UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

**DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE**

**CORSO DI LAUREA IN INGEGNERIA BIOMEDICA**

# "USO DI TECNICHE DI MACHINE LEARNING NELLA PREDIZIONE DI COMPLICANZE CARDIOVASCOLARI DEL DIABETE DI TIPO 2"

**Relatore: Prof. GIOVANNI SPARACINO**

**Laureando/a: IEGOR TOPOROV**

**ANNO ACCADEMICO 2023 – 2024**

**Data di laurea 21/11/2023**

# Abstract

Nel contesto clinico del diabete di Tipo 2, una delle malattie metaboliche più diffuse a livello mondiale, le complicanze cardiovascolari rappresentano una delle sfide più significative per la salute pubblica. Questo scenario ha spinto la comunità medica a esplorare nuovi approcci, tra cui l'applicazione dell'intelligenza artificiale nel campo della medicina, nota come machine learning. Questo elaborato si focalizza sulle complicanze cardiovascolari che spesso ne derivano e specificatamente su come si possono trovare soluzioni efficaci per gestire questa crescente epidemia usando opportune metodologie di machine learning in medicina.

# Sommario

# 1. Diabete di Tipo 2 (T2D) e sue complicanze cardiovascolari

## 1.1. Prevalenza e impatto sulla salute pubblica

Il diabete di tipo 2 (T2D) è la forma più comune di diabete e colpisce oltre il 90 percento delle persone affette da diabete. È una patologia diffusa in tutto il mondo e la sua incidenza sta aumentando costantemente con stime che prevedono entro il 2030 quasi 600 milioni di casi nel mondo. Secondo la classificazione ufficiale, il T2D si verifica quando il corpo non produce abbastanza insulina o non risponde adeguatamente all'insulina prodotta. Questo difetto insulinico si sviluppa su una base di resistenza all'insulina, che significa che il corpo non risponde efficacemente all'azione dell'insulina, particolarmente nel fegato, nei muscoli e nel tessuto adiposo. Questo processo può peggiorare nel tempo, portando a problemi di gestione del glucosio nel sangue. La causa del T2D rimane ancora sconosciuta, ma è riconosciuto come un problema complesso influenzato da molteplici fattori genetici e ambientali. Non è considerato come una singola malattia, ma piuttosto come un insieme di diverse condizioni. Il rischio di sviluppare questa patologia aumenta con l'età, in genere si manifesta negli adulti sopra i 30-40 anni, soprattutto se associato all'obesità e alla mancanza di attività fisica. Tuttavia, T2D può anche colpire bambini e adolescenti. La familiarità sembra svolgere un ruolo importante; circa il 40 percento dei pazienti ha parenti di primo grado, come genitori o fratelli, affetti dalla stessa malattia. Il T2D spesso non mostra segni evidenti per molti anni, poiché l'incremento graduale dei livelli di zucchero nel sangue, chiamato iperglicemia, non è inizialmente abbastanza grave da causare i sintomi caratteristici della malattia, tra cui stanchezza, aumento della sete, aumento della diuresi, perdita di peso non ricercata, malessere generale. Nei casi più gravi le maggiori complicanze possono arrecare al paziente danni anche importanti a livello neurologico (neuropatia), renale (nefropatia), oculare (retinopatia) e cardio-cerebrovascolare (ictus, malattia coronarica, arteriopatia degli arti inferiori) [5]. In Italia, circa il 6 percento della popolazione, ovvero quasi 4 milioni di persone, sono affette da T2D. Tuttavia, si stima che potrebbero esserci circa 1,5 milioni di persone con questa malattia che non sono ancora consapevoli della loro condizione. La prevalenza del diabete aumenta con l'età, raggiungendo il 21 percento nelle persone di 75 anni o più (dati ISTAT 2020). A livello globale, nel 2021, l'International Diabetes Federation (IDF) ha calcolato che oltre 530 milioni di persone nel mondo tra i 20 e i 79 anni sono affette da T2D [14]. Solo negli Stati Uniti, circa 35 milioni di persone sono affette da diabete, tra le quali il rischio di mortalità maggiore è causato dalle malattie cardiovascolari [7]. Altri 57 milioni di persone presentano una forma di pre-diabete, con livelli elevati di glucosio nel sangue che aumentano il loro rischio di sviluppare diabete, malattie cardiache e ictus [27].

## 1.2 Complicanze cardiovascolari associate al T2D

Le persone affette da diabete hanno da 2 a 4 volte più probabilità di sviluppare malattie cardiovascolari rispetto alla popolazione generale. In realtà, nei paesi industrializzati, le malattie cardiovascolari costituiscono la principale causa di morte nei pazienti con diabete mellito, creando un ciclo pericoloso di rischi reciproci e crescenti. Le principali complicanze cardiovascolari del diabete comprendono: cardiopatia ischemica, spesso complicata da scompenso cardiaco; ictus ischemico e/o emorragico. Livelli elevati di glucosio e insulina non utilizzata nel sangue, insieme a fattori di rischio aggiuntivi come dislipidemia, ipertensione e obesità, contribuiscono allo sviluppo di: disfunzione endoteliale, che coinvolge il tessuto che riveste l'interno dei vasi sanguigni; aterosclerosi precoce e veloce, che solitamente colpisce le arterie coronarie, responsabili del rifornimento di sangue al muscolo cardiaco. Le arterie coronarie, di piccolo calibro e tortuose, sono particolarmente suscettibili alle alterazioni metaboliche e al rischio di complicanze trombotiche associate al diabete. Queste complicazioni complicano le opzioni terapeutiche per le malattie coronariche e la conseguente cardiopatia ischemica. La complessità delle condizioni e l'alto rischio di inefficacia delle procedure di rivascolarizzazione, sia chirurgiche che percutanee, spesso portano allo sviluppo di scompenso cardiaco avanzato. Si stima che il 15-25 percento dei pazienti affetti da scompenso cardiaco sia anche affetto da diabete. Inoltre, una complicanza vascolare precoce è l'arteriopatia obliterata agli arti inferiori, che può causare ischemia e portare a ulcere al piede, gangrena e un significativo rischio di amputazione. Sul fronte cerebrovascolare, il diabete aumenta notevolmente il rischio di ictus ischemico, recidiva di ictus e deterioramento cognitivo [4].

## 1.3 Utilità del machine learning nella predizione delle complicanze cardiovascolari del T2D e scopo dell'elaborato

Lo scopo dell'elaborato è di indagare sulle tecniche e sui risultati attuali degli strumenti di apprendimento automatico nella predizione di complicanze cardiovascolari in casi di diabete di tipo 2. In un contesto di grande crescita di interesse verso i modelli predittivi del machine learning in ambito medico, dunque, il capitolo secondo cercherà di esplorare i concetti fondamentali di questa tecnologia rivoluzionaria  discutendo i vantaggi e gli svantaggi delle sue applicazioni nella diagnostica medica. Si approfondiranno le potenzialità di questa tecnologia nel migliorare la diagnosi e la gestione delle malattie, tra cui il diabete di Tipo 2 e le sue complicanze cardiovascolari saranno protagoniste nel terzo capitolo. Nella terza sezione, si esaminerà una metodologia specifica applicata a un caso di studio presente nella letteratura scientifica. Si analizzerà l'origine e la natura dei dati

clinici utilizzati nello studio, si esplorerà come siano state selezionate le codifiche ICD (International Classification of Diseases) e come sia stata creata una Disease Network per comprendere meglio le relazioni tra le condizioni di salute dei pazienti. Inoltre, si discuterà le tecniche di machine learning impiegate nel caso di studio, focalizzandosi sulla selezione delle caratteristiche (feature selection) e sugli score utilizzati per valutare le prestazioni dei modelli. A seguire, si entrerà nel cuore dell'analisi dei dati, esaminando il processo di sviluppo dei modelli di machine learning e valutando le loro prestazioni sia attraverso il training set mediante la cross-validation, sia utilizzando un dataset di test indipendente. Si analizzeranno anche le statistiche della ROC Curve (AUC) e l'importanza delle caratteristiche nel contesto del caso di studio esaminato. Si esplorerà come i modelli si comportano nella pratica, mettendo in evidenza le sfide e le opportunità riscontrate durante l'analisi dei dati del caso di studio. Si condurrà una discussione sul caso di studio esaminato analizzando l'importanza delle caratteristiche come età e genere nel contesto delle previsioni e si compareranno i risultati ottenuti con altri studi presenti in letteratura. Si esplorerà anche le limitazioni delle tecniche di machine learning nel contesto specifico del caso di studio, gettando le basi per una breve riflessione sul futuro promettente del machine learning in medicina e sulle sfide che ancora dovranno essere superate.

# 2. Machine Learning e esempi di applicazione in Medicina

## 2.1 Concetti fondamentali di machine learning

Esistono alcune differenze nelle definizioni disponibili degli algoritmi di machine learning. Questi algoritmi sono generalmente suddivisi in categorie in base al risultato ricercato, per cui la classificazione generale può essere la seguente.

- *Supervised Learning*: in questo contesto, esperti umani fungono da insegnanti, ovvero forniscono al computer dati di addestramento contenenti gli input (o *predictors*) e le relative risposte corrette (*output*). Utilizzando queste informazioni, il computer dovrebbe essere in grado di apprendere i modelli sottostanti. Il processo di apprendimento dai dati e l'utilizzo delle conoscenze acquisite per orientare decisioni future sono estremamente potenti. Le tecniche di apprendimento supervisionato ampiamente utilizzate includono algoritmi di classificazione e regressione.

    - *Classificazione*: in questo tipo di apprendimento, il computer impara a categorizzare dati in gruppi definiti. Ad esempio, Gmail organizza le email in categorie come social, promozioni o aggiornamenti. Può separare dati in due gruppi (binari) come sì o no, o in più gruppi (multiclasse) come diverse categorie di email.

    - *Regressione*: l'obiettivo è far imparare al computer a prevedere valori continui. Più il computer è in grado di prevedere un valore vicino al risultato reale, migliore è il modello. L'accuratezza del modello viene misurata calcolando l'errore: più l'errore è basso, più preciso è il modello di previsione.

