino alternative vantaggiose per i lavoratori dell'industria del tabacco compresi coltivatori e rivenditori

Studiare l'efficacia dell'implementazione di queste strategie non è semplice perché i risultati potrebbero essere modificati dall'effetto di fattori come politiche preesistenti (volte allo stesso fine), status socio-economico della nazione e capacità dello stato nel mettere in atto piani più efficaci anche in assenza di FCTC (Hiilamo and Glantz, 2022). Tuttavia, recentemente Paraje et al. (2024) hanno stimato, attraverso analisi statistiche, che dopo la ratifica di 170 stati a livello globale (escludendo la Cina), la prevalenza di fumatori sotto i 25 anni è diminuita, mentre il rapporto di cessazione nei fumatori tra 45 e 49 anni è aumentato. Anche nei 50 paesi europei che hanno aderito al trattato la prevalenza è in calo; con un tasso che dipende fortemente dalle condizioni socio-economiche della popolazione, minacciando la possibilità della nascita di nuove disuguaglianze (Willemsen et al., 2022).

I ricercatori Joossens and Raw (2006) hanno elaborato un sistema per misurare il livello di aderenza alle 6 misure definite dalla Banca Mondiale:

- incremento dei prezzi attraverso la tassazione di prodotti del tabacco;
- divieto di fumare in luoghi pubblici e nei locali lavorativi;
- informare i consumatori;
- divieto di pubblicità e promozioni per tutti i prodotti del tabacco;
- messaggi di avvertimento per i rischi della salute su tutte le confezioni;
- accesso ai trattamenti necessari per contrastare la dipendenza.

Attraverso la compilazione di un questionario riguardante le politiche adottate si ottiene un punteggio all'interno di una scala ("Tobacco Control Scale" o "TCS") che ha valore massimo 100 (Joossens and Raw, 2006). Feliu et al. (2019) hanno osservato che, nei 27 paesi dell'Unione Europea analizzati, il valore del punteggio ottenuto è correlato direttamente al tasso di cessazione e alla diminuzione della prevalenza di fumatori.

Tra le possibili misure, in diverse analisi si osserva che la tassazione è la misura più efficace per ridurre la prevalenza, soprattutto nelle fasce di età più giovani e di reddito inferiore (Paraje et al., 2024; Gravely et al., 2017).

1.3 Interventi nelle scuole

Come discusso nel paragrafo 1.1, la maggior parte dei fumatori inizia durante l'adolescenza, rendendo le scuole un contesto ideale per le campagne di prevenzione, poiché permettono di raggiungere un gran numero di individui in questa

fascia di età (Bafunno et al., 2019). Per questo motivo, gli interventi di prevenzione in ambito scolastico hanno una lunga storia e hanno portato allo sviluppo di diverse strategie (Morison et al., 1964; Andrus et al., 1964; Flay, 1985).

I tre principali modelli sono (Lantz et al., 2000):

- deficit di informazione o razionale;
- educazione affettiva;
- resistenza all'influenza sociale.

Il modello razionale si fonda sul presupposto che i giovani siano disinformati e pertanto debbano essere istruiti sui rischi di salute e sociali causati dal fumo (Lantz et al., 2000). Questo modello è stato utilizzato in diversi programmi fino ai primi anni '70, poi progressivamente abbandonati perché meno efficace di altri (Lantz et al., 2000).

I programmi basati sul secondo modello, invece, si concentrano maggiormente sul miglioramento dell'autostima e dell'immagine di sé, su tecniche di gestione dello stress, abilità decisionali, chiarificazione dei valori e definizione degli obiettivi (Lantz et al., 2000). Anche in questo caso l'impatto di questi programmi è stato debole o insignificante (Lantz et al., 2000).

Nell'ultimo caso si enfatizza l'importanza dell'ambiente come fattore critico nell'uso del tabacco, come ad esempio il comportamento dei pari e l'ambiente familiare (Lantz et al., 2000). I programmi pertanto si concentrano nello sviluppare tecniche per riconoscere e resistere alle influenze negative, abilità comunicative e assertività (Lantz et al., 2000).

Negli stati europei sono stati effettuati diversi interventi di prevenzione. Tra questi alcuni dei più importanti sono stati (Bafunno et al., 2019):

- "Luoghi di Prevenzione" (Italia): molto efficace grazie all'approccio basato sulla resistenza alle pressioni dei pari;
- "Eigenstandigwerden 5-6" (Germania): simile al precedente;
- "Smoke Free Sports - SFS" (Regno Unito): basato su molteplici teorie cognitive e implementato nella scuola primaria attraverso l'attività sportiva
- "Education Against Tobacco" (Germania): suddiviso in più moduli, uno dei quali ha visto l'uso del photoaging che è risultato efficace soprattutto tra le femmine e gli studenti con basso livello educativo o background migratorio
- "A Smoking Prevention Interactive Experience" (Romania): multimediale attraverso l'uso di video, animazioni e immagini accoppiate ad attività interattive

1.3.1 "Eueropean Drug Abuse Prevention" (EU-Dap)

EU-Dap è uno studio multicentrico con 9 partner in 7 stati europei finanziato dalla Commissione Europea. Lo scopo dello studio è sia elaborare un programma un programma preventivo per l'uso di droghe, tabacco e alcol, sia misurarne la sua efficacia attraverso uno studio sperimentale (Faggiano et al., 2006).

Il programma, chiamato Unplugged, ha lo scopo di ritardare l'inizio dell'uso di queste sostanze. Il modello "Life-skills" su cui è basato Unplugged prevede l'acquisizione di informazioni riguardanti fumo, tabacco e alcol affianco allo sviluppo di capacità di problem solving, pensiero critico, abilità comunicative, decisionali e di gestione delle emozioni (Faggiano et al., 2006).

Unplagged è stato disegnato in 3 curricula:

- base: solo intervento base;
- con pari: intervento base e coinvolgimento di pari;
- con genitori: intervento base e coinvolgimento di genitori.

L'intervento base è costituito da 3 parti per un totale di 12 unità. Il primo modulo è informativo e dopo un'unità introduttiva, i ragazzi imparano i fattori che li influenzano a iniziare l'uso di sostanze e i loro rischi per la salute (Faggiano et al., 2006). Il secondo, invece, è dedicato all'apprendimento di abilità interpersonali come la distinzione tra comunicazione verbale e non verbale, la capacità di esprimere le proprie emozioni, di valutare quanto il proprio comportamento è influenzato dal gruppo e di valutare criticamente le informazioni (Faggiano et al., 2006). Il terzo, infine, ha lo scopo di sviluppare le abilità intrapersonali come quelle di problem solving, decision making e stabilimento di obiettivi (Faggiano et al., 2006).

1.4 Intelligenza artificiale e trial clinici

Per valutare l'efficacia di un trattamento in medicina uno dei disegni di studio più utilizzati è il trial clinico, da molti considerato un "gold-standard" (Hansson, 2014). Oggi i moderni sistemi di intelligenza artificiale stanno portando numerose innovazioni nel modo in cui questi studi potranno essere condotti in futuro (Harrer et al., 2019).

La storia dell'utilizzo dell'intelligenza artificiale in ambito medico inizia negli anni '70 del secolo scorso con i primi sistemi esperti, come MYCIN (Shortliffe, 1977) e INTERNIST (Miller et al., 1985). Programmare questi sistemi era un processo complesso e lento che richiedeva l'aiuto di esperti del dominio per mappare la conoscenza in una serie di regole logiche. Quest'ultime risultavano spesso troppo rigide e difficili da aggiornare successivamente (McCauley and Ala, 1992). Con l'aumento dei dati a disposizione e la disponibilità di hardware per computarli efficientemente, i sistemi esperti hanno visto un progressivo abbandono in favore di algoritmi di Machine Learning (ML) e Deep Learning (DL) in grado di apprendere autonomamente le regole dai dati (Wang et al., 2019; Harrer et al., 2019).

Le loro applicazioni in ambito biomedico sono numerose e l'approvazione di molti dispositivi medici che utilizzano queste tecniche è in aumento (Muehlematter et al., 2021). Tra i campi in cui si stanno studiando ci sono la radiologia e la radiomica (Meng et al., 2023), le scienze omiche (Li et al., 2022), le procedure di meta-analisi (Hughes et al., 2020), la scoperta di nuovi farmaci (Sarkar et al., 2023) e la gestione delle varie fasi dei trial clinici (Askin et al., 2023).

Un trial clinico è generalmente composto da 5 diverse fasi: pre-clinica, disegno di studio, reclutamento, svolgimento e analisi. Nello sviluppo di nuovi farmaci l'intero processo dura in media 10-15 anni con costi elevati e i problemi che possono insorgere in ognuna di esse può aumentarli ulteriormente (Harrer et al., 2019).

In fase pre-clinica si stanno studiando algoritmi di ML e di DL per identificare nuove molecole target, possibili candidati farmaci e testarne la tossicità in silico (Zhavoronkov et al., 2020; Basile et al., 2019). Nella fase del disegno di studio alcuni gruppi di ricerca hanno utilizzato algoritmi di ML per predire l'outcome e il dropout. Questi approcci potrebbero permettere di accorciare la durata del trial e fornire maggior supporto ai partecipanti intenti ad abbandonare lo studio (Askin et al., 2023). Inoltre, alcuni ricercatori stanno studiando tecniche per predire quali trial sono più propensi a fallire nelle varie fasi e identificarne i possibili fattori causali (Feijoo et al., 2020).

Per quanto riguarda il disegno di studio è in fase di ipotesi la possibilità di usare reti neurali sufficientemente addestrate per predire la progressione della malattia nel braccio di controllo che, di conseguenza, potrebbe essere sostituito con uno virtuale (Lee and Lee, 2020).

Il processo di reclutamento, invece, potrebbe essere ottimizzato automatizzando i processi per determinare l'eleggibilità di un paziente con criteri di inclusione che considerino dati -omici, immagini mediche, dati di laboratorio e demografici (Askin et al., 2023).

Attraverso nuove tecnologie come le pillole "intelligenti" e l'uso di smartphone per filmarsi, sta diventando possibile monitorare la compliance al trattamento da parte dei pazienti in fase di svolgimento del trial. Addestrando reti neurali e algoritmi di machine learning con i dati ottenuti si potranno fare predizioni sulla costanza da parte dei pazienti di assumere i farmaci seguendo lo schema terapeutico oggetto di studio (Koesmahargyo et al., 2020; Martani et al., 2020; Story et al., 2019).

Infine, in fase di analisi, è possibile esaminare gli effetti eterogenei all'interno di specifici sottogruppi della popolazione e prevedere l'impatto del trattamento su ciascun individuo. (Goldstein and Rigdon, 2019). Inoltre le tecniche di ML e text mining consentono anche di automatizzare l'estrazione dei dati e l'imputazione di quelli mancanti (Gates et al., 2021; Zame et al., 2020).

1.5 Algoritmi di intelligenza artificiale

Nel paragrafo 1.4 sono stati introdotti diversi concetti relativi all'intelligenza artificiale (IA), con particolare attenzione al contesto dei trial clinici. Dopo una breve panoramica storica, i prossimi paragrafi approfondiranno diversi approc-

ci che vengono utilizzati, come l'IA simbolica e quella statistica, insieme alle tecniche più rilevanti per questa trattazione.

1.5.1 Storia dell'intelligenza artificiale

Nel corso della sua storia, l'intelligenza artificiale ha alimentato un ampio dibattito sulla sua definizione. Alcuni studiosi sostenevano che bisognasse considerare l'intelligenza in termini di fedeltà alla prestazione umana, in contrapposizione con coloro che preferivano una definizione formale di razionalità, sintetizzabile in "fare la cosa giusta". Inoltre, si svilupparono posizioni divergenti anche sulla natura dell'intelligenza: mentre alcune correnti la consideravano si focalizzavano sui processi del pensiero, altre preferivano concentrarsi sul comportamento intelligente (Russell and Norvig, 2021).

In questo panorama, la storia di questa disciplina i cui punti salienti sono mostrati in Fig. 3, ha visto un susseguirsi di diverse visioni.