- *Unsupervised Learning*: in questo approccio, un modello di apprendimento automatico viene addestrato utilizzando dati non etichettati e senza una destinazione specifica. Il computer può scoprire nuove informazioni sui modelli nei dati che gli esseri umani potrebbero non aver notato. Questi modelli sono particolarmente utili quando gli esperti umani non sanno cosa cercare nei dati. Tali algoritmi cercano di trovare regole, modelli, riassunti e gruppi nei dati di input, aiutando così a ottenere comprensioni profonde e a rendere i dati più significativi per gli utenti finali. Alcuni algoritmi comuni di apprendimento non supervisionato includono K-

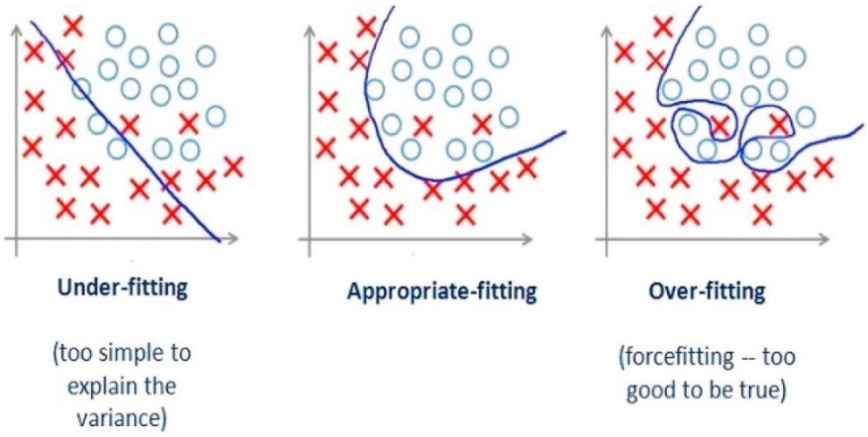
Means, Self-Organiying Maps (SOM), Autoencoder e Generative adversarial network (GAN), tra molti altri.

- *Semi-supervised Learning*: nei due precedenti paradigmi, o non c'erano etichette per tutte le osservazioni o erano presenti etichette per tutte. L'apprendimento semi-supervisionato si colloca tra questi due scenari. In molte situazioni pratiche, etichettare i dati richiede tempo e risorse considerevoli. Quindi, quando la maggior parte dei dati non ha etichette, gli algoritmi semi-supervisionati sono la scelta migliore per creare dei modelli. Questi algoritmi sfruttano l'idea che anche i dati senza etichette contengono informazioni preziose sulle caratteristiche di gruppo, anche se non conosciamo l'appartenenza precisa del gruppo per ciascun dato non etichettato.

- *Reinforcement Learning*: il reinforcement learning si propone di utilizzare le informazioni raccolte interagendo con un ambiente simulato tramite computer per imparare ad eseguire azioni che massimizzino le ricompense o minimizzino i rischi. Gli algoritmi di apprendimento per rinforzo (chiamati agenti) apprendono costantemente dall'ambiente in un processo iterativo. Durante questo processo, un agente impara dalla sua esperienza fino a esplorare tutte le possibili situazioni. Questo approccio consente a macchine e software di determinare automaticamente il comportamento ideale in uno specifico contesto, al fine di massimizzare le prestazioni. Per far apprendere all'agente, è necessario un feedback semplice sotto forma di ricompensa; ciò è chiamato segnale di rinforzo. Alcune applicazioni degli algoritmi di apprendimento per rinforzo comprendono giochi da tavolo giocati al computer (come scacchi, Go), mani robotiche e auto a guida autonoma [22].

In ambito medico vengono utilizzati maggiormente algoritmi supervised, e per capire ancora meglio il training process bisogna definire i seguenti aspetti tecnici. Per generalizzazione si intende quanto bene un modello di machine learning può applicare ciò che ha imparato da dati di training specifici a nuovi esempi non visti durante l'addestramento. Un buon modello di machine learning dovrebbe essere in grado di applicare le sue conoscenze a situazioni mai incontrate prima. Il problema principale si verifica quando un modello è eccessivamente specializzato (*overfitting*) o troppo semplice (*underfitting*) rispetto ai dati forniti. L'overfitting si verifica quando il modello impara non solo i modelli importanti nei dati, ma anche il rumore o le variazioni casuali che potrebbero non essere presenti nei nuovi dati. Questo può danneggiare la capacità del modello di fare previsioni accurate su

nuovi dati, poiché può essere influenzato da informazioni irrilevanti apprese durante l'addestramento. L'underfitting si verifica quando il modello non riesce né a rappresentare adeguatamente i dati di allenamento né a generalizzare su nuovi dati. Un modello sottoadattato è inefficace a causa delle sue prestazioni scadenti sui dati di allenamento. Questo problema spesso non viene discusso tanto quanto il sovrallenamento, poiché è facile rilevarlo grazie a buone metriche di valutazione. La soluzione consiste nel cercare approcci di apprendimento automatico alternativi. Tuttavia, l'underfitting offre un contrasto importante rispetto al problema del sovrallenamento. Entrambi i fenomeni, sia il sovrallenamento sia l'underfitting, possono portare a prestazioni del modello scarse. Tuttavia, il sovrallenamento è il problema più comune nell'implementazione pratica del machine learning. Questo è critico perché valutare le prestazioni del modello sui dati di addestramento è diverso dal valutarle su dati mai visti prima, che è essenzialmente il motivo per cui creiamo i modelli in primo luogo [22].




Differenza tra underfitting e overfitting, con esempio di fit utile [22].

L'elemento chiave per creare un modello di machine learning è il dataset. Il dataset viene diviso in tre parti, ognuna con un compito preciso. Il Dataset di Addestramento (*Training Set*) svolge un ruolo cruciale nel processo di apprendimento del modello. È la palestra del nostro modello, dove impara dai dati per comprendere le relazioni e i pattern. In altre parole, questi dati fungono da insegnanti per il modello, consentendogli di acquisire conoscenze e adattarsi alle peculiarità dei dati. Il Dataset di Validazione (*Validation Set*) non è coinvolto nell'addestramento diretto del modello, ma svolge un ruolo cruciale nel perfezionare le sue prestazioni. Utilizzando il dataset di validazione, possiamo regolare le impostazioni del modello, chiamate iperparametri, per trovare la configurazione ottimale. È come un campo di prova che ci permette di esplorare diverse opzioni e assicurarci che il nostro modello sia ben calibrato per affrontare nuovi compiti. Una volta che il modello è stato addestrato e ottimizzato utilizzando il dataset di addestramento e quello di validazione, viene sottoposto al verdetto finale: il dataset di test (*Test Set*). Questo insieme di dati è completamente

separato da quelli utilizzati per l'addestramento e la validazione. Serve come una sorta di esame finale per il nostro modello. Solo quando supera questo esame, dimostrando di poter fare previsioni accurate e generalizzare bene su nuovi dati, possiamo dire che il nostro modello è pronto per l'uso in situazioni del mondo reale. Per evitare overfitting il metodo più utilizzato è il K-fold cross-validation [22].

Un ultimo sguardo importante va dato alle tecniche di valutazione dei modelli sulla base dei risultati ottenuti. Le metriche di valutazione misurano le prestazioni di un modello specifico, aiutando nella selezione ottimale del modello nella fase di test. Diverse metriche possono riflettere obiettivi diversi quando si progetta un determinato modello. Ad esempio, ci sono diverse metriche per i compiti di classificazione, regressione, ranking, clustering, topic modelling e altri. Alcune metriche, come la *precision* e il *recall*, sono utili per più compiti. La scelta delle metriche di valutazione influisce fortemente sulla capacità del modello di apprendere dai dati. Pertanto, in base alle caratteristiche del dataset e alla formulazione del problema, viene effettuata la scelta della metrica di valutazione. Qua si analizzano brevemente le metriche più usate. Qui e in tutta la trattazione si indicheranno TP (True Positives), FP (False Positives), TN (True Negatives), FN (False Negatives).

- *Model Accuracy*: si può definire come il rapporto tra i campioni correttamente classificati e il numero
  totale di predizioni fatte.

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

- *Precision* e *Recall*: in task di classificazione, si tratta dei due seguenti rapporti:

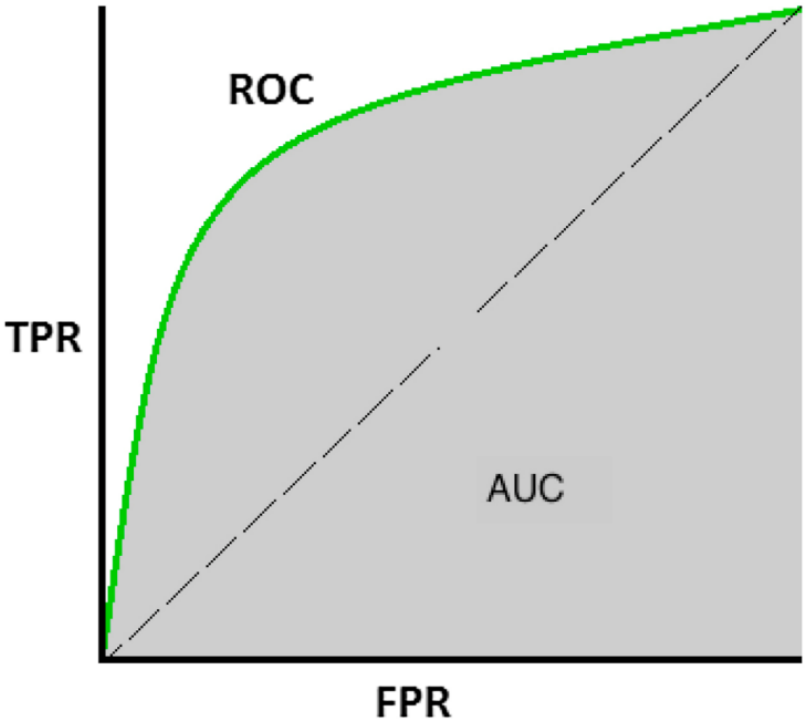
$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

  Più semplicemente, un'alta precision indica che l'algoritmo restituisce significativamente di più i risultati rilevanti piuttosto che quelli irrilevanti, e un alto recall indica che l'algoritmo ha restituito la maggior parte dei risultati rilevanti.
- *F1 Score*: si tratta della media armonica della precision e del recall, con massimo valore 1 e minimo 0.

$$F1 = \frac{2 \times (Precision \times Recall)}{Precision + Recall}$$

- *ROC Curve* e *AUC (Receiver Operating Characteristic Curve* e *Area Under the Curve)*: una curva ROC è un grafico che illustra la capacità diagnostica di un sistema di classificazione binaria variando la soglia di discriminazione. La curva ROC viene creata rappresentando il tasso di veri positivi (TPR) rispetto al tasso di falsi positivi (FPR) in diverse impostazioni di soglia ed offre una rappresentazione visuale delle prestazioni del classificatore. Per confrontare diversi classificatori, è possibile ridurre le prestazioni della curva ROC a un unico valore scalare rappresentante le prestazioni attese. Un metodo comune consiste nel calcolare l'area sotto la curva ROC, abbreviata come AUC. Un modello eccellente ha un AUC vicino a 1, indicando una buona capacità di separare le classi. Un modello scadente ha un AUC vicino a 0, segno di una scarsa capacità di separazione. Un AUC pari a zero indica che il modello sta invertendo i risultati, prendendo 0 come 1 e viceversa. Infine, un AUC di 0,5 indica che il modello non ha alcuna capacità di separare le classi [22].




Relazione tra ROC e AUC. ROC è una curva di probabilità e AUC rappresenta il grado di separabilità. Ci dice quanto è in grado un modello di distinguere le classi. *TPR*: True Positives Rate; *FPR*: False Positives Rate [22].

## 2.2 Alcune applicazioni di machine learning nella diagnostica medica

La diagnosi medica è un compito complesso che è ampiamente considerato come un processo empirico, ma è scarsamente compreso come un compito cognitivo. Quindi, per quanto possa sembrare complessa, la diagnosi delle malattie tramite l'uso di un computer, come nel nostro caso mediante l'apprendimento automatico, è suddivisa in diverse fasi. Il primo passaggio nella diagnosi di una malattia è l'acquisizione dei dati. Questi dati possono presentarsi in forme diverse, tra cui interviste mediche, dati clinici, demografici, immagini, registrazioni vocali, dati storici del paziente o persino suoni cardiaci. Il passo successivo coinvolge l'elaborazione dei dati. Durante questa fase, i dati vengono preparati, gestendo i valori mancanti, riducendo la dimensionalità, affrontando dati rumorosi, e così via. Successivamente, vengono identificate le variabili target e le variabili predittive. Questi dati vengono quindi alimentati in uno dei modelli per l'addestramento. Una volta che il modello è stato addestrato, viene utilizzato per la diagnosi. Le limitazioni dovute alla presenza di numerose strutture sovrapposte, casi complessi, distrazioni, stanchezza e ai limiti del sistema visivo umano rendono preziosa l'introduzione di una 'seconda opinione'. Questa necessità ha favorito l'adozione dei sistemi CAD (Computer-Aided Design) nei processi di diagnosi. Il CAD è un concetto che assegna ruoli paritari sia ai medici che ai computer, assistendo i medici nel prendere le migliori decisioni/pratiche cliniche. Inoltre, a causa dell'incremento della complessità dei pazienti, degli errori diagnostici elevati e dell'abbondanza di dati disponibili, i sistemi EHR vengono utilizzati per assistere nella presa di decisioni cliniche [12].

Weng et al. hanno utilizzato i dati cardiaci di carattere clinico acquisiti elettronicamente di 370.000 pazienti in cura nel Regno Unito per costruire modelli di apprendimento automatico per prevedere l'insorgenza di malattie cardiovascolari (CVD) [19]. Hanno confrontato quattro metodi di apprendimento automatico (regressione logistica, random forest, alberi decisionali potenziati con il gradiente e rete neurale) con l'algoritmo stabilito dalle linee guida dell'ACC per prevedere l'insorgenza di CVD nel corso di un periodo di 10 anni. I loro risultati hanno mostrato che gli algoritmi di machine learning hanno superato l'algoritmo di previsione del rischio ACC/AHA stabilito, in termini di metriche di discriminazione del modello. In alcuni casi, le caratteristiche che non erano presenti negli algoritmi esistenti sono state pesate fortemente dall'algoritmo di machine learning. Inoltre, la mancanza di dati nei record elettronici della salute è stata rivelatrice per prevedere l'outcome. Ciò suggerisce che gli algoritmi di apprendimento automatico sono stati in grado di individuare modelli nuovi nei dati. Altre ricerche hanno integrato i dati clinici con dati biologici. Halim et al. hanno investigato quali dei 53 biomarcatori proteici precedentemente segnalati fossero più predittivi delle CVD [20]. Hanno addestrato vari modelli di Logistic Regression, con diversi

numeri di variabili cliniche, per individuare i biomarcatori più importanti per prevedere la morte o l'infarto miocardico. Utilizzando metodi di machine learning, sono stati in grado di identificare un pannello più ristretto di proteine rilevanti che possono essere studiate per lo sviluppo di terapie aggiuntive per la cura cardiovascolare. Un'altra applicazione dell'apprendimento automatico nella rilevazione delle malattie coronariche è stata nel supportare la selezione dei pazienti per le immagini diagnostiche non invasive. Al'Aref et al. hanno esplorato se un modello di apprendimento automatico possa migliorare la stima della probabilità pre-test di un paziente per la malattia coronarica ostruttiva (CAD) e quindi aumentare la resa diagnostica delle modalità di imaging non invasive come l'angiografia coronarica tramite tomografia computerizzata [21].

Grazie alla disponibilità di strumenti intelligenti per l'analisi dei dati, i metodi di ML contribuiscono a svelare relazioni interessanti nei dati. Come seconda opinione, possono confermare o contestare le decisioni dei clinici. L'integrazione di strumenti basati su ML che monitorano costantemente un crescente flusso di dati alla ricerca di pattern, assistendo nella presa di decisioni per i clinici o regolando automaticamente le impostazioni dei dispositivi a letto, ha migliorato i risultati del trattamento dei pazienti e ridotto in modo significativo i costi globali del trattamento. Attraverso il machine learning, è possibile sviluppare modelli predittivi personalizzati in grado di anticipare le possibili complicazioni post-operatorie, come la predisposizione alle infezioni, basandosi sulle caratteristiche specifiche del paziente. L'intelligenza artificiale trova impiego anche nella diagnostica per immagini, come ad esempio in radiologia, e nella ricerca farmacologica e negli studi epidemiologici, contribuendo a individuare nuove terapie e a identificare i fattori di rischio per le malattie. La medicina di precisione e la robotica sono altri due settori in cui l'intelligenza artificiale sta rivoluzionando la pratica medica. Nel campo dell'oncologia, l'intelligenza artificiale può fungere da sistema di supporto decisionale: gli algoritmi aiutano i medici nell'individuare la terapia più appropriata per un paziente, considerando l'efficacia dei trattamenti in casi simili e valutando gli eventuali effetti collaterali [3].

Tuttavia, sebbene il machine learning prometta la miglior assistenza clinica, finora non ha dimostrato appieno la sua utilità, probabilmente a causa dell'opacità, ovvero la chiarezza, negli algoritmi e nell'analisi dei modelli di machine learning. Inoltre, la qualità dei dati e la generalizzazione dei modelli di ML rimangono tra le sfide ancora irrisolte [12]. Nonostante tutto ciò, il settore del machine learning e dell'intelligenza artificiale in medicina sta attraversando un periodo di grande interesse e il tasso di crescita annuo raggiungerà presto un valore del 42,2 percento. Questo dato impressionante riflette il grandissimo potenziale di questi strumenti nella sanità e il crescente interesse di investitori e professionisti [3].

# 3. Un caso di studio di letteratura: predizione del rischio di CVD in pazienti con T2D

Per questa analisi di predizione di complicanze cardiovascolari del diabete di tipo 2 tramite machine learning viene utilizzato lo studio condotto in [15].

## 3.1 Data base esaminato

Questo studio ha utilizzato un set di dati amministrativi raccolti nell'intervallo di anni dal 1995 al 2018 proveniente dalla health fund CBHS in Australia per sviluppare un modello di previsione del rischio. Il set di dati comprendeva informazioni mediche di circa 124.000 pazienti de-identificati che erano membri di quella health fund durante quel periodo indicato. Ciascuna cartella clinica conteneva un ID paziente univoco, l'età, il sesso, le date di ammissione e dimissione e i codici di malattia. I codici di malattia in questo set di dati erano definiti secondo la Classificazione Internazionale delle Malattie 9a e 10a *Australian Modification* (ICD-9-AM e ICD-10-AM). I codici di malattia in ogni ricovero riflettevano i problemi di salute del paziente durante quel determinato episodio. Nel dataset, i codici di malattia (ICD) venivano utilizzati per identificare i pazienti affetti da diabete di tipo 2 (T2D) e malattie cardiovascolari (CVD), poiché lo scopo era prevedere il rischio di CVD nei pazienti con T2D. Per creare il set di dati di ricerca, sono stati applicati diversi criteri di filtro al dataset iniziale, tra cui la selezione di pazienti con almeno un episodio di ricovero, l'uso di codici di malattia validi, l'eliminazione dei dati duplicati e l'esclusione di codici di malattia troppo generici, come lesioni normali, raffreddori e influenza, ecc. Successivamente, sono stati selezionati due gruppi di pazienti per studiare il rischio di CVD nei pazienti con T2D. Il primo gruppo, chiamato CT2D&CVD, comprendeva pazienti che avevano prima ricevuto una diagnosi positiva di T2D e successivamente una diagnosi positiva di CVD. Il secondo gruppo, chiamato CT2D, includeva pazienti con diagnosi positiva solo di T2D. CVD e T2D avevano i loro codici specifici nei sistemi di codifica ICD e per selezionare i pazienti dello studio è stata cercata la presenza di almeno un codice ICD della malattia nei dati dell'episodio di ricovero. Per identificare i pazienti con CVD, sono stati utilizzati i codici ICD per cinque diverse malattie specifiche. La tabella seguente fornisce bene una lista dei codici ICD cercati nel dataset originale [15].

| Comorbidities | ICD-10-AM Codes | ICD-9-AM Codes |
|---|---|---|
| Congestive Heart Failure | I09.9, I1.0, I13.0, I13.2, I25.5, I42.0, I42.5–I42.9, 143.x, 150.x, P29.0 | 398.91, 402.11, 402.91, 404.11, 404.13, 404.91, 404.93, 428.x |
| Cardiac Arrhythmias | I44.1–I44.3, I45.6, I45.9, I47.x, R00.0, Roo.1, R00.8, T82.1, Z45.0, Z95.0 | 426.10, 426.11, 426.13, 426.2–426.53, 426.6–426.28, 427.0, 427.2427.31, 427.60, 427.9, 785.0, V45.0, V53.3 |
| Valvular Disease | A52.0, I05.x–108.x, I09.1, I09.8, I34.x–I39.x, Q23.0–Q23.3, Z95.2–Z95.4 | 093.2, 394.0–397.1, 424.0–424.91, 746.3–746.6, V42.2, V43.3 |
| Pulmonary Circulation Disorders | I26.x, I27.x, I28.0, I28.8, I28.9 | 416.x, 417.9 |
| Peripheral Vascular Disorders | I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9 | 440.x, 441.2, 441.4, 441.7, 441.9, 443.1–443.9, 447.1, 557.1, 557.9, V43.4 |
| Type 2 Diabetes | E11.0, E11.1, E11.2–E11.9 | 250.0–250.3, 250.4–250.7, 250.9 |