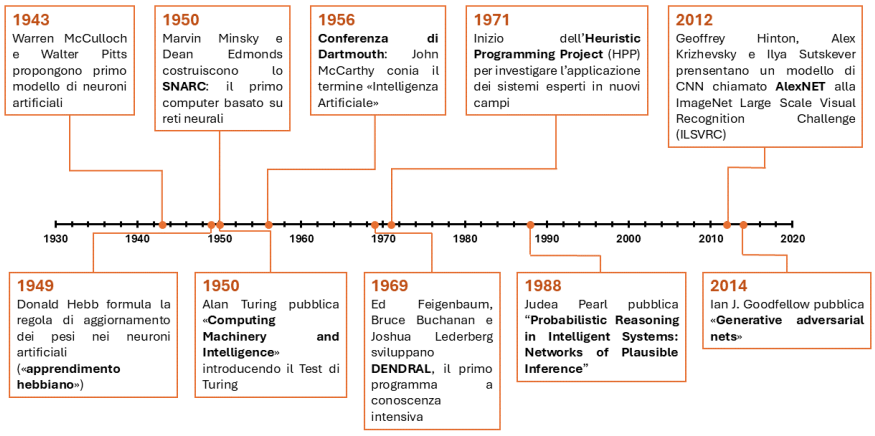

Figura 3: Linea temporale con alcune delle principali innovazioni nell'ambito dell'IA (Goodfellow et al., 2014; Russell and Norvig, 2021)

Il termine intelligenza artificiale venne coniato nel 1956 durante la conferenza di Dartmouth, un workshop di due mesi durante l'estate ideato da John McCarthy e che riunì 10 ricercatori dell'epoca (Russell and Norvig, 2021). Tuttavia, il primo lavoro riconosciuto come intelligenza artificiale risale al 1943, quando i ricercatori Warren McCulloch e Walter Pitts unirono conoscenze provenienti dalla fisiologia, dalla loica proposizionale e dalla teoria della computazione per modellare il primo neurone artificiale. Sei anni dopo, Donald Hebb formulò una regola per l'aggiornamento dei pesi delle connessioni tra tali neuroni (Russell and Norvig, 2021).

Tra gli anni '50 e gli anni '60 troviamo i primi successi della nuova materia che andava sviluppandosi portando entusiasmo e grandi aspettative. Un esempio è il General Problem Solver (GPS) di Newell e Simons, che risolveva alcuni problemi scomponendoli in sotto-obiettivi e analizzandone le possibili azioni in un processo simile a quello di un umano. Pochi anni dopo Herbert Gelernter scrisse il Geometry Theorem Prover, un programma in grado di dimostrare alcuni problemi matematici. Tuttavia, il lavoro più influente per gli anni a venire fu quello di Arthur Samuel, il cui limite era il tempo di computazione: usando il metodo che oggi definiamo "apprendimento per rinforzo" creò un programma in grado di imparare il gioco della dama a livello di un buon dilettante(Russell and Norvig, 2021).

I primi entusiasmi cominciarono a smorzarsi dopo pochi anni. La scelta di non analizzare le attività ma mimare il modo in cui gli umani le svolgevano portò i primi sistemi a fallire non appena i problemi divennero più complessi. Per risolvere questi problemi, tra la fine degli anni '60 e la metà degli anni '80, i ricercatori iniziarono a cambiare punto di vista. I metodi deboli, basati su una successione di passi elementari e generali, vennero sostituiti dai sistemi esperti. Questi sistemi si fondavano su una conoscenza specifica del dominio e utilizzavano regole che permettevano un ragionamento in grado di risolvere problemi più complessi in ambiti più ristretti(Russell and Norvig, 2021).

Negli anni successivi avvenne un altro cambio di paradigma basato sulla teoria della probabilità, l'apprendimento automatico e i risultati sperimentali. In questo periodo venne reinventato l'algoritmo della retropropagazione da parte di diversi gruppi di studio. Gli approcci basati sui modelli nascosti di Markov dominarono il campo grazie alla solida trattazione matematica alla base e le ottime prestazioni. Infine, il formalismo delle reti bayesiane permise una modellizzazione della conoscenza incerta e del ragionamento probabilistico (Russell and Norvig, 2021).

Per concludere questa breve introduzione storica, nel nuovo millennio lo sviluppo del World Wide Web e la maggiore potenza di calcolo dei computer hanno permesso la creazione di set di dati di grandi dimensioni: i Big Data. Per trarre vantaggio da questa grande mole di dati sono stati sviluppati numerosi algoritmi di apprendimento i cui successi portarono l'IA a recuperare appetibilità commerciale. Oggi i sistemi di deep learning hanno permesso di raggiungere (e in alcuni casi superare) le prestazioni umane in molti compiti di visione, riconoscimento vocale, diagnosi mediche, traduzione automatica e nei giochi (Russell and Norvig, 2021).

1.5.2 Approcci nell'intelligenza artificiale

Nel paragrafo 1.5.1 sono stati riassunti alcuni momenti chiave nella storia dell'IA, che hanno portato allo sviluppo di vari approcci come illustrato in Fig. 4. In generale essi possono essere classificati in due categorie: quelli simbolici e quelli basati sull'apprendimento statistico.

Riferendosi agli approcci simbolici, John Haugeland coniò l'acronimo GOFAI (Good Old-Fashioned Artificial Intelligence) Haugeland (1989). In questa cate-

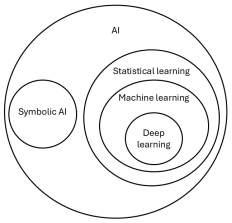

Figura 4: Approcci nell'intelligenza artificiale

goria rientrano i programmi in cui la conoscenza dell'esperto del settore viene codificata in un sistema formale per mimare il ragionamento umano nel risolvere i problemi. Le principali limitazioni dei sistemi simbolici sono nei contesti in cui la conoscenza è incompleta o le regole sono soggette a cambiamenti, risultando difficili da aggiornare (Boucher, 2019). Ad esempio, nel campo medico, sistemi basati su approcci simbolici per la gestione delle linee guida come GLARE (GuideLine Acquisition Representation and Execution), possono facilitarne l'uso soprattutto in casi complessi come, ad esempio, quello delle comorbidità. In queste situazioni possono verificarsi interazioni tra più linee guida che, se non venissero individuate e gestite correttamente, potrebbero portare a gravi complicazioni (Anselma et al., 2011; Terenziani et al., 2001; Piovesan et al., 2018). Allo stesso tempo, tuttavia, questi sistemi non sono in grado di replicare l'esperienza e l'intuito di un medico, fattori spesso importanti nel processo decisionale (Boucher, 2019).

I sistemi di apprendimento statistico, invece, hanno un approccio "data-driven". L'apprendimento può essere di due tipi: supervisionato e non supervisionato. In quello supervisionato la macchina impara a predire un output da un insieme di input. In quello non supervisionato, invece, la macchina apprende le relazioni e i pattern presenti nei dati, senza che le venga fornito uno specifico output (Boucher, 2019; James et al., 2021).

In conclusione del paragrafo, si evidenzia la differenza tra machine learning (ML) e deep learning (DL). Nel machine learning tradizionale, le feature di input sono spesso rappresentazioni dei dati grezzi, ottenute tramite selezione ed elaborazione manuale da parte di esperti del settore (feature engineering). Al contrario, nel deep learning, le reti neurali profonde sono in grado di apprendere automaticamente nuove rappresentazioni dell'input su più livelli, mantenendo solo le informazioni più rilevanti per il compito da svolgere (Goodfellow et al., 2016; LeCun et al., 2015). Un esempio per chiarire questa differenza è la classificazione di una lesione (ad esempio benigna o non benigna) nelle immagini mediche. Con il machine learning tradizionale, il radiologo deve estrarre manualmente informazioni dall'immagine come la circolarità, l'area della regione di interesse e il contrasto. L'algoritmo, utilizza queste caratteristiche per apprendere come distinguere le diverse classi. Al contrario, nel deep learning, una Rete Neurale Convoluzionale (CNN) può analizzare direttamente i valori

dei pixel dell'immagine e effettuare la classificazione senza necessità di estrarre manualmente le caratteristiche (Suzuki, 2017).

1.5.3 Algoritmi di apprendimento supervisionato

Nel paragrafo 1.5.2 è stata introdotta la differenza tra apprendimento supervisionato e non supervisionato. Nelle prossime pagine verranno introdotti gli algoritmi di machine learning più utilizzati per entrambe le tipologie di apprendimento.

Per l'apprendimento supervisionato gli algoritmi principali sono: regressione logistica, regressione lineare, regressione polinomiale, metodi basati su alberi, Support Vector Machine (SVM) e K-Nearest Neighbors (KNN) (James et al., 2021). Ci sono due tipologie di problemi che possono essere risolti tramite l'utilizzo di questi algoritmi: classificazione e regressione. In entrambi i casi è noto un outcome che si vuole predire ma mentre nella classificazione è qualitativo nella regressione è quantitativo. Un esempio di un problema di regressione può essere la previsione del costo delle cure mediche di un soggetto (Taloba et al., 2022). Un esempio di classificazione, invece, è la predizione degli outcome dell'ipertensione (Chang et al., 2019).

Regressione logistica e regressione lineare La regressione logistica è un metodo utilizzato per la classificazione di variabile binarie, anche se il risultato ottenuto è una probabilità calcolata con la Equazione 1:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}} \tag{1}$$

dove Y è la variabile dipendente, β_i sono i coefficienti e i valori delle variabili indipendenti sono x_i (Kurt et al., 2008).

La regressione lineare si utilizza per il problemi di regressione. In generale la relazione vera che intercorre tra la variabile dipendente e quelle indipendenti si può scrivere come nella Equazione 2:

$$Y = f(X) + \varepsilon \tag{2}$$

dove Y è la variabile dipendente, X è l'insieme delle variabili indipendenti, f una funzione sconosciuta e ε è un termine di errore con media zero. Nella regressione lineare la funzione f viene approssimata in una retta come nella Equazione 3:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3}$$

dove β_0 , chiamato anche intercetta, è il valore che Y assume quando $X = 0$, mentre β_1 è il coefficiente angolare della retta (James et al., 2021).

In alcuni casi, la regressione lineare non è in grado di predire efficacemente la relazione tra la variabile dipendente e quelle indipendenti poiché non sempre è lineare. Aumentare il grado del polinomio permette di ottenere curve più complesse e la funzione f viene approssimata come Nella Equazione 4 (James et al., 2021).

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_d X^d + \varepsilon \tag{4}$$

Algoritmi basati su alberi Gli algoritmi basati su alberi possono essere utilizzati sia per i problemi di regressione sia per quelli di classificazione anche se in seguito per semplicità si fa riferimento solo alla seconda tipologia (James et al., 2021).

L'albero di decisione è l'algoritmo più semplice ed è composto da nodi e rami che rappresentano il processo decisionale. Grazie a questa struttura gerarchica, gli alberi di decisione sono facilmente interpretabili: i nodi che rappresentano le decisioni da prendere, i rami le possibili decisioni e le foglie il risultato finale del processo. Il nodo viene costruito in modo da suddividere efficacemente un insieme di dati, originariamente eterogenei per quanto riguarda le classi, in due insiemi che sono il più possibile omogenei rispetto alle classi stesse. Generalmente le misure più utilizzate per fare questa separazione sono l'indice di Gini e l'entropia (Kingsford and Salzberg, 2008).

Il principale problema degli alberi di decisione è che piccoli cambiamenti nei dati possono portare ad alberi molto differenti, variazioni dell'accuratezza e over-fitting. Per migliorare la stabilità di questi algoritmi, si possono utilizzare metodi ensemble che combinano le previsioni di più classificatori. In questo modo, gli effetti di overfitting dei singoli classificatori sono bilanciati tra loro, aumentando l'accuratezza predittiva del modello complessivo (Zimmermann, 2008). Due metodi di ensamble sono: bagging e boosting.

Nelle tecniche di bagging come il random forest si utilizzano più alberi. Gli alberi sono addestrati su campioni indipendenti e identicamente distribuiti estratti con reimmissione dal dataset iniziale. Inoltre, per costruire i nodi non vengono prese in considerazioni tutte le variabili indipendenti ma solo un loro sottoinsieme estratto casualmente. La predizione più finale è il valore più frequente tra quelle dei singoli alberi (Breiman, 2001).

Come nel bagging, anche nel boosting vengono combinati più alberi. Per semplicità il processo verrà spiegato applicandolo ad un problema di regressione, ma è possibile usare questo metodo anche con la classificazione. In questo caso un primo modello (albero) viene addestrato per fare previsioni nei valori della variabile dipendente. Un nuovo albero, invece, si addestra sui residui del modello precedente a cui verrà aggiunto in modo da correggerne gli errori. Aggiungendo gradualmente nuovi alberi più piccoli, il modello finale migliora le previsioni nelle aree dove il modello iniziale era meno preciso (James et al., 2021).

Support Vector Machine (SVM) Il Support Vector Machine (SVM) è un metodo di apprendimento automatico nato originariamente per i problemi di classificazione binaria e successivamente esteso per risolvere anche quelli di regressione (Support Vector Regressor). L'obiettivo di questo algoritmo è individuare l'iperpiano che separa meglio le due classi. In molti casi non è possibile trovarlo direttamente quindi si utilizza una tecnica matematica chiamata kernel per proiettare i dati in uno spazio di dimensioni superiori. L'utilizzo dei kernel consente di trovare un piano in cui i dati diventino separabili, consentendo di trovare un iperpiano che separi le due classi. Tuttavia, attraverso questa trasformazione, aumenta il rischio di over-fitting. Per minimizzare questo pro-

blema l'algoritmo sceglie l'iperpiano che massimizza il margine, ossia la distanza geometrica tra l'iperpiano stesso e i punti più vicini delle due classi (chiamati Support Vectors) (Boswell, 2002).

K-Nearest Neighbors (KNN) Il K-Nearest Neighbors (KNN) è un algoritmo non parametrico utilizzato per problemi di classificazione e regressione. Una volta definita una funzione di distanza viene calcolata la distanza tra il punto che si vuole classificare e tutti gli altri punti. La classe di appartenenza predetta sarà la classe più rappresentata nei K punti più vicini. Rispetto ai metodi precedenti questo algoritmo è definito "lazy" non costruisce un modello generale attraverso una fase di training ma prende le decisioni utilizzando tutti i dati (Kataria and Singh, 2013; Song et al., 2017).

1.5.4 Algoritmi di apprendimento non supervisionato

Clustering Uno dei problemi principali nell'apprendimento non supervisionato è il clustering. L'obiettivo è suddividere i dati in sottoinsiemi non sovrapposti chiamati cluster (James et al., 2021). Inoltre, i cluster dovrebbero avere quanto più possibile le seguenti caratteristiche (Xu and Tian, 2015):

- le istanze all'interno dello stesso cluster devono essere il più simili possibile;
- le istanze in cluster diversi devono essere il più diverse possibile;
- la misurazione della similarità e della dissimilarità deve essere chiara e avere un significato pratico.

Ci sono diverse strategie per trovare i cluster e in ognuna di esse si sono sviluppati diversi algoritmi.

Nelle strategie basate sul partizionamento, ad esempio, due algoritmi molto utilizzati sono K-means e una sua evoluzione chiamata K-medoids. In K-means si genera una sequenza randomica di K punti nel piano chiamati centroidi. Ogni punto nel set dei dati viene assegnato al cluster corrispondente al centroide più vicino (secondo una distanza definita all'inizio). A questo punto, su ogni cluster si calcola il nuovo centroide come punto medio di tutti i punti del cluster e si riassegnano i dati nei nuovi cluster. Il processo termina quando l'algoritmo converge, ossia quando i centroidi non cambiano significativamente tra un'iterazione e l'altra (Xu and Tian, 2015).

In K-medoids il processo di costruzione dei cluster è simile ma anziché utilizzare i centroidi usa i medoid che sono punti presenti nel set di dati. Per la selezione iniziale dei medoid, viene calcolata la distanza di tutte le coppie di punti possibili. Successivamente, per ogni punto, si sommano tutte le distanze tra esso e gli altri punti. Questo valore prende il nome di dissimilarità totale. I punti si ordinano in ordine crescente secondo il loro valore di dissimilarità totale e si selezionano i primi K come medoid iniziali. Analogamente a K-means ogni punto viene assegnato al cluster del medoid più vicino e si aggiornano i medoid. I nuovi medoid sono calcolati come i punti che minimizzano la distanza totale

rispetto a tutti gli altri negli stessi cluster. I punti vengono riassegnati ai nuovi cluster e il processo itera fino a convergenza (Park and Jun, 2009).

Negli algoritmi di clustering gerarchico, invece, si considera inizialmente ogni punto come un cluster. Il processo procede in maniera iterativa, unendo i cluster più vicini a ogni passo. L'algoritmo termina quando si ottiene un unico cluster che contiene tutti i punti (Xu and Tian, 2015). Esempi di algoritmi che usano questo approccio sono BIRCH (Zhang et al., 1996), CURE (Guha et al., 1998) , ROCK (Guha et al., 2000) e Chameleon (Karypis et al., 1999).

Ci sono molti altri approcci per calcolare i cluster che non verranno approfonditi. Alcuni utilizzano la densità dei punti nello spazio (zone più dense di punti formano un cluster) come l'algoritmo DBSCAN (Ester et al., 1996), altri la distribuzione (dati che provengono dalla stessa distribuzione appartengono allo stesso cluster) (Xu et al., 1998), fino a quelli più moderni basati ad esempio sulla densità e la distanza (Rodriguez and Laio, 2014). Infine, sono state anche proposte delle strategie basate sulla logica fuzzy in cui ogni punto a un grado di appartenenza a ogni cluster nell'intervallo continuo [0, 1] (Dunn, 1973).

Riduzione della dimensionalità Grazie alle tecnologie moderne, si devono spesso gestire dati con molte variabili in input come, ad esempio, quelli generati dai microarray nelle analisi genomiche (Sorzano et al., 2014). Quando questo accade, può sorgere un problema noto in letteratura come "curse of dimensionality" e che si può manifestare in vari modi: ad esempio tramite sparsità dei dati, multicollinearità, test multipli e overfitting (Altman and Krzywinski, 2018).

La sparsità dei dati indica le regioni di spazio che non vengono occupate dai dati. Per tornare al caso dei microarray, si pensi a una situazione con 5 polimorfismi a singolo nucleotide (SNPs) dove l'allele mutato ha una prevalenza del 10%. Considerando solamente una variante, su un campione di 1000 persone allora ci si attende di trovare 100 persone che portano l'allele mutato. Tuttavia, se considerassimo tutte e 5 le varianti per studiare gli individui che presentano tutti gli alleli mutati, allora è molto probabile che su un campione di 1000 persone non si trovi nessuno e quindi il campione non è più sufficiente a rappresentare la popolazione (Altman and Krzywinski, 2018).

La multicollinearità, invece, nasce quando si hanno più dimensioni che osservazioni allora una delle variabili sarà necessariamente una combinazione lineare delle precedenti. Il problema è che risulterà più complicato interpretare i risultati. Passando dalla genetica alla biochimica, si può pensare a una situazione in cui si hanno 3 metaboliti che permettono di determinare la concentrazione di un ormone. Aggiungendo un 4 metabolita che è una combinazione lineare dei precedenti, qualunque combinazione di 3 metaboliti dei 4 possibili consentirà di predire la concentrazione dell'ormone, tuttavia non è possibile determinare l'impatto che ognuno di essi ha (Altman and Krzywinski, 2018).

Per spiegare il problema dei test multipli si può utilizzare nuovamente l'esempio degli SNPs. Per testare l'associazione di un polimorfismo a un tratto fenotipico è necessario effettuare un test. Supponendo di testare l'associazione di un tratto con 100 SNPs e di tenere il livello di significatività di 0.05, allora

si otterrebbero circa 5 falsi positivi. Per questa ragione è necessario fare delle correzioni che complicano l'analisi (Altman and Krzywinski, 2018).

Per gestire questo problema è necessario ridurre il numero di variabili. Per farlo è possibile utilizzare due metodi differenti: la feature selection e la riduzione della dimensionalità. Nel primo caso si selezionano solo le variabili (feature) di interesse mentre nel secondo si crea un gruppo di variabili dimensioni minori all'originale e che riesca a rappresentarle quanto meglio (Sorzano et al., 2014).

Una delle tecniche di riduzione della dimensionalità più utilizzate è l'analisi delle componenti principali (PCA). Ogni componente principale è una combinazione lineare di tutte le variabili originali, progettata per massimizzare la varianza spiegata. Selezionando un numero sufficiente di componenti principali, inferiore al numero delle variabili iniziali, è possibile approssimare i dati originali (Greenacre et al., 2022).

Oltre alla PCA, esistono algoritmi come Laplacian Eigenmap, Stochastic Neighbor Embedding (t-SNE) e Uniform Manifold Approximation and Projection (UMAP) che si concentrano meno sul mantenere la struttura globale e più sul preservare le piccole distanze (Njue and Franklin, 2020).

2 Obiettivo

L'obiettivo di questa tesi è identificare l'eventuale presenza di effetti eterogenei del trattamento Unplugged nello studio EU-Dap descritto nel paragrafo 1.3.1. Per farlo si utilizza una strategia basata sull'apprendimento non supervisionato, come spiegato nel capitolo 3. Le analisi statistiche hanno l'obiettivo di individuare le caratteristiche dei soggetti che differiscono tra i vari gruppi individuati, con l'intento di spiegare le differenze di efficacia del trattamento. Con questo sistema sarà possibile valutare anche se tra i gruppi ci sono differenze nella prevalenza di soggetti che hanno ridotto o mantenuto i livelli di fumo costanti nel braccio di controllo.

3 Materiali e metodi

3.1 Selezione delle features

La prima fase è stata la preparazione dei dati estratti dalla base di dati. Tramite un processo di feature selection sono state selezionate solo le risposte (date al baseline) alle domande del questionario che riguardavano il fumo.

E' stato scelto di considerare un unico braccio di intervento (indipendentemente dal curriculum) e uno di controllo: quindi i soggetti appartenenti a uno qualsiasi dei tre curricula descritti nel paragrafo 1.3.1 sono stati considerati come parte dello stesso braccio di intervento.

La variabile di outcome è stata definita utilizzando la quantità di sigarette (o altri prodotti di tabacco) fumata negli ultimi 30 giorni al baseline e al primo follow-up. In particolare, è stata codificata con valore 1 nel caso in cui al follow-up l'individuo fumasse la stessa quantità o una quantità inferiore di sigarette e 0 in caso contrario. Fanno eccezione i soggetti che al baseline e al follow-up hanno fumato più di 30 sigarette negli ultimi 30 giorni; in questo caso la variabile assume valore 0.

Le variabili di esposizione sono state codificate nel seguente modo:

- sex: indica il sesso della persona, assume 0 nei soggetti maschi e 1 nei soggetti femmine;
- age: indica l'età calcolata al baseline;
- smoking: indica la quantità di sigarette fumate negli ultimi 30 giorni. Assume i seguenti valori: 1 (0 sigarette fumate), 2 (1-2 sigarette fumate), 3 (3-5 sigarette fumate), 4 (6-9 sigarette fumate), 5 (10-19 sigarette fumate), 6 (20-29 sigarette fumate), 7 (30 o più sigarette fumate);
- risk perceived: indica i rischi legati al fumo percepiti dal soggetto (problemi con amici, problemi con genitori, sviluppo di dipendenza, problemi economici). Assume valori interi nell'intervallo [4, 16]. Più è alto è il valore e maggiori sono i rischi che l'individuo riconosce;
- benefit perceived: indica la percezione di conseguenze positive del fumo da parte del soggetto (rilassamento, maggiore popolarità, divertimento e fiducia in se stessi). Assume valori interi nell'intervallo [4, 16]. Più è alto è il valore e maggiori sono i benefici che l'individuo riconosce;
- other smoker: indica la presenza di fumatori tra i conoscenti più stretti come genitori, fratelli e amici. Assume valore 1 se ci sono fumatori nella cerchia di conoscenti e in 0 in caso contrario.
- family support: indica il livello di sostegno e regole che il soggetto riceve dalla famiglia. Assume valori interi nell'intervallo [5, 20]. Più è basso il valore, più l'individuo si ritiene in una situazione familiare in cui possa sentirsi supportato e in presenza di regole a cui ritiene importante non disobbedire;

- decision 1: indica se il soggetto ha la tendenza a portare o meno a termine quanto ha iniziato. Assume valori interi nell'intervallo [1, 4]. Valore minori indicano che l'individuo porta a termine più spesso i propri obiettivi;
- decision 2: indica l'abitudine del soggetto a prendere decisioni senza pensare alle conseguenze. Assume valori interi nell'intervallo [1, 4]. Valori maggiori indicano che l'individuo tende a non pensare alle conseguenze più spesso;
- decision 3: indica l'abitudine del soggetto a prendere decisioni pesando le conseguenze. Assume valori interi nell'intervallo [1, 4]. Valori maggiori indicano che l'individuo tende a pesare le conseguenze;
- decision 4: indica l'abitudine del soggetto a rimpiangere le proprie scelte. Assume valori interi nell'intervallo [1, 4]. Valori maggiori indicano che l'individuo tende a rimpiangere le proprie scelte;
- decision 5: indica l'influenza degli amici nel processo decisionale del soggetto. Assume valori interi nell'intervallo [1, 4]. Valori maggiori indicano che l'individuo tende a non farsi influenzare dagli amici;
- family both parents: l'individuo vive con entrambi i genitori. Assume valore 1 nel caso in cui il soggetto viva con entrambi i genitori e 0 altrimenti;
- family one parent: l'individuo vive con un solo genitore. Assume valore 1 in caso in cui il soggetto viva uno e un solo genitore e 0 altrimenti.
- family others: l'individuo non vive con i genitori. Assume valore 1 nel caso in cui il soggetto viva senza i genitori e 0 in caso contrario.