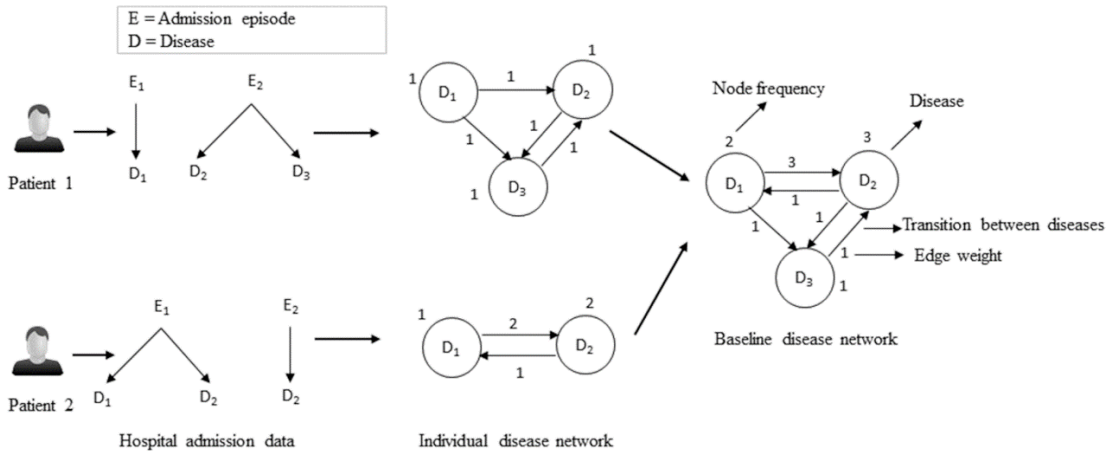
Tabella indicante i raggruppamenti in comorbilità e tutti i relativi codici ICD-9-AM e ICD-10-AM di CVD e T2D [15].

### 3.2 Selezione ICD e creazione *Disease Network* nel caso di studio esaminato

Come già accennato in precedenza, i codici di malattia nei dati di ammissione dei pazienti sono stati registrati sia nel formato ICD-9-AM che nel formato ICD-10-AM. La decisione su quali malattie e relativi codici ICD analizzare rappresenta una delle principali sfide nello sviluppo dei modelli di previsione del rischio. Ciascuna coorte comprende circa 1000 codici ICD diversi. Se fossero stati considerati tutti i codici ICD disponibili, l'analisi della rete di malattie (*Disease Network*) sarebbe diventata troppo complessa. Per semplificare la rete, i codici ICD sono stati raggruppati in comorbilità, riducendo così il numero di nodi (cioè, codici ICD) nella rete. Pertanto, è stata creata una tabella di mappatura ICD per associare ogni comorbilità ai corrispondenti codici ICD. Nella letteratura scientifica, esistono diversi indici di comorbilità (Charlson et al., 1987; Elixhauser et al., 1998; Deyo, Cherkin & Ciol 1992). E' stato utilizzato l'indice di comorbilità di Elixhauser per generare l'elenco delle comorbilità e la tabella di mappatura, poiché questo indice è stato proposto per misurare la comorbilità basandosi sui dati amministrativi. L'indice di comorbilità adattato di Elixhauser (Garland et al., 2012) comprendeva 31 comorbilità. Sono state escluse le comorbilità relative al diabete di tipo 2 (in totale 2) e alle malattie cardiovascolari (in totale 5) poiché lo studio aveva già suddiviso il dataset in base alla presenza di T2D e CVD. I codici ICD di 24 comorbilità su 31 sono stati mappati con la tabella di traduzione ottenuta dallo studio di H. Quan et al. (2005) [9]. Questi codici ICD nella tabella di traduzione erano presenti sia nella versione ICD-9 che in versione ICD-10, e sono stati successivamente verificati manualmente con le versioni ICD-9-AM e ICD-10-

AM del dataset di questo studio. Poiché i codici ICD-9 e ICD-10 nella tabella di traduzione corrispondevano ai corrispondenti codici ICD-9-AM e ICD-10-AM del dataset di questo lavoro, non è stato necessario apportare ulteriori modifiche [15].

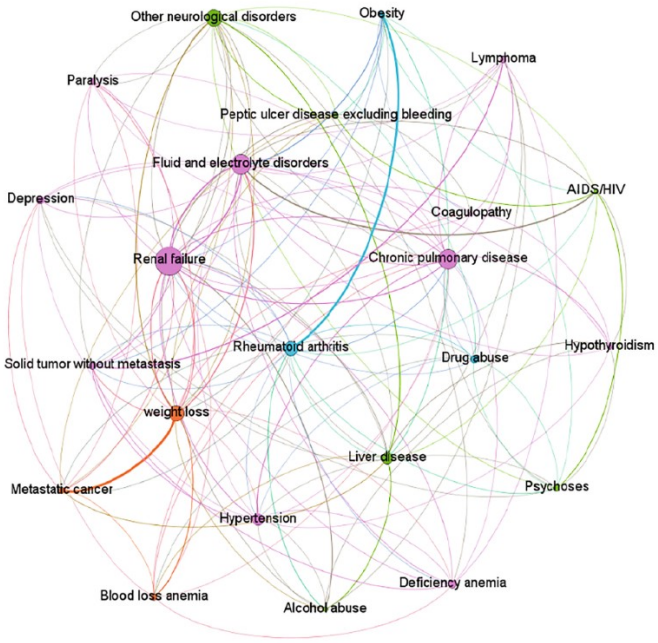
Ciò che si indica con traiettoria di salute dei pazienti serve a mostrare la transizione delle malattie durante le successive ammissioni nel tempo. La prima rete di malattie, chiamata rete di malattie individuali (*disease network*), rappresenta la traiettoria di salute di un singolo paziente. Questa rete è costituita da nodi e archi, dove ciascun nodo rappresenta una malattia e l'arco tra due nodi indica l'occorrenza successiva di queste due malattie. Nella rete, l'arco tra due nodi indica che un paziente è passato da una malattia cronica all'altra durante le successive ammissioni. Due attributi chiave di questa rete sono la frequenza del nodo, che indica la prevalenza delle malattie riscontrate da un paziente considerando tutte le ammissioni, e il peso dell'arco, che indica il numero di volte in cui due malattie sono occorse simultaneamente o in ammissioni consecutive. Successivamente, le reti di malattie individuali per i pazienti delle varie coorti vengono fuse per creare un'altra rete di malattie chiamata rete di malattie di base. Sono state create due di queste reti dalle due coorti (cioè, $C_{T2D\&CVD}$ e $C_{T2D}$). La prima rete di malattie di base, denominata $BN_{T2D\&CVD}$, è stata creata dalla coorte $C_{T2D\&CVD}$. L'altra rete di malattie di base, chiamata $BN_{T2D}$, è stata creata dalla coorte $C_{T2D}$. Ciascuna di queste reti mostra le traiettorie di salute delle rispettive coorti di pazienti. I nodi e gli archi della rete di malattie di base sono calcolati sommando i nodi e gli archi delle reti di malattie individuali delle coorti corrispondenti. La successiva illustra il processo completo di generazione della rete di malattie di base [15].




Processo di generazione del disease network di base [15].

Come si nota nel diagramma soprastante, ogni paziente porta con sé nodi (indicati con D = Disease) ed archi (occorrenze di comorbilità). Più è ricco il dataset di pazienti, e più preciso è il

disease network che è possibile ricavarne. La mappa risultante dalle due mappe dei pazienti racchiude tutte le informazioni dei due diagrammi di base "sommate" per occorrenza (occorrenze: D1: 2 volte; D2: 3 volte; D3: 1 volta). Se si avesse una mappa contenente dati di 120 mila pazienti, si avrebbero informazioni precise a livello statistico sull'occorrenza di determinati ICD in relazione con altri. Nella valutazione del rischio di CVD a partire da T2D, avere a disposizione questo genere di dati da mandare in input a un solido modello di machine learning, supervisionato o no, può essere di grande aiuto per trovare pattern importanti [15].




Esempio di disease network finale. Le dimensioni dei nodi sono proporzionali alla prevalenza delle corrispondenti condizioni di comorbilità. Nodi dello stesso colore appartengono allo stesso cluster/community [15].

### 3.3 Feature selection e score utilizzati nel caso di studio esaminato

Questo studio ha selezionato cinque feature da due diverse categorie: demografiche e di rete di malattie. Due delle caratteristiche demografiche, ovvero l'età e il sesso, sono considerate fattori di rischio per CVD nei pazienti con T2D. Le altre tre caratteristiche (ossia, *network node*, *network edge* e *cluster match score*) sono calcolate dal disease network finale. Nonostante ci siano altre misure derivanti dal disease network utilizzate nell'analisi dello stesso, in questo studio ci si è concentrati solo su queste tre caratteristiche. Ciò è dovuto al fatto che le sequenze di comparsa delle malattie e le loro frequenze sono cruciali per esplorare le comorbilità delle malattie, e altre caratteristiche di rete (come *core-periphery*, *closeness* e *betweenness*) non riescono a catturare molti di questi dettagli. Le tre caratteristiche basate sulla rete rappresentano il grado di somiglianza che un paziente di test ha in

termini di presenza di malattie, progressione e cluster di malattie presenti nella rete di malattie finale. Queste caratteristiche possono essere utilizzate come potenziali predittori per la valutazione del rischio dei pazienti con T2D riguardo alla progressione verso le CVD [15].

Il punteggio di rischio per la caratteristica dell'età ($F_{age}$) è assegnato da 0 a 1. Per ottenere questi punteggi, l'età (in anni) dei pazienti viene divisa per la differenza tra l'età minima e massima nella popolazione dello studio. Il punteggio di rischio per la caratteristica del sesso ($F_{sex}$) è assegnato 1 per i maschi e 0 per le femmine. Le caratteristiche selezionate estratte dalla rete di malattie includono il punteggio di corrispondenza del nodo di rete (*network node match score*), il punteggio di corrispondenza del bordo di rete (*network edge match score*) e il punteggio di corrispondenza del cluster di rete (*network cluster match score*). Il network node match score viene calcolato confrontando la rete di malattie individuale del paziente di test ($N_{test}$) con la rete di malattie finale ($N_{FD}$) in termini di comparsa delle malattie. Un punteggio alto indica che il paziente di test è stato diagnosticato con più malattie presenti sia in $N_{FD}$ che in $N_{test}$, e che queste malattie sono più frequenti in entrambe le reti. La caratteristica basata sul nodo ($F_{node}$) è definita come segue. Tutte le notazioni e i nomi delle variabili sono stati riportati come nello studio da cui è citata la ricerca.