3.2 Pre-processamento dei dati

Imputazione dei valori mancanti La sostituzione dei valori mancanti nel dataframe è stata effettuata tramite il metodo IterativeImputer della libreria di python scikit-learn versione 1.4.2 (Pedregosa et al., 2011).

Normalizzazione dei dati I dati sono stati normalizzati con il metodo StandardScaler della libreria di python scikit-learn versione 1.4.2 (Pedregosa et al., 2011).

Riduzione della dimensionalità La riduzione della dimensionalità è stata effettuata con il metodo UMAP della libreria di python umap-learn versione 0.5.6 (McInnes et al., 2018). Per rendere il risultato riproducibile è stato fissato il valore del parametro random_state.

Per considerare le differenze tra i diversi tipi di variabili, è stata utilizzata la distanza di Gower (Gower, 1971), implementata nella libreria Python gower (versione 0.1.2).

3.3 Clustering

L'algoritmo di clustering utilizzato è K-means, implementato nella libreria di python scikit-learn versione 1.4.2 (Pedregosa et al., 2011). Per ottimizzare il numero di cluster è stato utilizzato il metodo del gomito descritto in Nainggolan et al. (2019). Infine, per valutare la qualità del clustering è stato calcolato il silhouette score, anch'esso implementato nella libreria di python scikit-learn versione 1.4.2 (Pedregosa et al., 2011).

3.4 Analisi statistiche

Per valutare se l'intervento sia stato o meno efficace, è stato utilizzato il test del Chi-quadrato implementato nella libreria di python SciPy versione 1.11.4 (Virtanen et al., 2020). Il test statistico è stato effettuato in ogni cluster sulla variabile di outcome confrontando il gruppo di intervento e di controllo.

Per valutare differenze della variabile di outcome tra i cluster nei gruppi di controllo sono stati utilizzati il test del Chi-quadrato e la regressione logistica, implementati nelle librerie di python SciPy versione 1.11.4 e statsmodels versione 0.14.2 (Virtanen et al., 2020; Seabold and Perktold, 2010).

Per valutare le differenze di efficacia dell'intervento tra i vari cluster, calcolare gli odds ratio e gli intervalli di confidenza, è stata condotta una regressione logistica utilizzando la libreria Python statsmodels (versione 0.14.2) (Seabold and Perktold, 2010).

Infine, per trovare le variabili che potessero spiegare le differenze di efficacia tra coppie di cluser sono stati effettuati tre test statistici a seconda del tipo di variabile. Per le variabili binarie è stato condotto il test del chi-quadrato implementato nella libreria di python SciPy versione 1.11.4 (Virtanen et al., 2020). Per le variabili ordinarie, invece, è stato condotto il test di Mann-Whitney U, anch'esso implementato nella libreria di python SciPy versione 1.11.4 (Virtanen et al., 2020). Infine, per le variabili continue è stata valutata la normalità con il test di Shapiro: nel caso la variabile sia risultata normale allora è stato condotto il test ANOVA; in caso contrario il test Kruskal-Wallis. Tutti e tre i test sono implementati nella libreria di python SciPy versione 1.11.4 (Virtanen et al., 2020).

3.5 Altri strumenti utilizzati

La versione di python utilizzata in tutte le analisi è Python 3.12.3. Altre librerie di python utilizzate sono: pandas versione 2.1.4 (pandas development team, 2020), numpy versione 1.26.4 (Harris et al., 2020), seaborn versione 0.13.2 (Waskom, 2021) e matplotlib versione 3.8.4 (Hunter, 2007). I grafici in figura 1 e figura 2 sono stati creati con il pacchetto R versione 4.3.3 ggplot2.

4 Risultati

4.1 Clustering

In Figura 5 è rappresentato il grafico della variazione del valore di Sum Square Error (SSE) al variare del numero dei cluster.

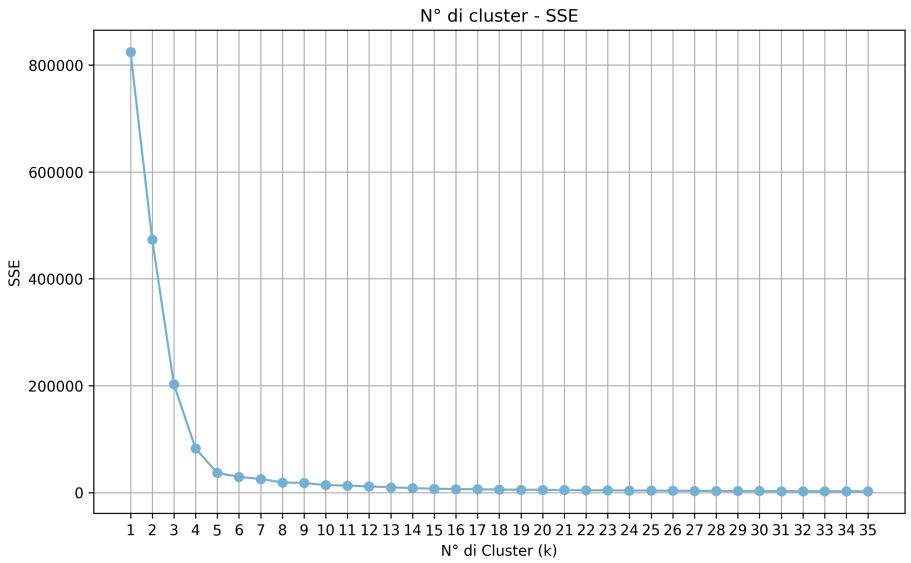

Figura 5: Variazione del valore di SSE in base al numero di cluster. Il punto rosso rappresenta il numero di cluster scelti in base al metodo del gomito

In figura 6 è sono mostrati i cluster trovati in seguito alla riduzione della dimensionalità del set di dati.

Il valore di silhouette score ottenuto per k=5 vale: 0.76.

In Tabella 1 è rappresentata la distribuzione all'interno dei cluster dei soggetti nel gruppo di controllo e nel gruppo di intervento.

Cluster	**Controllo**	**Intervento**
0	1280	987
1	566	452
2	1029	869
3	540	527
4	468	361

Tabella 1: Distribuzione dei soggetti nel gruppo di controllo e nel gruppo di intervento per cluster

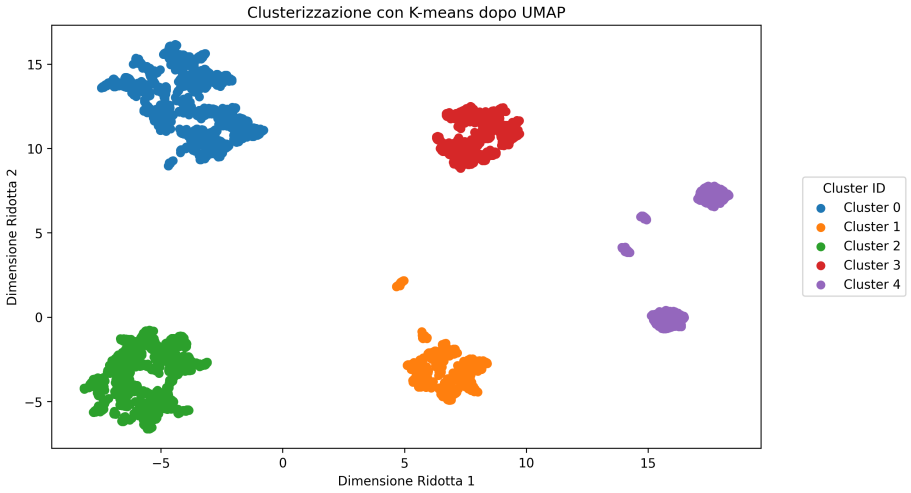

Figura 6: Cluster ottenuti con metodo K-means impostando K=6 dopo la riduzione della dimensionalità

Nella Tabella 2 sono riportate le medie e le deviazioni standard (σ) del numero di individui per cui la variabile di outcome assume valore 1, suddivisi nei gruppi di controllo e intervento all'interno di ciascun cluster. In Figura 7 è presentato un istogramma che mostra i valori della Tabella 2. Poiché la variabile di outcome è binaria, la media e la percentuale di individui con valore 1 sono equivalenti.

Cluster	**Controllo**	**Intervento**
0	0.575000 ± 0.494536	0.789260 ± 0.408041
1	0.763251 ± 0.425463	0.887168 ± 0.316738
2	0.629738 ± 0.483110	0.752589 ± 0.431756
3	0.759259 ± 0.427930	0.889943 ± 0.313258
4	0.523504 ± 0.499982	0.739612 ± 0.439455

Tabella 2: Medie e deviazioni standard del numero di individui con valore 1 della variabile di outcome nei gruppi di controllo e intervento, suddivisi per cluster.

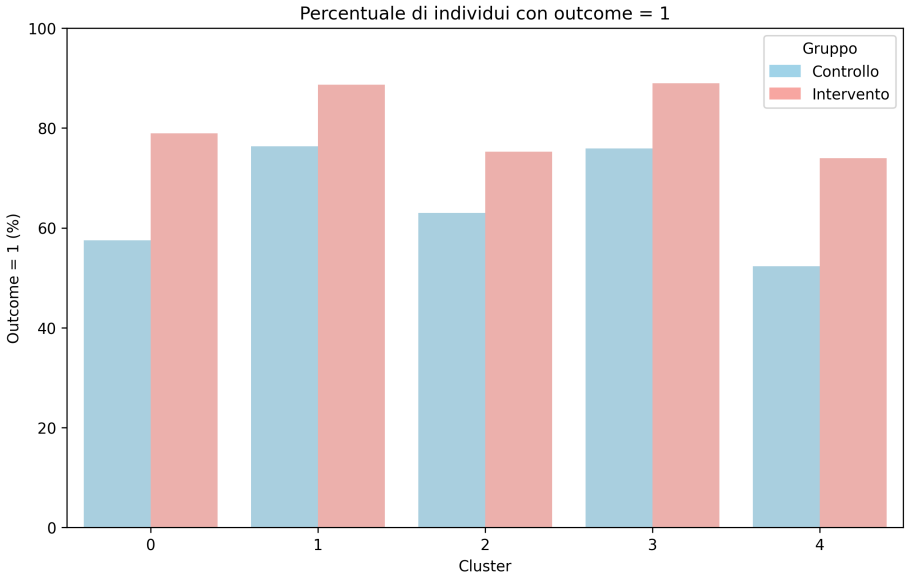

Figura 7: Percentuali di soggetti con la variabile di outcome pari a 1 nei gruppi di intervento e controllo per ciascun cluster

Nelle Tabelle 3 - 6 sono mostrate suddivise per cluster la media, la mediana, la deviazione standard e i valori massimi e minimi per le variabili age, risk perceived, benefit perceived e family support.

Cluster	**Mean**	**Median**	**Std**	**Min**	**Max**
0	13.37	13.0	1.12	11	19
1	13.13	13.0	0.93	11	18
2	13.29	13.0	0.99	12	18
3	13.16	13.0	0.87	11	17
4	13.38	13.0	1.06	12	18

Tabella 3: Statistiche descrittive per l'età (age) suddivise per cluster.

Cluster	**Mean**	**Median**	**Std**	**Min**	**Max**
0	10.53	10	4.21	4	32
1	9.78	9	4.38	4	32
2	10.08	10	3.41	4	32
3	9.68	9	4.18	4	32
4	10.59	10	4.16	4	32

Tabella 4: Statistiche descrittive per il rischio percepito (risk perceived) suddivise per cluster.

Cluster	Mean	Median	Std	Min	Max
0	13.44	13	4.78	4	32
1	14.18	14	4.13	4	32
2	13.05	13	3.69	4	32
3	13.94	14	4.08	4	32
4	13.48	13	4.10	4	32

Tabella 5: Statistiche descrittive per il beneficio percepito (benefit perceived) suddivise per cluster.

Cluster	Mean	Median	Std	Min	Max
0	9.35	9	4.92	5	40
1	8.15	8	3.05	5	40
2	8.83	8	3.56	5	40
3	8.15	8	3.49	5	40
4	9.15	9	3.77	5	40

Tabella 6: Statistiche descrittive per il supporto familiare (family support) suddivise per cluster.

Nelle Tabelle 7 - 12 è mostrata la distribuzione percentuale dei soggetti tra le diverse categorie delle variabili smoking, decision_1, decision_2, decision_3, decision_4 e decision_5.

Cluster	1	2	3	4	5	6	7
0	0.79	0.06	0.03	0.02	0.02	0.02	0.06
1	0.95	0.03	0.01	0.00	0.00	0.00	0.01
2	0.75	0.08	0.04	0.02	0.02	0.02	0.06
3	0.95	0.02	0.01	0.00	0.00	0.00	0.00
4	0.76	0.07	0.03	0.02	0.03	0.02	0.06

Tabella 7: Distribuzione percentuale dei soggetti per i valori della variabile fumo (smoking) per cluster.

Cluster	1	2	3	4
0	0.28	0.54	0.15	0.03
1	0.25	0.59	0.15	0.01
2	0.27	0.56	0.15	0.01
3	0.24	0.59	0.16	0.02
4	0.28	0.56	0.14	0.02

Tabella 8: Distribuzione percentuale dei soggetti per i valori della variabile decisione 1 (decision_1) per cluster.

Cluster	1	2	3	4
0	0.13	0.38	0.38	0.11
1	0.07	0.32	0.46	0.15
2	0.10	0.39	0.39	0.12
3	0.07	0.34	0.43	0.15
4	0.11	0.39	0.38	0.12

Tabella 9: Distribuzione percentuale dei soggetti per i valori della variabile decisione 2 (decision_2) per cluster.

Cluster	1	2	3	4
0	0.24	0.48	0.23	0.04
1	0.22	0.54	0.21	0.03
2	0.21	0.49	0.27	0.04
3	0.25	0.53	0.20	0.03
4	0.20	0.46	0.30	0.04

Tabella 10: Distribuzione percentuale dei soggetti per i valori della variabile decisione 3 (decision_3) per cluster.

Cluster	1	2	3	4
0	0.16	0.42	0.34	0.09
1	0.10	0.38	0.43	0.09
2	0.14	0.42	0.37	0.07
3	0.10	0.36	0.42	0.11
4	0.13	0.37	0.39	0.11

Tabella 11: Distribuzione percentuale dei soggetti per i valori della variabile decisione 4 (decision_4) per cluster.

Cluster	1	2	3	4
0	0.21	0.34	0.33	0.12
1	0.13	0.35	0.39	0.13
2	0.16	0.31	0.40	0.13
3	0.19	0.38	0.32	0.11
4	0.20	0.37	0.33	0.11

Tabella 12: Distribuzione percentuale dei soggetti per i valori della variabile decisione 5 (decision_5) per cluster.

Nelle Tabelle 13 - 17 è mostrata la distribuzione percentuale dei soggetti tra i valori delle le variabili sex, other_smoker, family_one_parent, family_both_parent e family_other .

Cluster	Maschi	Femmine
0	1.00	0.00
1	0.03	0.97
2	0.00	1.00
3	1.00	0.00
4	0.48	0.52

Tabella 13: Distribuzione del sesso nei vari cluster.

Cluster	N	Y
0	0.00	1.00
1	0.96	0.04
2	0.00	1.00
3	1.00	0.00
4	0.18	0.82

Tabella 14: Distribuzione degli altri fumatori nei vari cluster.

Cluster	N	Y
0	1.00	0.00
1	1.00	0.00
2	1.00	0.00
3	1.00	0.00
4	0.00	1.00

Tabella 15: Distribuzione degli individui con un solo genitore nei vari cluster.

Cluster	N	Y
0	0.00	1.00
1	0.06	0.94
2	0.00	1.00
3	0.00	1.00
4	1.00	0.00

Tabella 16: Distribuzione degli individui con entrambi i genitori nei vari cluster.

Cluster	N	Y
0	1.00	0.00
1	0.94	0.06
2	1.00	0.00
3	1.00	0.00
4	1.00	0.00

Tabella 17: Distribuzione degli individui con altre strutture familiari nei vari cluster.

4.2 Analisi statistiche

In Tabella 18 sono riportati i risultati del test Chi-quadrato, che evidenziano le differenze nella variabile di outcome tra i gruppi di controllo e intervento all'interno di ciascun.

Cluster	Chi-quadrato	P-value	DF	Frequenze Attese
0	114.442437	1.042490e-26	1	[[424.60, 855.40], [327.40, 659.60]]
1	25.124098	5.375688e-07	1	[[102.86, 463.14], [82.14, 369.86]]
2	32.441390	1.228420e-08	1	[[323.12, 705.88], [272.88, 596.12]]
3	30.487044	3.361037e-08	1	[[95.15, 444.85], [92.85, 434.15]]
4	39.391493	3.468046e-10	1	[[178.96, 289.04], [138.04, 222.96]]

Tabella 18: Risultati del test Chi-quadrato per i vari cluster, comprensivi di valori Chi-quadrato, P-value, gradi di libertà e frequenze attese.

In Tabella 19 sono riportati i risultati della regressione logistica, utilizzata per analizzare le differenze nella variabile di outcome tra i vari cluster, considerando solo il gruppo di controllo.

Variabile	Coefficiente	Errore Standard	z	$P > \|z\|$	IC
Costante	0.3023	0.057	5.346	0.000	[0.191, 0.413]
Cluster 1	0.8683	0.114	7.623	0.000	[0.645, 1.092]
Cluster 2	0.2288	0.086	2.666	0.008	[0.061, 0.397]
Cluster 3	0.8463	0.115	7.331	0.000	[0.620, 1.073]
Cluster 4	-0.2082	0.108	-1.920	0.055	[-0.421, 0.004]

Tabella 19: Risultati della regressione logistica per analizzare le differenze significative nella variabile di outcome tra i diversi cluster, considerando solo il gruppo di controllo

In Figura 8 è rappresentato un grafico con i risultati della regressione logistica per evidenziare le differenze di efficacia dell'intervento nei vari cluster. Sono mostrati gli Odd Ratio (OR) e gli intervalli di confidenza (IC) al 95%.

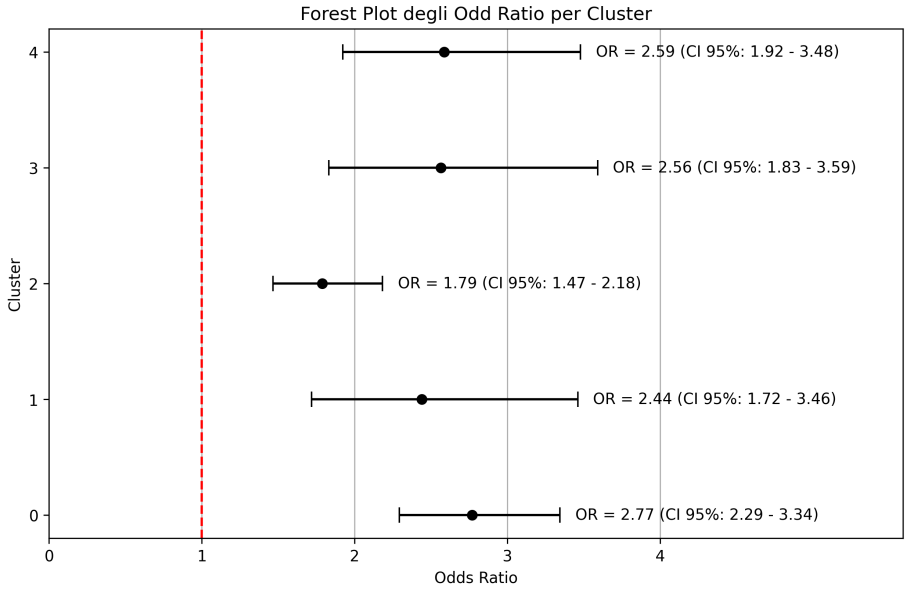

Figura 8: Grafico degli Odds Ratio (OR) e degli intervalli di confidenza (IC) al 95% ottenuti dalla regressione logistica, che evidenzia le differenze di efficacia dell'intervento nei vari cluster.

Nelle Tabelle 20, 21 e 22 sono riportati i risultati dei test statistici (Kruskal-Wallis, Mann-Whitney U e Chi-quadro) per analizzare quali variabili differiscono in modo statisticamente significativo tra il cluster 0 e il cluster 2.

Variabile	**H-Statistic**	**P-Value**
age	1.854831	0.173223
risk_perceived	5.931790	0.014870
benefit_perceived	1.036777	0.308572
family_support	4.959972	0.025941

Tabella 20: Risultati del test Kruskal-Wallis per le variabili *age*, *risk_perceived*, *benefit_perceived*, e *family_support*.

Variabile	U-Statistic	P-Value
smoking	2078165.0	0.009782
decision_1	2171880.0	0.554777
decision_2	2081704.0	0.054988
decision_3	2058849.5	0.009791
decision_4	2112789.5	0.287031
decision_5	1973533.5	0.000001

Tabella 21: Risultati del test Mann-Whitney U per le variabili *smoking*, *decision_1*, *decision_2*, *decision_3*, *decision_4*, e *decision_5*.

Variabile	Chi2-Statistic	P-Value
sex	4140.874662	0.000000
other_smoker	0.000000	1.000000
family_Both parents	0.000000	1.000000
family_One parent	0.000000	1.000000
family_Others	0.000000	1.000000

Tabella 22: Risultati del test Chi-quadro per le variabili binarie *sex*, *other_smoker*, *family_Both parents*, *family_One parent*, e *family_Others*.

In Figura 9 è mostrata la differenza della distribuzione dei valori della variabile risk_perceived nel cluster 0 e nel cluster 5

In Figura 10 è mostrata la differenza della distribuzione dei valori della variabile family_support nel cluster 0 e nel cluster 5

In Figura 11 è rappresentata la distribuzione percentuale dei valori che assume la variabile smoking nel cluster 0 e nel cluster 2.

In Figura 12 è mostrata tramine la differenza della distribuzione dei valori della variabile decision_3. Per semplificare la visualizzazione è stata scelta una configurazione di colori specifica. In particolare, i valori 1 e 2 della variabile codificano l'informazione che lo studente è d'accordo con l'affermazione, mentre i valori 3 e 4 che non è d'accordo. Per questa ragione sono state utilizzate tonalità di azzurro quando per i valori 1 e 2 e di rosso per i valori 3 e 4. La stessa idea è stata applicata anche per la variabile decision_5, come si può vedere in Figura 13

In Figura 14 è mostrata la distribuzione percentuale dei maschi e delle femmine nel cluster 0 e nel cluster 2.

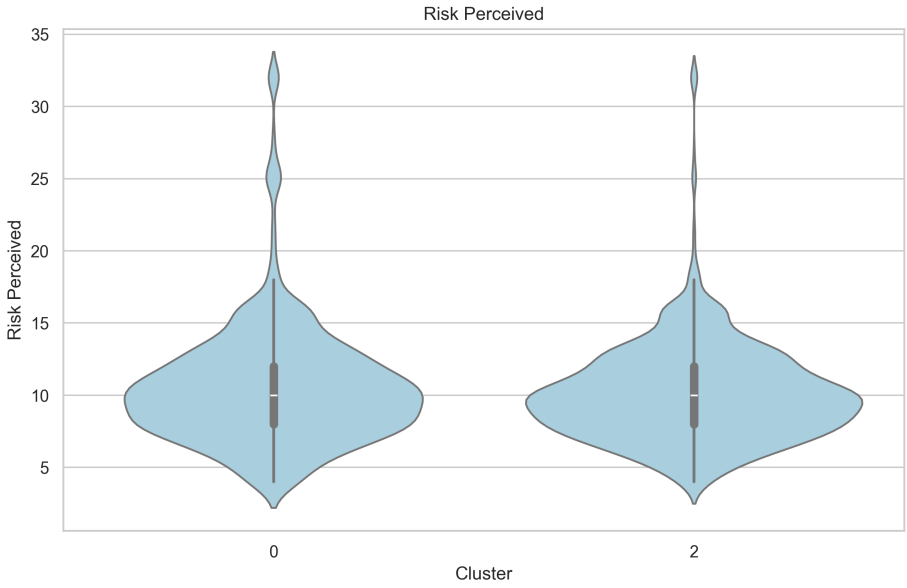

Figura 9: Violin plot della variabile risk_perceived nel cluster 0 e nel cluster 5