Imponendo:

$match_score\ (node) = \sum_{i=1}^{N} freq(d_{iN_{test}}) * freq(d_{iN_{FD}})$, con $d_{iN_{test}} = d_{iN_F}$ per $i = 1,2,3, \dots, N$

e anche:

$total_preval\ (node) = \sum_{i=1}^{N_1} freq(d_{iN_{test}})$, per $i = 1,2,3, \dots, N_1$

La caratteristica sul nodo è: $F_{node_{N_{test}}} = \frac{match_score\ (node)}{total_preval\ (node)}$.

Qui, $d_i$ indica un nodo di rete. N indica il numero totale di nodi in comune (o malattie) tra $N_{FD}$ e $N_{test}$. $N_1$ è il numero totale di malattie nel $N_{test}$. Il $match_score\ (node)$ è determinato moltiplicando il numero totale di malattie comuni tra $N_{FD}$ e $N_{test}$ e sommandole, mentre il $total_preval\ (node)$ rappresenta la somma complessiva della prevalenza delle malattie in $N_{test}$.

Allo stesso modo, il edge match score è calcolato confrontando la rete di malattie individuale del paziente di test ($N_{test}$) con la rete di malattie finale ($N_{FD}$) in termini di prevalenza dei bordi (ossia, transizione tra malattie). La caratteristica basata sul bordo ($F_{edge}$) è definita come segue.

Imponendo:

$match_score\ (edge) = \sum_{i=1, j=1, i\neq j}^{N} freq\left(e_{(d_i,d_j)N_{test}}\right) * freq\left(e_{(d_i,d_j)N_{FD}}\right)$, con $e_{(d_i,d_j)N_{test}} = e_{(d_i,d_j)N_{FD}}$ per $i, j = 1,2,3, \dots, N$

e anche:

$total_preval\ (edge) = \sum_{i=1, j=1, i\neq j}^{P} freq\left(e_{(d_i,d_j)N_{test}}\right)$, per $i, j = 1,2,3, \dots, P$

La caratteristica sul bordo è: $F_{edge_{N_{test}}} = \frac{match_score\,(edge)}{total_preval\,(edge)}$. In questa formula, $e_{(d_i,d_j)}$ rappresenta il bordo della rete e P è il numero totale di bordi di $N_{test}$.

Per calcolare il punteggio di corrispondenza del cluster di rete, le malattie di $N_{FD}$vengono divise in diversi cluster mediante l'applicazione di un algoritmo di clustering. Il cluster di una rete è un gruppo di "attori" che condividono molte interazioni tra loro. Gli attori di un cluster hanno meno interazioni con gli attori di altri cluster. Gli attori all'interno dello stesso gruppo possono avere caratteristiche simili. Questo vale anche per la rete di malattie. Le malattie tendono a comparire insieme poiché condividono fattori di rischio comuni. Pertanto, le malattie croniche comorbide hanno collegamenti comuni tra di loro e possono costituire cluster. Se i nodi di un bordo di $N_{test}$ rientrano nello stesso cluster in $N_{FD}$, viene conteggiato come corrispondenza di cluster. Ad esempio, quando un bordo tra due malattie (ad esempio, depressione e insufficienza renale) di $N_{test}$ rientra nello stesso cluster in $N_{FD}$, quel bordo viene considerato come una corrispondenza. In questo modo, viene calcolato il conteggio complessivo delle corrispondenze e poi viene diviso per il numero totale di bordi in NFD. La caratteristica basata sul cluster ($F_{cluster}$) è definita come segue:
$F_{cluster_{N_{test}}} \frac{Numero\ di\ bordi\ in\ N_{test}\ i\ quali\ nodi\ hanno\ lo\ stesso\ cluster\ in\ N_{FD}}{Numero\ totale\ di\ bordi\ in\ N_{FD}}$ [15].

### 3.4 Tecniche di machine learning utilizzate

Sebbene siano stati sviluppati molti modelli per modellare il rischio di malattie cardiovascolari (CVD), questo studio ha preso in considerazione quei classificatori basati sull'apprendimento automatico. Nella letteratura sono state proposte diverse tecniche di apprendimento automatico [24]. Questo studio ha utilizzato sei di queste tecniche (ovvero logistic regression, support vector machine, decision tree, random forest, *k*-nearest neighbour e Naïve Bayes) per sviluppare un modello predittivo del rischio di CVD nei pazienti con T2D. L'uso di questi sei algoritmi di apprendimento automatico è ampiamente raccomandato anche nella letteratura per la predizione del rischio di CVD [18].

La *logistic regression* (LR) è un metodo di classificazione ben consolidato. Può essere utilizzata per stimare i valori binari (ad esempio, 0 o 1) basati su un dato insieme di variabili indipendenti. La LR predice la probabilità della presenza o assenza di un evento ed essendo una probabilità, i valori di output si collocano tra 0 e 1. Ad esempio, se il punteggio di probabilità è maggiore di 0,5, la LR lo classificherà come '1'; altrimenti come '0'.

Il *support vector machine* (SVM) rappresenta una tecnica di apprendimento supervisionato utilizzata sia per la classificazione dei dati che per la regressione. Nel caso della classificazione, SVM può mappare ciascun elemento del dato in uno spazio delle caratteristiche ad alta dimensionalità,

generando un iperpiano che separa i punti dati in due classi. Per i dati lineari, l'iperpiano (*hyperplane*) ottimale massimizza la distanza marginale tra le due classi e minimizza gli errori di generalizzazione [24]. Quando si tratta di dati non lineari, SVM utilizza una funzione kernel non lineare per mappare i dati in uno spazio delle caratteristiche ad alta dimensionalità, consentendo così la classificazione dei dati.

Il *decision tree* (DT) un modello a forma di albero che comprende il nodo interno (cioè, le variabili di input), il ramo (cioè, gli esiti) e il nodo foglia (cioè, la classe) [10]. L'albero decisionale è costituito da nodi con diversi livelli, il livello più basso dei quali è chiamato foglie. Inoltre, presenta un nodo radice, che costituisce il livello più alto dell'albero.

Random Forest (RF) è un algoritmo di classificazione che consiste in numerosi Decision Trees (DT). Per classificare un nuovo campione di dati, RF passa il vettore delle caratteristiche di input di quei dati a tutti i suoi alberi decisionali. Ciascun DT produce un risultato di classificazione per quel vettore delle caratteristiche di input. La classificazione che riceve il maggior numero di 'voti' è considerata l'esito finale della classificazione [24].

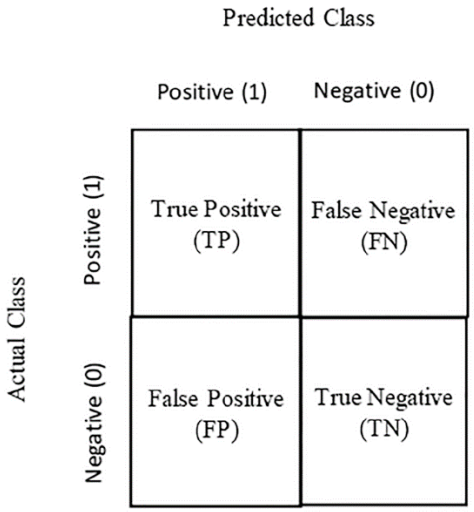
Naïve Bayes (NB) è un classificatore probabilistico sviluppato utilizzando il teorema di Bayes. Questo teorema illustra la probabilità di un evento considerando la conoscenza precedente delle condizioni correlate a quell'evento. Il classificatore NB prevede che la presenza di una specifica caratteristica in una classe non sia correlata alla presenza di qualsiasi altra caratteristica, anche se le caratteristiche per quella classe potrebbero avere indipendenza tra di loro.

Il *k*-nearest neighbour (KNN) è un semplice approccio di apprendimento automatico per problemi di classificazione e regressione. KNN funziona basandosi sulla distanza più vicina dai punti campione rispetto a tutti i punti dell'insieme di dati di addestramento. Nell'algoritmo, k si riferisce al numero di vicini più prossimi considerati per il 'voto' di appartenenza contro i punti campione. La classe che riceve il maggior numero di voti verrà utilizzata per etichettare il punto campione [15].

### 3.5 Struttura della fase di *validation* e *evaluation*

Nello studio in questione inizialmente è stato utilizzato il 65% dei pazienti da ciascuna coorte (cioè CT2D&CVD e CT2D) selezionati casualmente dal dataset per costruire la disease network finale. Tre caratteristiche basate sulla rete (node, edge, cluster match score) sono state calcolate dalla rete di malattie finale, mentre due caratteristiche demografiche (cioè età e sesso) sono state selezionate per il restante 35% dei pazienti. Questo 35% dei pazienti con le relative caratteristiche è stato poi suddiviso in due dataset: un dataset di addestramento e un dataset di test. Lo studio ha impiegato la tecnica di *k-fold* Cross-Validation (CV) utilizzando il dataset di addestramento per

valutare la robustezza dei modelli di classificazione dell'apprendimento automatico. Nel metodo di *k-fold* CV, il dataset di addestramento è stato diviso casualmente in *k* sottoinsiemi di dimensioni uguali. Questo metodo prevede *k* iterazioni. In ogni iterazione, il modello viene addestrato utilizzando *k-1* sottoinsiemi e viene testato utilizzando l'ultimo sottoinsieme. Diversi sottoinsiemi vengono assegnati alle diverse iterazioni. Il dataset di test è stato utilizzato per valutare le prestazioni dei classificatori. La validazione dei classificatori tramite il dataset di test previene il bias della stima delle prestazioni dovuto all'overfitting del modello sul dataset di addestramento. Per tutti i modelli di apprendimento automatico, sono stati utilizzati gli stessi dataset di addestramento e di test per evitare il bias. Questo studio ha calcolato la matrice di confusione che riassume le prestazioni di classificazione di un classificatore basandosi sul dataset di test. Questa matrice è ampiamente utilizzata per calcolare le misure di prestazione (come accuracy, recall, precision e F1 Score). In questo studio è stata anche usata la curva ROC come indicatore delle prestazioni del classificatore. La figura successiva mostra il modello di base di una matrice di confusione [15].




Matrice di confusione, esempio generale con definizioni [15].

### 3.6 Development dei modelli

In [15] si sono utilizzate tutte le sei diverse tecniche di apprendimento automatico menzionate nel paragrafo precedente per sviluppare dei modelli predittivi. Dopo aver selezionato pazienti con diabete di tipo 2 (T2D) e malattie cardiovascolari (CVD) attraverso un processo di filtraggio, ciascuna coorte è stata divisa in due gruppi: uno per creare un disease network di base e l'altro per l'addestramento e il test del modello. Sono state confrontate le reti di malattie individuali dei pazienti di addestramento e test con la rete di malattie finale e calcolati i punteggi per valutare la corrispondenza tra le reti. Inoltre, sono state considerate le feature dell'età e del sesso come fattori di rischio, assegnando loro punteggi appropriati. Successivamente, sono stati normalizzati i punteggi

utilizzando un metodo di trasformazione Z-score e sono stati suddivisi in un dataset di addestramento (utilizzato per addestrare i modelli) e un dataset di test (utilizzato per testare i modelli). Sono state valutate le prestazioni dei modelli utilizzando un approccio di convalida incrociata a 10-fold, dove il dataset è stato suddiviso in 10 gruppi uguali, e i modelli sono stati testati su ciascun gruppo. I risultati del modello sono stati poi confrontati con il dataset di test per valutare l'efficacia della previsione. Sono anche stati condotti test statistici per valutare la significatività dei risultati ottenuti per entrambe le coorti, confermando la validità delle nostre conclusioni con un alto livello di significatività [15].

### 3.7 Performance dei modelli con il training set attraverso la cross-validation

Andando in ordine dei modelli presentati prima, le performance sono state le seguenti. Il modello di Logistic Regression (LR) ha mostrato una accuracy media del 84,21 percento. Il modello Support Vector Machine (SVM) ha mostrato una accuracy media di 84,23 percento. Il modello di classificazione binaria Decision Tree (DT) ha mostrato prestazioni ancora migliori, con una accuracy media di 86 percento. Il Random Forest (RF) ha ottenuto la accuracy media migliore con un valore di 88,94 percento. Il Naïve Bayes (NB) ha ottenuto un valore di 82,21 percento. Il sesto e ultimo modello, il K-nearest neighbour ha raggiunto una accuracy media del 81,77 percento.

Complessivamente, le prestazioni dei sei modelli di apprendimento automatico nel dataset di validazione utilizzando la convalida incrociata a 10 fold per prevedere le CVD nei pazienti con T2D erano quasi simili. Tuttavia, si è riscontrato che il modello predittivo basato su RF ha ottenuto l'accuratezza migliore (88,94 percento), seguito da DT (86,00 percento). Questo studio ha selezionato il miglior modello durante il processo di CV e lo ha testato utilizzando il dataset di test per valutare ulteriormente le prestazioni dei modelli predittivi. Di seguito la tabella con i risultati [15].

| Model name | Accuracy (%) 10-fold CV | Variance |
|---|---|---|
| LR | 84.21 | 0.0019 |
| SVM | 84.23 | 0.0054 |
| DT | 86.00 | 0.0022 |
| RF | 88.94 | 0.0011 |
| NB | 82.21 | 0.0022 |
| KNN | 81.77 | 0.0016 |

Tabella dei risultati dopo la 10-fold cross validation sui vari modelli testati con il training set [15].

### 3.8 Performance dei modelli usando il dataset di test

A questo punto viene utilizzato il dataset di test per valutare ulteriormente le performance dei 6 modelli. La convalida utilizzando questo dataset può evitare eventuali distorsioni nei risultati predetti in precedenza usando il dataset di addestramento. La seguente tabella presenta i vari parametri di prestazione per ogni modello, ovvero accuracy, precision, recall e F1-Score. Si osserva che il modello basato su RF mostra l'accuratezza più alta del 87,50 percento.

L'accuracy per i modelli LR, DT e SVM è del 83,33%, che è leggermente inferiore rispetto al modello RF (87,50 percento). Gli altri modelli hanno comunque prestazioni soddisfacenti nel dataset di test. Uno dei modelli mostra una specificità o recall del 100 percento, indicando che questi modelli non generano falsi negativi per i gruppi con entrambe le patologie: CVD e T2D. Questo è preferibile per il framework proposto poiché dal punto di vista della previsione del rischio di presenza di entrambe le CVD e T2D, è più sicuro identificare i pazienti con T2D come pazienti affetti sia da CVD che da T2D piuttosto che non essere al 100 percento sicuri di aver identificato bene la presenza di solo una delle patologie. Questo approccio conservativo riduce al minimo il rischio di sottovalutare la gravità delle condizioni di salute dei pazienti [15].

Nella letteratura, diversi modelli di previsione del rischio basati su dati sanitari hanno ottenuto la migliore accuratezza utilizzando algoritmi RF [16] [11]. In questo studio il modello di previsione basato su RF ha superato gli altri cinque di poco. Un RF è composto da molti singoli DT che agiscono come un ensemble. La bassa correlazione tra i modelli DT offre vantaggi nel modello RF perché i modelli non correlati possono generare previsioni di insieme più accurate rispetto alle previsioni individuali. Il motivo di questo effetto è che i DT si proteggono reciprocamente dai propri errori individuali. Pertanto, come gruppo, i DT possono convergere verso la classe di previsione corretta nel modello RF. Questa caratteristica non è stata dimostrata negli altri modelli di apprendimento automatico poiché lavorano come modelli costituenti individuali. Tuttavia, l'accuratezza di previsione dei sei modelli di apprendimento automatico è stabile e soddisfacente, considerando le caratteristiche eterogenee del dataset amministrativo [15].

| Model name | Accuracy (%) | Precision (%) | Recall (%) | $F_1$ Score (%) |
|---|---|---|---|---|
| LR | 83.33 | 83.33 | 83.33 | 83.33 |
| SVM | 83.33 | 83.33 | 83.33 | 83.33 |
| DT | 83.33 | 84.62 | 91.67 | 88.00 |
| RF | 87.50 | 80.00 | 100.00 | 88.89 |
| NB | 79.17 | 81.82 | 75.00 | 78.26 |
| KNN | 79.17 | 76.92 | 83.33 | 80.00 |

Tabella che indica i valori degli score ottenuti con il test dataset sui 6 diversi modelli [15].

### 3.9 Statistiche della ROC curve (AUC) e feature importance

Questo studio ha calcolato la AUC sul dataset di test per ogni modello di machine learning utilizzato. Qua di seguito si trova il grafico comparativo dei risultati ottenuti. Per la previsione di CVD in pazienti con T2D il risultato migliore è stato ottenuto nuovamente dal modello RF (Random Forest), con un valore AUC di 0.83, nonostante tutti i modelli abbiano raggiunto un AUC di almeno 0.70. Si tratta di un risultato che permette di affermare che il modello RF è capace di predire il rischio di complicanze cardiovascolari in pazienti diabetici di tipo 2 con efficacia [15].

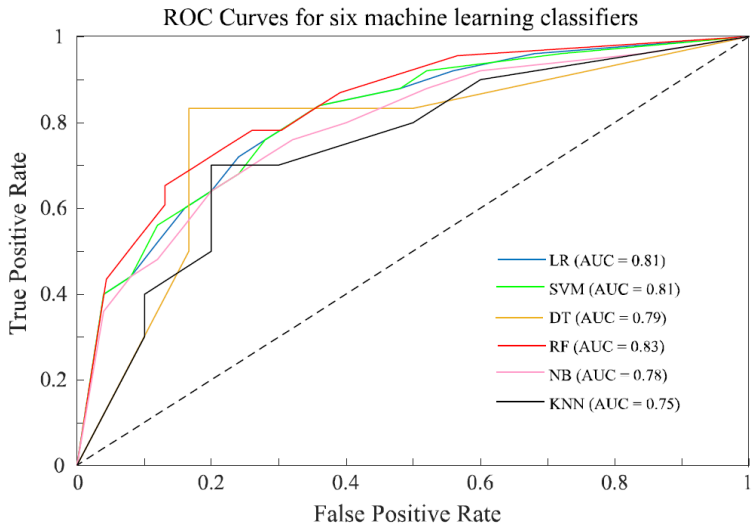



Grafico di tutti i valori di ROC e AUC ottenuti sul dataset di test per ogni modello di ML testato [15].

Di seguito si trova il grafico della feature importance riscontrata in questo studio. È stato utilizzato un MATLAB feature selection toolbox (Feature Selection Library [8]) con il quale si è visto che ogni modello trova delle piccole variazioni di importanza per feature per tutti i pazienti del dataset di addestramento e di test. Si osserva che però risultano sempre rilevanti le feature di età e node match score, e risulta sempre non rilevante o quasi la feature del genere. I modelli RF e SVM sono gli unici due a dare importanza a tutte le feature ed è importante notare come questa caratteristica li porti ad avere risultati sempre consistenti nella predizione di CVD in pazienti T2D [15].

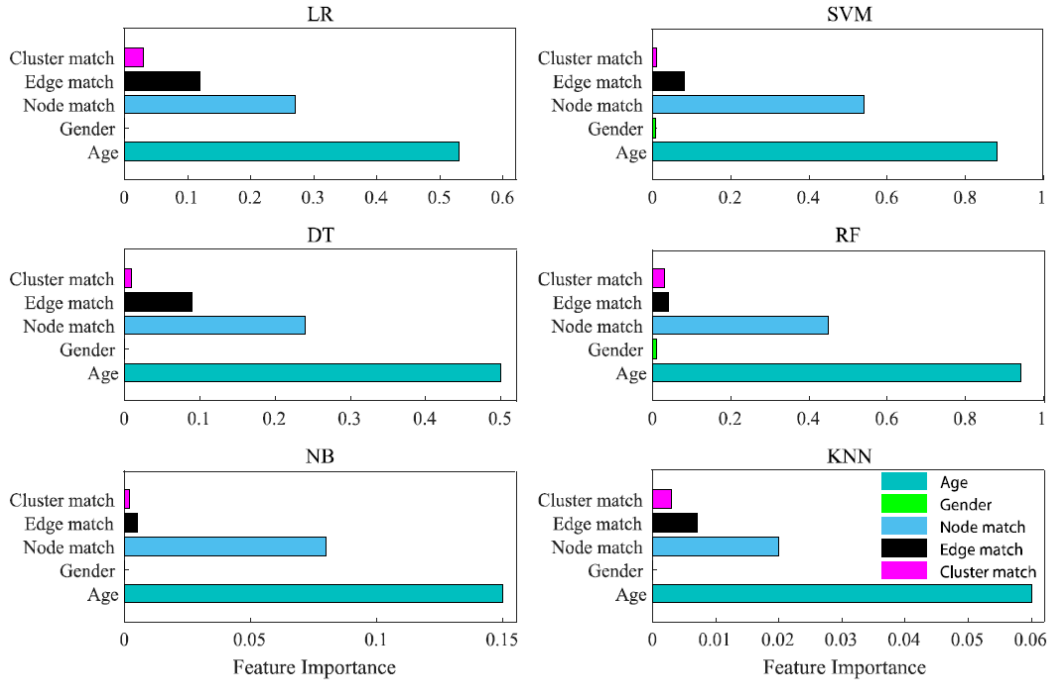



Tabella dei risultati di feature selection sui modelli di ML considerati nello studio esaminato. Si osserva l'importanza della feature dell'età (*age*) e node match score, che ottengono risultati molto simili in ogni valutazione [15].