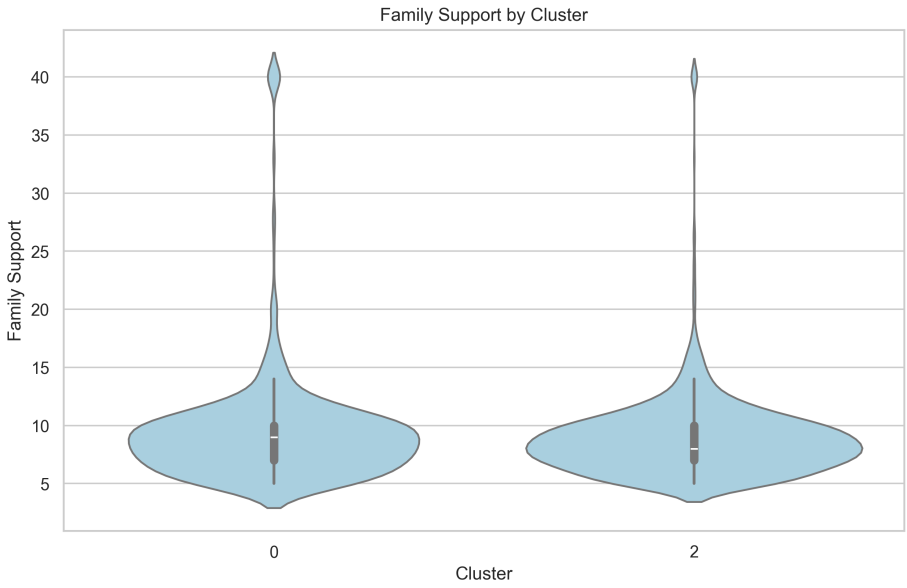

Figura 10: Violin plot della variabile family_support nel cluster 0 e nel cluster 5

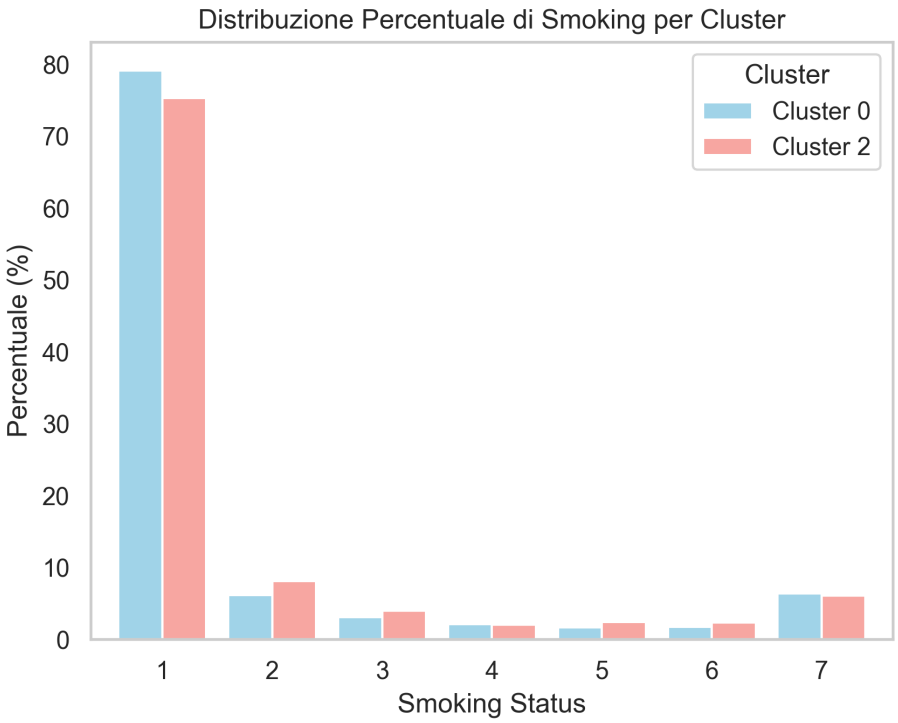

Figura 11: Distribuzione percentuale dei valori che assume la variabile smoking nel cluster 0 e nel cluster 2

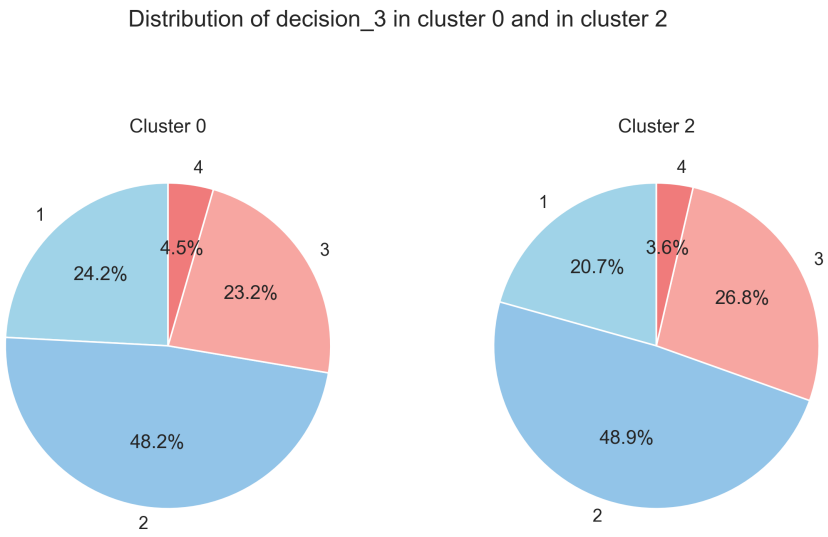

Figura 12: Distribuzione percentuale nei valori di decision_3 nel cluster 0 e nel cluster 2

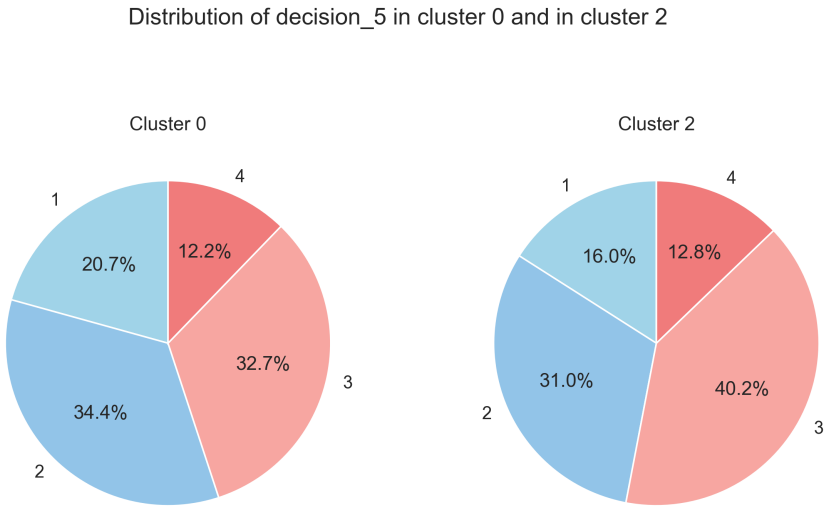

Figura 13: Distribuzione percentuale nei valori di decision_3 nel cluster 0 e nel cluster 2

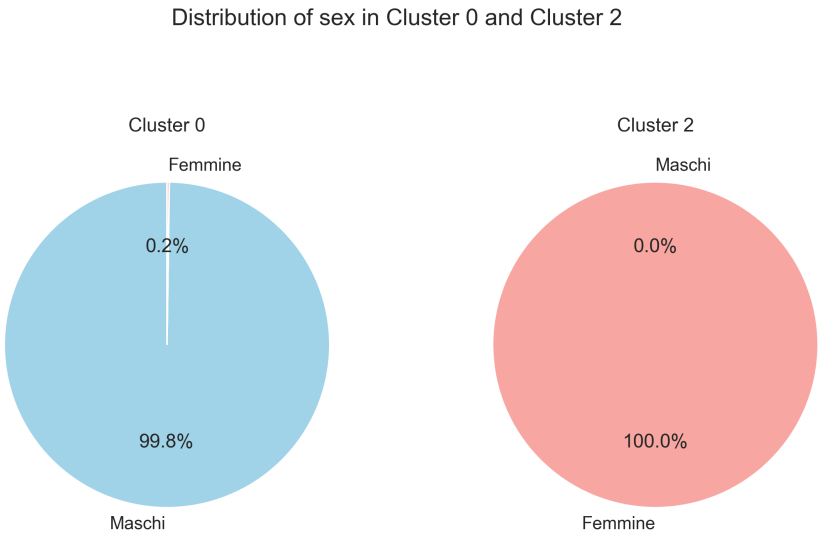

Figura 14: Distribuzione percentuale di maschi e di femmine nel cluster 0 e nel cluster 2

5 Discussione

5.1 Clustering

Attraverso il metodo descritto nel capitolo 3 è stato scelto di impostare l'algoritmo K-means con $K = 5$, ottenendo quindi 5 cluster. Sono stati testati altri valori di K ($K = 6$ e $K = 7$) nella regione del gomito della curva in Figura 5. I valori di silhouette score per $K > 5$ risultano maggiori a causa di una separazione dei punti nel cluster 4 in Figura 6, tuttavia è stato scelto di mantenere K=5. La ragione di questa scelta è che i nuovi cluster che si sarebbero formati avevano un numero di punti molto basso; di conseguenza nelle analisi statistiche venivano intervalli di confidenza molto ampi e poco significativi per l'interesse dello studio. Con questa configurazione, tramite le Tabelle 3 - 17 è possibile dare una iniziale descrizione qualitativa degli individui che caratterizzano i cluster.

Nel cluster 0 gli studenti sono tutti maschi, con almeno una persona vicina che fuma e con vivono entrambi i genitori.

Nel cluster 1 si hanno prevalentemente studentesse (97%), senza persone vicine che fumano, che vivono nel 94% dei casi con entrambi i genitori e nel restante 6% senza genitori.

Nel cluster 2 si hanno solo studentesse, con almeno una persona vicina che fuma e che vivono con entrambi i genitori.

Nel cluster 3 gli studenti sono tutti maschi, senza persone vicine che fumano e vivono con entrambi i genitori.

Nel cluster 4 troviamo tutti gli studenti e le studentesse che vivono con un solo genitore e nell'82% dei casi con almeno una persona vicina che fuma.

Per quanto riguarda le altre variabili è difficile verificare se ci siano particolari differenze senza usare test statistici adeguati che sono stati effettuati solo in casi specifici.

Un'ultima osservazione di carattere qualitativo riguarda le affermazioni sui metodi decisionali. In particolare per la variabile decision_1 che si ricorda indicare la determinazione nel raggiungere i propri obiettivi, la maggior parte degli studenti in tutti i cluster si trova d'accordo (i valori 1 e 2 indicano che gli studenti concordano quanto affermato, mentre i valori 3 e 4 che non concordano). Una distribuzione simile anche se meno marcata (soprattutto nel cluster 4) si osserva per la variabile decision_3, che si ricorda essere riferita alla tendenza degli individui a ponderare le proprie decisioni.

5.2 Analisi statistiche

Attraverso la regressione logistica presentata in Tabella 19, è emerso che esistono differenze significative tra i cluster riguardo alla probabilità di ridurre o mantenere costante la quantità di sigarette fumate negli ultimi 30 giorni. In particolare, analizzando i coefficienti e gli intervalli di confidenza si osserva che rispetto al cluster 0 (impostato come riferimento), nel cluster 1 e nel cluster 3 c'è una maggiore probabilità che gli individui senza alcun intervento, diminuiscano o mantengano costante la quantità di sigarette consumate. Anche nel

cluster 2 si osserva lo stesso comportamento, sebbene con un effetto più attenuato rispetto ai casi precedenti. Infine, non sono emerse differenze statisticamente significative tra il cluster 4 e il cluster 0.

Nella Tabella 6 si può osservare che sia nel cluster 0, sia nel cluster 4 gli studenti hanno un valore con media e mediana maggiore rispetto ai colleghi, indicando che questi soggetti potrebbero avere meno regole in famiglia, essere meno seguiti e trovare minore supporto da parte di genitori e amici. In particolare, nel cluster 4 la situazione potrebbe essere spiegata anche dal fatto che i soggetti non vivano con entrambi i genitori.

Il confronto tra il cluster 3 e il cluster 0 invece potrebbe mostrare l'effetto della presenza di un fumatore nella cerchia ristretta delle proprie conoscenze.