### 3.10 Discussione sul caso di studio

Esistono numerose complicanze delle condizioni presenti tra CVD e T2D a causa delle complesse relazioni tra loro. Queste comorbilità sono spesso causate sia dalle CVD che dal T2D, o da entrambi. Poiché le comorbilità delle malattie croniche condividono frequentemente fattori di rischio comuni, si sta sempre più considerando la loro interazione durante la valutazione delle condizioni di salute. L'identificazione precoce dei pazienti affetti da queste malattie croniche potrebbe aiutare a prevenire le relative complicanze, rendendo estremamente utili gli strumenti di analisi di grosse quantità di dati come le tecniche di machine learning.

Per riassumere il lavoro del caso di studio esaminato, si può dire quanto segue. Con lo scopo di individuare il rischio di CVD in pazienti T2D, è stato utilizzato un insieme di dati fornito da una cassa di assistenza sanitaria privata. Sono stati selezionati 172 pazienti diagnosticati sia con CVD che con T2D e 172 pazienti diagnosticati solo con T2D. Successivamente, è stata creata la rete delle malattie finale per calcolare le misure basate sulla rete. Queste caratteristiche o punteggi di rischio sono stati utilizzati per sviluppare il modello predittivo basato sull'apprendimento automatico. Nello studio sono state utilizzate tecniche di machine learning, ovvero LR, SVM, DT, RF, NB e KNN. I modelli sono stati valutati in termini di accuracy, precision, recall e F1-Score. I risultati mostrano che le prestazioni di ciascun modello di previsione del rischio sono simili. Il modello predittivo basato su

RF ha fornito una previsione con un'alta accuracy (87.5 percento) sui dati di test. In generale, la maggior parte dei modelli ha raggiunto una accuracy superiore al 79 percento, dimostrando così il suo potenziale nell'utilizzo pratico. La figura sottostante fornisce un'analisi approfondita delle differenze tra le tecniche di machine learning prese in considerazione. Come rivelato in questa figura, solo RF ha una accuracy del 100% per i casi veri positivi. Ha un'accuratezza relativamente più bassa (75%) per i casi falsi positivi, ma è solo del 8.33% al di sotto del valore massimo ottenuto tra tutte le tecniche. È importante capire anche che ciascun modello potrebbe essere utile per compiti diversi, in base ai risultati diversi che ha ottenuto in questo studio. Ad esempio, secondo questa figura, se l'obiettivo è identificare i pazienti con T2D che sono veramente a rischio di CVD, allora RF è la migliore tecnica da considerare, seguita da DT. Questo perché il modello riesce a trovare con una accuratezza del 100 percento i soggetti True Positive, in un contesto in cui non ha importanza se i falsi positivi riscontrati possono crescere. Allo stesso modo, se l'obiettivo è identificare i pazienti con T2D che non sono a rischio di CVD, allora ci sono quattro opzioni (LR, SVM, DT e NB) [15].

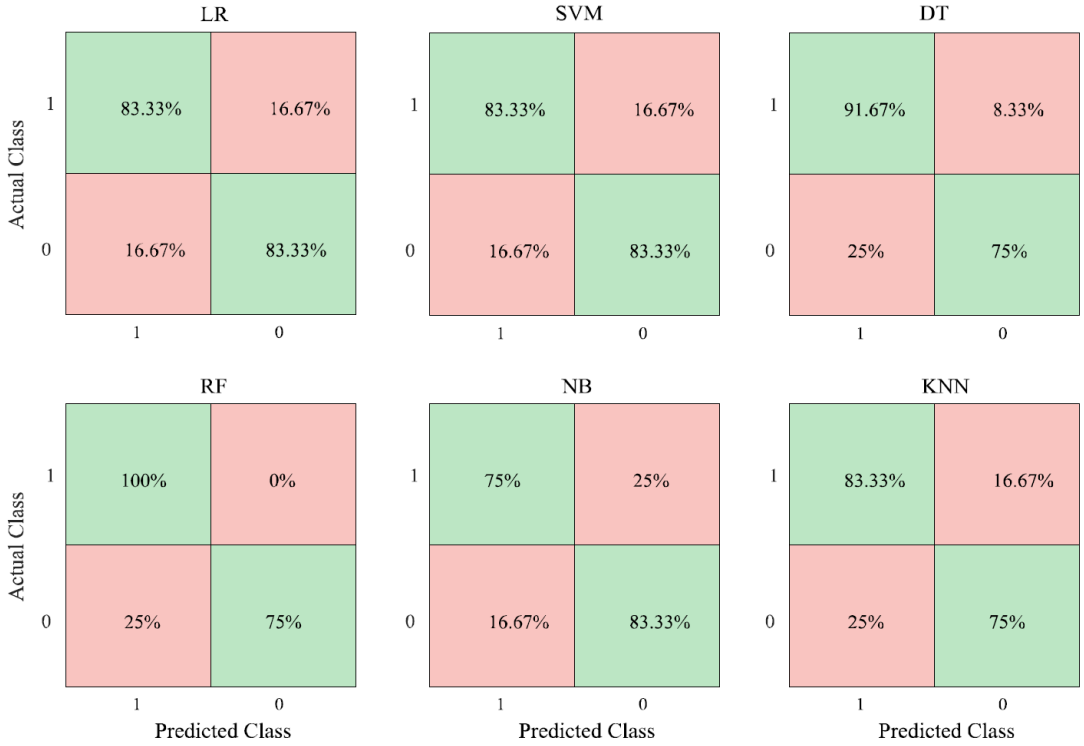



Tabella di tutte le confusion matrix ottenute dal caso di studio [15].

### 3.10.1 Importanza della feature dell'età e del genere

La tabella sottostante mostra le statistiche delle due coorti (cioè $C_{T2D\&CVD}$ e $C_{T2D}$) in termini di età e genere. Si è riscontrato che la percentuale di persone anziane (cioè ≥60 anni) è maggiore rispetto agli altri nella coorte selezionata. Inoltre, la proporzione di pazienti di sesso femminile con malattie croniche multiple è maggiore rispetto alla proporzione di pazienti di sesso femminile con una singola malattia cronica. Al contrario, il rischio di CVD nei pazienti maschi con T2D è relativamente più basso. Per i pazienti diabetici, si osserva che le pazienti di sesso femminile hanno una maggiore probabilità di sviluppare CVD rispetto ai pazienti di sesso maschile [25]. Inoltre, l'Istituto Australiano per la Salute e il Benessere (AIHW) ha riportato un numero maggiore di pazienti maschi rispetto alle pazienti femminili [1]. Dati statistici simili sono stati riscontrati anche in questo caso di studio esaminato. Pertanto, il dataset utilizzato in questo studio rispecchia le statistiche governative australiane [15].

| | $C_{T2D\&CVD}$ Population (%) | $C_{T2D}$ Population (%) |
|---|---|---|
| *Age* | | |
| 0– 29 | 0 | 0 |
| 30–39 | 0.56 | 0.21 |
| 40–49 | 1.18 | 0.68 |
| 50–59 | 4.64 | 16.66 |
| 60–69 | 18.61 | 25.00 |
| 70–79 | 32.40 | 27.64 |
| 80–89 | 34.46 | 23.42 |
| 90–99 | 7.46 | 4.20 |
| ≥100 | 0.68 | 2.19 |
| *Gender* | | |
| Male | 59.30 | 69.65 |
| Female | 40.70 | 30.35 |

Tabella di confronto tra le due coorti esaminate. Si osservano le differenze di età nel primo rilevamento di T2D tra i due gruppi [15].

È stata condotta anche un'analisi bivariata per comprendere eventuali relazioni potenziali tra età e numero di anni tra la prima diagnosi di T2D e CVD. Ciò potrebbe rivelare una relazione potenziale tra l'età e la progressione verso il CVD dai dati. L'età è stata calcolata al momento della prima diagnosi di T2D. La maggior parte delle persone ha sviluppato il CVD entro un intervallo temporale di 0-5 anni. È stato anche adattato un modello lineare con un intervallo di confidenza del 95 percento mostrato in blu. I punti sembrano distribuiti uniformemente sul grafico, suggerendo una buona adattabilità al modello lineare. Complessivamente, l'analisi mostra che i pazienti anziani con T2D sembrano sviluppare il CVD prima rispetto ai pazienti più giovani con T2D. Pertanto, queste due variabili potrebbero essere importanti caratteristiche nello sviluppo di modelli predittivi per la progressione del CVD nei pazienti con T2D. Segue il grafico dei dati di questa analisi [15].

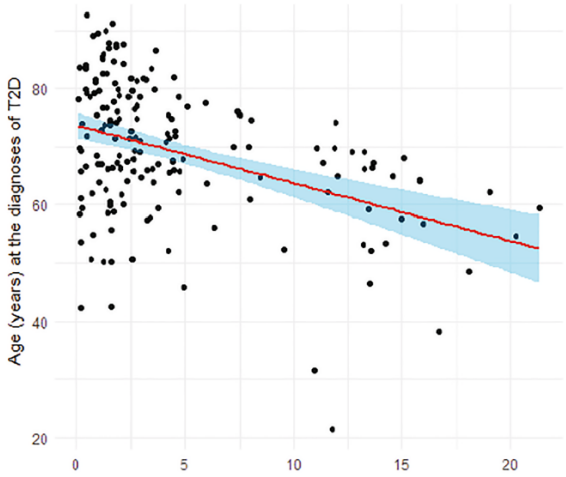



Grafico dell'analisi bivariata su età al momento della diagnosi di T2D (asse y) e intervallo di tempo, in anni, tra diagnosi di T2D e diagnosi di CVD (asse x) [15].

### 3.10.2 Confronto con altri studi in letteratura

Questa ricerca ha utilizzato dati che vengono comunemente raccolti per l'elaborazione delle richieste di assistenza sanitaria ed è facilmente accessibile per le indagini di ricerca. Apparentemente, questa ricerca introduce per la prima volta l'analisi di rete nella modellazione del rischio di CVD per i pazienti con condizioni di T2D. Anche se questa ricerca utilizza solo informazioni ICD dalle storie di ricovero e dimissione degli ospedali dei pazienti, rivela un'accuratezza molto buona per i modelli sviluppati [15]. Si tratta di una differenza sostanziale, in quanto altri studi utilizzano dati alternativi, come ad esempio informazioni demografiche, informazioni di anamnesi medica e dati comportamentali, raccolti dai partecipanti durante interviste a domicilio. Questi partecipanti sono anche invitati a sottoporsi a esami fisici, fisiologici e di laboratorio dettagliati, eseguiti da personale qualificato in centri mobili appositamente attrezzati [27]. I database in questione spesso sono presi da NHANES (National Health and Nutrition Examination Survey) [2]. L'utilizzo di questi database comporta la soluzione di nuovi problemi che fanno parte del Data Mining/Modeling: estrarre i dati, rendere i dati completi, gestire i dati mancanti, affrontare i nomi di attributi diversi tra annate diverse dello stesso database rendendo omogenea la continuità tra di loro, e altri casi. Il Data Mining si può suddividere in pre-processing, extraction e normalization. Altri database simili si differenziano per regioni di raccolta dei dati, come ad esempio ANDIS, SDR, ANDIU e DIREVA per le regioni scandinave [6], PIMA (Pima Indians Diabetes Database) negli Stati Uniti (Arizona) [17], ARIC (Atherosclerosis Risk in Communities) e CHS (Cardiovascular Health Study) a livello globale [7], e tanti altri.