Dopo avere analizzato le differenze tra i gruppi di controllo è stata approfondita l'efficacia dell'intervento. In Figura 8 è mostrato il risultato convertito in Odds Ratio (OR) della regressione logistica per valutare l'efficacia intracluster dell'intervento. Il primo risultato ottenuto è che in tutti i cluster l'intervento è risultato efficace. Inoltre tra il cluster 2 e il cluster 0 non c'è sovrapposizione degli intervalli di confidenza, pertanto nel cluster 0 l'intervento è stato più efficace. Per spiegare queste differenze, sono state eseguite ulteriori analisi statistiche per individuare le variabili in cui c'è una differenza significativa tra i due gruppi. I risultati di queste analisi sono consultabili nelle Tabelle 20 - 22. Le variabili che sono risultate diverse in maniera statisticamente significativa sono:

- risk_perceived
- family_support
- smoking
- decision_3
- decision_5
- sex

In Figura 9 – Figura 9 sono mostrate tramite una rappresentazione grafica queste variabili. Per le variabili risk_perceived (Figura 9), smoking (Figura 11) e decision_3 (Figura 12) è difficile stabilire il modo in cui cambia la distribuzione. Di conseguenza è difficile formulare ipotesi sulle relazioni tra le variazioni e l'efficacia dell'intervento.

Per la variabile decision_5, è possibile individuare un cambio della percentuale di persone nei due cluster che sono in accordo o in disaccordo con l'affermazione: "There are several possible way to take decisions. How the following apply to you: No matter what friends think". Nel cluster 0 il 55.1% degli studenti è in accordo con questa affermazione, mentre nel cluster 2 la percentuale scende al 47%. Anche se i valori rimangono vicini al 50% nel primo gruppo gli studenti potrebbero non avere la tendenza a farsi influenzare dai propri pari.

I violin plot in Figura 10 e i valori nella Tabella 6 mostrano che nel cluster 0 c'è un minore supporto familiare percepito da parte degli studenti. Questo

sembra essere causato dal fatto che c'è una maggiore densità di individui nei valori più estremi (in alto riferendosi alla Figura 10) che spostano sia la media sia la mediana. L'intervento pertanto potrebbe avere influito positivamente su quei soggetti che, seppur costituiscano una minoranza della popolazione nel cluster, potrebbero avere avuto più fatica a ricevere un aiuto da parte della famiglia e degli amici.

Il sesso, a causa della marcata differenza tra i due gruppi come evidenziato nel grafico di Figura 14 e nella Tabella 13, sembra essere uno dei fattore che hanno maggiormente influenzato la differenza di successo tra i cluster. In particolare, nel cluster 0, composto esclusivamente da maschi, l'intervento è risultato più efficace rispetto al cluster 2, costituito invece da sole femmine. È comunque importante considerare che non in tutti i cluster composti esclusivamente da maschi l'intervento è stato più efficace rispetto a quelli composti solo da femmine. Questo risultato, di conseguenza, può essere spiegato sia dalle variabili che caratterizzano i soggetti dei due gruppi, sia, soprattutto, dal maggior numero di individui presenti nei cluster di interesse (Tabella 1), il che ha contribuito a ridurre l'ampiezza degli intervalli di confidenza e a migliorare la precisione delle stime statistiche.

6 Conclusioni

In questa tesi sono state utilizzate tecniche di apprendimento non supervisionato per individuare gruppi di studenti simili nello studio EU-Dap e valutare l'efficacia dell'intervento Unplugged sull'abitudine al fumo in ognuno di essi.

Prima di optare per UMAP, è stata testata anche la tecnica della PCA per la riduzione della dimensionalità. Tuttavia, per raggiungere una varianza spiegata di almeno l'85%, erano necessarie numerose dimensioni. Inoltre, quando si è provato a eseguire il clustering con diversi algoritmi, i valori di silhouette score che si ottenevano erano bassi e confrontabili a quelli ottenuti in assenza di riduzione della dimensionalità. Per questa ragione è stato scelto di provare ad utilizzare una tecnica di riduzione della dimensionalità non lineare come UMAP. Tra le varie tecniche disponibili, l'algoritmo UMAP è stato precedentemente utilizzato in diverse analisi biologiche e cliniche (Grollemund et al., 2020; Sakaue et al., 2020; Becht et al., 2019; Greenwood et al., 2022). Questa metodologia, inoltre, consente di mappare nuove osservazioni nello spazio latente senza necessità di riaddestrare il modello (Greenwood et al., 2022).

Per la fase di clustering, non è stato necessario testare algoritmi diversi da K-means, grazie all'efficacia con cui questo metodo ha individuato i cluster di interesse. All'interno di ciascun cluster si sono formati gruppi più omogenei per le variabili categoriali, mentre le variabili continue hanno mostrato una maggiore eterogeneità. Se si fosse optato per un numero di cluster più elevato, probabilmente si sarebbero ottenute classi più omogenee anche per le variabili continue. Tuttavia, questo avrebbe comportato un numero ridotto di individui in ogni cluster, soprattutto in presenza dei valori più estremi, rendendo più complessa l'interpretazione delle analisi statistiche. Dai risultati è emerso che in assenza dell'intervento la presenza di fumatori nella cerchia ristretta di conoscenze (amici e familiari) ha avuto un impatto negativo, portando gli studenti ad aumentare la quantità di sigarette fumate durante il periodo tra il baseline e il primo follow-up. Questo esito è in accordo con quanto trovato in diversi gruppi di ricerca dello stesso periodo (Hill et al., 2005; Irles et al., 2013) e più recenti (Vitória et al., 2020). Questo tipo di fenomeno viene spesso spiegato tramite la Social Learning Theory di Albert Bandura secondo cui le persone apprendono nuovi comportamenti osservando e imitando gli altri (Bandura, 1977; Irles et al., 2013).

L'analisi condotta per indagare la differenza di efficacia dell'intervento nei diversi cluster ha avuto due principali esiti. In primo luogo, è che l'intervento è risultato efficace in tutti i gruppi sebbene con delle differenze tra essi. In secondo luogo, il sesso potrebbe avere avuto un'influenza sull'efficacia dell'intervento. In particolare, il confronto tra un cluster composto esclusivamente maschi e uno formato da sole studentesse, ha mostrato un'efficacia superiore nel primo in modo statisticamente significativo. Tuttavia, il sesso potrebbe non essere stato l'unico fattore determinante vista la presenza di altre variabili in questi due gruppi che differiscono in maniera significativa e altri cluster con simili ripartizioni tra i sessi senza una differenza di efficacia significativa. Gli studi condotti in precedenza hanno comunque messo in evidenza una maggiore

efficacia dell'intervento nei maschi utilizzando tecniche diverse e una definizione di efficacia differente rispetto a quella utilizzata in questa tesi (Vigna-Taglianti et al., 2014). La stessa disuguaglianza è emersa anche in un trial condotto in Brasile con il programma "Education Against Tobacco" (Lisboa et al., 2019). Un recente studio in Spagna ha messo in evidenza che tra le adolescenti, rispetto alla controparte maschile, c'è un consumo maggiore di sigarette e una maggiore suscettibilità mediata per lo più dalla presenza di fumatori tra le amicizie (Santano-Mogena et al., 2023).

Uno dei principali limiti in questa analisi è che le feature selezionate non permettono di studiare associazioni note in letteratura come quella tra il consumo di tabacco e di alcol (Grucza and Bierut, 2006). In generale, questo processo ha ridotto la quantità di informazioni rispetto a quelle disponibili, permettendo di individuare solo associazioni già note in letteratura, non consentendo di sfruttare al meglio le potenzialità degli strumenti di intelligenza artificiale utilizzati.

Nonostante i limiti sopracitati di questo lavoro, attraverso questa pipeline sono stati trovati risultati che trovano conferma nelle precedenti analisi statistiche. Grazie alla scelta di usare algoritmi che funzionano bene anche in presenza di dataset di grandi dimensioni, nelle prossime analisi sarà possibile ampliare le informazioni utilizzate mantenendo il flusso di lavoro pressoché invariato con modifiche che saranno prevalentemente nella fase di preprocessamento dei dati e nella scelta dell'algoritmo di clustering.

Per concludere, gli strumenti di intelligenza artificiale in epidemiologia possono rappresentare un valido supporto a quelli della statistica classica. A differenza dei modelli tradizionali, come la regressione di Cox (Cox, 1972), gli approcci di machine learning non richiedono spesso assunzioni di linearità, consentendo di evidenziare interazioni più complesse, soprattutto in presenza di molte variabili, come accade nei fascicoli sanitari elettronici o negli studi sull'esposoma (Atehortúa et al., 2023; Hamilton et al., 2021).

Riferimenti bibliografici

Altman, N. and Krzywinski, M. (2018). The curse (s) of dimensionality. *Nat Methods*, 15(6):399–400.

Andrus, L. H., Hyde, D. F., and Fischer, E. (1964). Smoking by high school students—failure of a campaign to persuade adolescents not to smoke. *California Medicine*, 101(4):246.

Anselma, L., Bottrighi, A., Molino, G., Montani, S., Terenziani, P., and Torchio, M. (2011). Supporting knowledge-based decision making in the medical context: The glare approach. *International Journal of Knowledge-Based Organizations (IJKBO)*, 1(1):42–60.

Askin, S., Burkhalter, D., Calado, G., and El Dakrouni, S. (2023). Artificial intelligence applied to clinical trials: opportunities and challenges. *Health and technology*, 13(2):203–213.

Atehortúa, A., Gkontra, P., Camacho, M., Diaz, O., Bulgheroni, M., Simonetti, V., Chadeau-Hyam, M., Felix, J. F., Sebert, S., and Lekadir, K. (2023). Cardiometabolic risk estimation using exposome data and machine learning. *International Journal of Medical Informatics*, 179:105209.

Bafunno, D., Catino, A., Lamorgese, V., Pizzutilo, P., Di Lauro, A., Petrillo, P., Lapadula, V., Mastrandrea, A., Ricci, D., and Galetta, D. (2019). Tobacco control in europe: a review of campaign strategies for teenagers and adults. *Critical Reviews in Oncology/Hematology*, 138:139–147.

Bandura, A. (1977). Social learning theory. *Englewood Cliffs*.

Basile, A. O., Yahi, A., and Tatonetti, N. P. (2019). Artificial intelligence for drug toxicity and safety. *Trends in pharmacological sciences*, 40(9):624–635.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44.

Boswell, D. (2002). Introduction to support vector machines. *Departement of Computer Science and Engineering University of California San Diego*, 11:16–17.

Boucher, P. (2019). How artificial intelligence works. *EPRS–European Parliamentary Research Service, In: europarl. europa*, 2.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Buchmann, A. F., Blomeyer, D., Jennen-Steinmetz, C., Schmidt, M. H., Esser, G., Banaschewski, T., and Laucht, M. (2013). Early smoking onset may promise initial pleasurable sensations and later addiction. *Addiction biology*, 18(6):947–954.

Chang, W., Liu, Y., Xiao, Y., Yuan, X., Xu, X., Zhang, S., and Zhou, S. (2019). A machine-learning-based prediction method for hypertension outcomes based on medical data. *Diagnostics*, 9(4):178.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters.

ESPAD Group (2020). *ESPAD Report 2019: Results from the European School Survey Project on Alcohol and Other Drugs*. EMCDDA Joint Publications. Publications Office of the European Union, Luxembourg.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Faggiano, F., Siliquini, R., Vigna Taglianti, F., Panella, M., Group, E.-D. S., et al. (2006). Unplugged, an effective school-based program for the prevention of substance use among adolescents. eudap final technical report n. 1.

Feijoo, F., Palopoli, M., Bernstein, J., Siddiqui, S., and Albright, T. E. (2020). Key indicators of phase transition for clinical trials through machine learning. *Drug discovery today*, 25(2):414–421.

Feliu, A., Filippidis, F. T., Joossens, L., Fong, G. T., Vardavas, C. I., Baena, A., Castellano, Y., Martínez, C., and Fernández, E. (2019). Impact of tobacco control policies on smoking prevalence and quit ratios in 27 european union countries from 2006 to 2014. *Tobacco Control*, 28(1):101–109.

Flay, B. R. (1985). Psychosocial approaches to smoking prevention: a review of findings. *Health psychology*, 4(5):449.

Gates, A., Gates, M., Sim, S., Elliott, S. A., Pillay, J., and Hartling, L. (2021). Creating efficiencies in the extraction of data from randomized trials: a prospective evaluation of a machine learning and text mining tool. *BMC medical research methodology*, 21:1–12.

Goldstein, B. A. and Rigdon, J. (2019). Using machine learning to identify heterogeneous effects in randomized clinical trials—moving beyond the forest plot and into the forest. *JAMA network open*, 2(3):e190004–e190004.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gorini, G., Gallus, S., Carreras, G., Cortini, B., Vannacci, V., Charrier, L., Cavallo, F., Molinaro, S., Galeone, D., Spizzichino, L., et al. (2019). A long way to go: 20-year trends from multiple surveillance systems show a still huge use of tobacco in minors in italy. *European Journal of Public Health*, 29(1):164–169.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871.

Gravely, S., Giovino, G. A., Craig, L., Commar, A., D'Espaignet, E. T., Schotte, K., and Fong, G. T. (2017). Implementation of key demand-reduction measures of the who framework convention on tobacco control and change in smoking prevalence in 126 countries: an association study. *The Lancet Public Health*, 2(4):e166–e174.

Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., and Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100.

Greenwood, D., Taverner, T., Adderley, N. J., Price, M. J., Gokhale, K., Sainsbury, C., Gallier, S., Welch, C., Sapey, E., Murray, D., et al. (2022). Machine learning of covid-19 clinical data identifies population structures with therapeutic potential. *Iscience*, 25(7).

Grollemund, V., Chat, G. L., Secchi-Buhour, M.-S., Delbot, F., Pradat-Peyre, J.-F., Bede, P., and Pradat, P.-F. (2020). Development and validation of a 1-year survival prognosis estimation model for amyotrophic lateral sclerosis using manifold learning algorithm umap. *Scientific reports*, 10(1):13378.

Grucza, R. A. and Bierut, L. J. (2006). Cigarette smoking and the risk for alcohol use disorders among adolescent drinkers. *Alcoholism: Clinical and Experimental Research*, 30(12):2046–2054.

Guha, S., Rastogi, R., and Shim, K. (1998). Cure: An efficient clustering algorithm for large databases. *ACM Sigmod record*, 27(2):73–84.

Guha, S., Rastogi, R., and Shim, K. (2000). Rock: A robust clustering algorithm for categorical attributes. *Information systems*, 25(5):345–366.

Hamilton, A. J., Strauss, A. T., Martinez, D. A., Hinson, J. S., Levin, S., Lin, G., and Klein, E. Y. (2021). Machine learning and artificial intelligence: applications in healthcare epidemiology. *Antimicrobial Stewardship 38; Healthcare Epidemiology*, 1(1):e28.

Hansson, S. O. (2014). Why and for what are clinical trials the gold standard? *Scandinavian Journal of Public Health*, 42(13_suppl):41–48.

Harrer, S., Shah, P., Antony, B., and Hu, J. (2019). Artificial intelligence for clinical trial design. *Trends in pharmacological sciences*, 40(8):577–591.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Haugeland, J. (1989). *Artificial intelligence: The very idea*. MIT press.

Hiilamo, H. and Glantz, S. (2022). Global implementation of tobacco demand reduction measures specified in framework convention on tobacco control. *Nicotine and Tobacco Research*, 24(4):503–510.

Hill, K. G., Hawkins, J. D., Catalano, R. F., Abbott, R. D., and Guo, J. (2005). Family influences on the risk of daily smoking initiation. *Journal of Adolescent health*, 37(3):202–210.

Hu, T., Gall, S. L., Widome, R., Bazzano, L. A., Burns, T. L., Daniels, S. R., Dwyer, T., Ikonen, J., Juonala, M., Kähönen, M., et al. (2020). Childhood/adolescent smoking and adult smoking and cessation: the international childhood cardiovascular cohort (i3c) consortium. *Journal of the American Heart Association*, 9(7):e014381.

Hughes, K. S., Zhou, J., Bao, Y., Singh, P., Wang, J., and Yin, K. (2020). Natural language processing to facilitate breast cancer research and management. *The Breast Journal*, 26(1):92–99.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Irles, D. L., Pertusa, M. G., Guijarro, Á. B., and Carbonell, M. J. F. (2013). Parent and peer influence models in the onset of adolescent smoking. *Salud y drogas*, 13(1):59–65.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer Texts in Statistics. Springer New York, 2 edition.

Joossens, L. and Raw, M. (2006). The tobacco control scale: a new scale to measure country activity. *Tobacco control*, 15(3):247–253.

Karypis, G., Han, E.-H., and Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *computer*, 32(8):68–75.

Kataria, A. and Singh, M. (2013). A review of data classification using k-nearest neighbour algorithm. *International Journal of Emerging Technology and Advanced Engineering*, 3(6):354–360.

Kingsford, C. and Salzberg, S. L. (2008). What are decision trees? *Nature biotechnology*, 26(9):1011–1013.

Koesmahargyo, V., Abbas, A., Zhang, L., Guan, L., Feng, S., Yadav, V., and Galatzer-Levy, I. R. (2020). Accuracy of machine learning-based prediction of medication adherence in clinical research. *Psychiatry research*, 294:113558.

Kurt, I., Ture, M., and Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert systems with applications*, 34(1):366–374.

Lantz, P. M., Jacobson, P. D., Warner, K. E., Wasserman, J., Pollack, H. A., Berson, J., and Ahlstrom, A. (2000). Investing in youth tobacco control: a review of smoking prevention and control strategies. *Tobacco control*, 9(1):47–63.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lee, C. S. and Lee, A. Y. (2020). How artificial intelligence can transform randomized controlled trials. *Translational vision science & technology*, 9(2):9–9.

Li, R., Li, L., Xu, Y., and Yang, J. (2022). Machine learning meets omics: applications and perspectives. *Briefings in Bioinformatics*, 23(1):bbab460.

Lisboa, O. C., Bernardes-Souza, B., Xavier, L. E. D. F., Almeida, M. R., Corrêa, P. C. R. P., and Brinker, T. J. (2019). A smoking prevention program delivered by medical students to secondary schools in brazil called "education against tobacco": Randomized controlled trial. *J Med Internet Res*, 21(2):e12854.

Martani, A., Geneviève, L. D., Poppe, C., Casonato, C., and Wangmo, T. (2020). Digital pills: a scoping review of the empirical literature and analysis of the ethical aspects. *BMC medical ethics*, 21:1–13.

McCauley, N. and Ala, M. (1992). The use of expert systems in the healthcare industry. *Information & Management*, 22(4):227–235.

McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Meng, Y., Yang, Y., Hu, M., Zhang, Z., and Zhou, X. (2023). Artificial intelligence-based radiomics in bone tumors: Technical advances and clinical application. In *Seminars in Cancer Biology*. Elsevier.

Miller, R. A., Pople Jr, H. E., and Myers, J. D. (1985). Internist-i, an experimental computer-based diagnostic consultant for general internal medicine. In *Computer-assisted medical decision making*, pages 139–158. Springer.

Morison, J. B., Medovy, H., and MacDonell, G. T. (1964). Health education and cigarette smoking: A report on a three-year program in the winnipeg school division, 1960-1963. *Canadian Medical Association Journal*, 91(2):49.

Muehlematter, U. J., Daniore, P., and Vokinger, K. N. (2021). Approval of artificial intelligence and machine learning-based medical devices in the usa and europe (2015–20): a comparative analysis. *The Lancet Digital Health*, 3(3):e195–e203.

Nainggolan, R., Perangin-angin, R., Simarmata, E., and Tarigan, A. F. (2019). Improved the performance of the k-means cluster using the sum of squared error (sse) optimized by using the elbow method. In *Journal of Physics: Conference Series*, volume 1361, page 012015. IOP Publishing.

Njue, M. and Franklin, B. (2020). Dimensionality reduction on mnist dataset using pca, t-sne and umap.

Organization, W. H. (n.d.). Tobacco. Accessed: 2024-10-01.

pandas development team, T. (2020). pandas-dev/pandas: Pandas.

Paraje, G., Flores Muñoz, M., Wu, D. C., and Jha, P. (2024). Reductions in smoking due to ratification of the framework convention for tobacco control in 171 countries. *Nature Medicine*, 30(3):683–689.

Park, H.-S. and Jun, C.-H. (2009). A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Piovesan, L., Terenziani, P., and Molino, G. (2018). Glare-sscpm: an intelligent system to support the treatment of comorbid patients. *IEEE Intelligent Systems*, 33(6):37–46.

Rodriguez, A. and Laio, A. (2014). Clustering by fast search and find of density peaks. *science*, 344(6191):1492–1496.

Roemer, R., Taylor, A., and Lariviere, J. (2005). Origins of the who framework convention on tobacco control. *American Journal of Public Health*, 95(6):936–938. PMID: 15914812.

Russell, S. J. and Norvig, P. (2021). *Intelligenza artificiale: Un approccio moderno*. Pearson, Londra, 4th edition. Traduzione italiana di *Artificial Intelligence: A Modern Approach*.

Sakaue, S., Hirata, J., Kanai, M., Suzuki, K., Akiyama, M., Lai Too, C., Arayssi, T., Hammoudeh, M., Al Emadi, S., Masri, B. K., et al. (2020). Dimensionality reduction reveals fine-scale structure in the japanese population with consequences for polygenic risk prediction. *Nature communications*, 11(1):1569.

Santano-Mogena, E., Franco-Antonio, C., and Cordovilla-Guardia, S. (2023). Gender differences in susceptibility to smoking among high school students. *Journal of Advanced Nursing*, 79(5):1912–1925.

Sarkar, C., Das, B., Rawat, V. S., Wahlang, J. B., Nongpiur, A., Tiewsoh, I., Lyngdoh, N. M., Das, D., Bidarolli, M., and Sony, H. T. (2023). Artificial intelligence and machine learning technology driven modern drug discovery and development. *International Journal of Molecular Sciences*, 24(3):2026.

Seabold, S. and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*. Available at: `https://github.com/statsmodels/statsmodels`.

Shortliffe, E. H. (1977). Mycin: A knowledge-based computer program applied to infectious diseases. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 66. American Medical Informatics Association.

Song, Y., Liang, J., Lu, J., and Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing*, 251:26–34.

Sorzano, C. O. S., Vargas, J., and Montano, A. P. (2014). A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*.

Story, A., Aldridge, R. W., Smith, C. M., Garber, E., Hall, J., Ferenando, G., Possas, L., Hemming, S., Wurie, F., Luchenski, S., et al. (2019). Smartphone-enabled video-observed versus directly observed treatment for tuberculosis: a multicentre, analyst-blinded, randomised, controlled superiority trial. *The Lancet*, 393(10177):1216–1224.

Suzuki, K. (2017). Survey of deep learning applications to medical image analysis. *Med Imaging Technol*, 35(4):212–226.

Taloba, A. I., Abd El-Aziz, R. M., Alshanbari, H. M., and El-Bagoury, A.-A. H. (2022). Estimation and prediction of hospitalization and medical care costs using regression in machine learning. *Journal of Healthcare Engineering*, 2022(1):7969220.

Terenziani, P., Molino, G., and Torchio, M. (2001). A modular approach for representing and executing clinical guidelines. *Artificial intelligence in medicine*, 23(3):249–276.

Vigna-Taglianti, F. D., Galanti, M. R., Burkhart, G., Caria, M. P., Vadrucci, S., and Faggiano, F. (2014). "unplugged," a european school-based program for substance use prevention among adolescents: Overview of results from the eu-dap trial. *New directions for youth development*, 2014(141):67–82.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Vitória, P., Pereira, S. E., Muinos, G., De Vries, H., and Lima, M. L. (2020). Parents modelling, peer influence and peer selection impact on adolescent smoking behavior: A longitudinal study in two age cohorts. *Addictive behaviors*, 100:106131.

Wang, F., Casalino, L. P., and Khullar, D. (2019). Deep learning in medicine—promise, progress, and challenges. *JAMA internal medicine*, 179(3):293–294.

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

Willemsen, M. C., Mons, U., and Fernández, E. (2022). Tobacco control in europe: progress and key challenges. *Tobacco control*, 31(2):160–163.

World Health Organization (2003). *WHO framework convention on tobacco control*. World Health Organization.

World Health Organization (2020). Summary results of the global youth tobacco survey in selected countries of the who european region (2020).

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of data science*, 2:165–193.

Xu, X., Ester, M., Kriegel, H.-P., and Sander, J. (1998). A distribution-based clustering algorithm for mining in large spatial databases. In *Proceedings 14th International Conference on Data Engineering*, pages 324–331. IEEE.

Zame, W. R., Bica, I., Shen, C., Curth, A., Lee, H.-S., Bailey, S., Weatherall, J., Wright, D., Bretz, F., and van der Schaar, M. (2020). Machine learning for clinical trials in the era of covid-19. *Statistics in biopharmaceutical research*, 12(4):506–517.

Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114.

Zhavoronkov, A., Vanhaelen, Q., and Oprea, T. I. (2020). Will artificial intelligence for drug discovery impact clinical pharmacology? *Clinical Pharmacology & Therapeutics*, 107(4):780–785.

Zimmermann, A. (2008). Ensemble-trees: Leveraging ensemble power inside decision trees. In *International conference on discovery science*, pages 76–87. Springer.