### 3.10.3 Limitazioni dell'uso di tecniche di machine learning nel caso di studio esaminato

Lo studio [13] presenta alcune limitazioni importanti da considerare. In primo luogo, il dataset utilizzato comprende dati sanitari reali, ma la precisione della codifica è influenzata da diversi sistemi di codifica utilizzati nei vari ospedali. Questa diversità può compromettere l'accuratezza complessiva del dataset. Inoltre, i cambiamenti frequenti nelle politiche e l'esperienza variegata dei codificatori clinici possono avere un impatto sulla coerenza dei dati. In secondo luogo, i dati utilizzati provengono principalmente dai rapporti di ammissione e dimissione ospedaliera, escludendo le visite dai medici di famiglia. Ciò potrebbe portare a una mancata inclusione di determinati codici di malattie o a una sottorappresentazione di alcune condizioni, soprattutto se legate alla cura primaria. Inoltre, i pazienti trattati in strutture ospedaliere pubbliche non sono inclusi, il che può influenzare la rappresentatività del campione. La terza limitazione riguarda le malattie cardiovascolari considerate. Alcune condizioni, come ictus e infarto miocardico, potrebbero non essere incluse nello studio poiché ci si è basati solo su alcune malattie cardiovascolari specifiche menzionate nell'indice di Elixhauser. Inoltre, non sono stati considerati alcuni importanti fattori di rischio comportamentali, come fumo e alcol, e informazioni farmaceutiche dettagliate, limitando così l'analisi. Infine, il numero di pazienti nelle due coorti è relativamente limitato, il che può limitare la generalizzabilità dei risultati. Studi futuri dovrebbero considerare una maggiore diversità nei dati e un campione più ampio per garantire risultati più robusti e rappresentativi. Inoltre, il modello di previsione del rischio proposto non può fornire un punteggio di regressione in quanto si basa su una classificazione binaria. Per avere una visione più completa, sarà necessario testare questo framework su altri dataset pubblici al di fuori del contesto australiano, allo scopo di confrontare e validare il modello di questa ricerca. Tuttavia, è importante sottolineare che il modello utilizzato potrebbe essere di grande utilità per le part interessate, consentendo loro di identificare il rischio di malattie cardiovascolari nei pazienti affetti da diabete di tipo 2 [15].

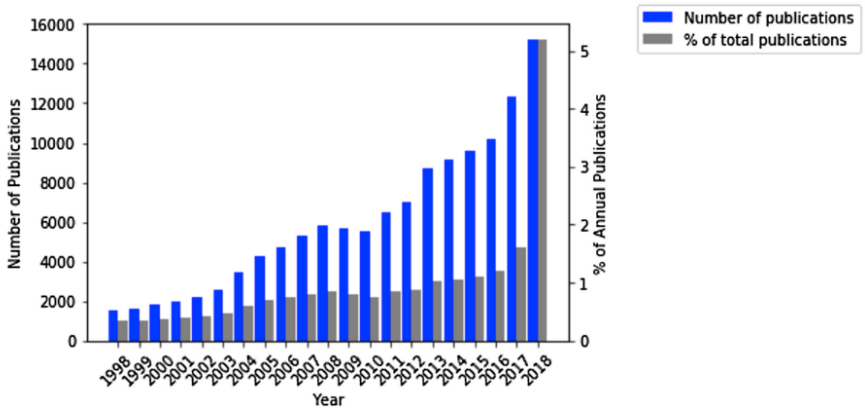
# 4. Conclusioni

## 4.1 Bilancio del lavoro svolto

In conclusione, l'esplorazione approfondita del contesto clinico del diabete di Tipo 2 e delle sue complicanze cardiovascolari, unitamente all'analisi delle applicazioni del machine learning in medicina e alla disamina dettagliata di uno specifico caso di studio, ci ha fornito una panoramica sulle possibilità e le sfide che questa innovativa tecnologia offre nel campo della salute. Attraverso l'analisi dei dati complessi e l'implementazione di modelli predittivi avanzati, il machine learning ha dimostrato di essere una risorsa importante nella diagnosi precoce, nella gestione personalizzata delle malattie e nell'ottimizzazione dei risultati clinici. Durante il nostro percorso, abbiamo compreso l'importanza vitale di selezionare accuratamente le caratteristiche e di valutare attentamente la qualità dei dati, elementi fondamentali per garantire modelli affidabili e generalizzabili. Abbiamo esaminato l'importanza dell'età, del genere e di altre caratteristiche specifiche nel contesto del machine learning, evidenziando come queste variabili influiscano significativamente sulla precisione delle previsioni. Nella nostra disamina del caso di studio specifico, abbiamo esaminato criticamente le prestazioni dei modelli attraverso l'analisi delle curve ROC (AUC) e l'importanza delle caratteristiche, offrendo così un quadro completo delle capacità predittive dei modelli implementati. Abbiamo anche discusso delle limitazioni, riconoscendo l'importanza di affrontare questioni quali la generalizzabilità dei modelli su diverse popolazioni e periodi temporali. Guardando al futuro, è evidente che il machine learning in medicina rappresenta una promettente area di sviluppo. Tuttavia, è essenziale affrontare le sfide legate alla qualità dei dati, alla generalizzabilità e alla comprensione delle variabili coinvolte per garantire l'applicazione sicura ed efficace di queste tecnologie nella pratica clinica quotidiana. Il costante sviluppo e l'adozione responsabile del machine learning nella medicina moderna possono rivoluzionare il modo in cui i medici affrontano il diabete di tipo 2 e le sue complicanze cardiovascolari, portando a possibili diagnosi più accurate, trattamenti personalizzati e miglioramenti tangibili nella qualità della vita dei pazienti. In definitiva, questa elaborato fornisce una visione chiara delle potenzialità del machine learning in medicina, mostrando solo uno dei numerosi aspetti positivi di questi strumenti.

### 4.2 Considerazioni su possibili scenari futuri

L'innovazione nell'ambito dell'intelligenza artificiale nel settore sanitario sta procedendo a passo veloce e sembra destinata a continuare così. Un'analisi della letteratura scientifica su PubMed utilizzando termini come "deep learning", "machine learning", "intelligenza artificiale" o "rete neurale" rivela un aumento costante sia nel numero assoluto di ricerche legate all'intelligenza artificiale, sia nella loro percentuale rispetto all'intera produzione scientifica nel campo medico. Nel 2000, sono stati pubblicati 1838 articoli (corrispondenti allo 0,38% di tutti gli articoli dell'anno), che soddisfacevano i criteri di ricerca. Nel 2010, il numero è salito a 5491 articoli (pari allo 0,74% di tutti gli articoli), mentre nel 2018 sono stati pubblicati 15.240 articoli (costituendo il 5,20% di tutte le pubblicazioni scientifiche dell'anno). Qua sotto si riporta un grafico che mostra i dati visivamente [22]. Queste tendenze mostrano che l'uso dell'intelligenza artificiale nella sanità sta crescendo rapidamente, grazie alla crescente consapevolezza della sua utilità e agli avanzamenti tecnologici. Inoltre, la disponibilità gratuita di pacchetti software standard e l'accesso conveniente a potenti risorse di calcolo online hanno reso più accessibile l'implementazione di algoritmi di intelligenza artificiale complessi. Questa democratizzazione degli strumenti ha reso più semplice per gli esperti del settore sanitario adottare l'intelligenza artificiale in vari ambiti specifici, senza richiedere investimenti finanziari elevati iniziali.




Pubblicazioni riguardanti l'IA nella sanità su PubMed. I numeri sono stati generati tramite Boolean PubMed ricercando i termini "deep learning", "machine learning", "artificial intelligence" o "neural network" [22].

Nessuna applicazione medica dell'IA ha registrato progressi recenti o attirato tanta attenzione quanto l'elaborazione delle immagini mediche, soprattutto nei campi della patologia e radiologia. Nei casi in cui l'IA ha prestazioni paragonabili a quelle dei lettori umani, alcuni compiti di radiologia di routine potrebbero diventare sempre più automatizzati in futuro, consentendo agli operatori umani di dedicare più tempo ai casi complessi o alle immagini in cui le previsioni basate sull'IA non

raggiungono una certa soglia di certezza. Attualmente, i carichi di lavoro dei radiologi limitano il numero di immagini che possono essere lette in un sistema sanitario, e le tariffe di lettura aumentano i costi medici. I sistemi automatizzati progettati per sollecitare la revisione umana per immagini selezionate potrebbero migliorare la velocità di interpretazione delle immagini, consentendo di dedicare più tempo ai casi difficili e riducendo i costi [13].

Come già menzionato in precedenza, ciò che distingue principalmente questo studio da altri è il tipo di dati raccolti e utilizzati per la ricerca. Al contrario di altri studi, questo si è focalizzato sulla disease network, composta dalle relazioni tra ICD diversi. Bisogna ricordare quindi che, per produrre un modello predittivo di qualità è fondamentale avere a disposizione dati di qualità. Inoltre, lo sviluppo del modello richiede spesso grandi quantità di dati per l'addestramento. Fortunatamente, nel settore sanitario spesso si dispone di grandi quantità di dati digitalizzati, riducendo questa limitazione. Tuttavia, queste risorse di dati sanitari sono spesso limitate a sistemi di cartelle cliniche e dati di registri e richieste di rimborso. Ottenere queste risorse è difficile, richiedendo spesso partnership interne ed esterne, nonché un enorme sforzo per la pulizia e la manipolazione dei dati [22]. Il futuro dell'applicazione di machine learning dipende perciò anche da questo fattore.

Un interessante ambito di dati da considerare per il futuro è quello degli *smart wearable data*. Come dice il nome, riguarda i dati raccolti da dispositivi wearable come smartphone e smartwatch. Il motivo per cui è estremamente interessante e attuale è la disponibilità di enormi quantità di dati di questo genere, e il facile accesso ad essi, dovuto alle multinazionali che li producono. La grande limitazione invece consiste nella qualità di dati e nella tipologia di questi. È naturale che dei dispositivi portatili di ridotte dimensioni non possano avere gli strumenti per misurare valori di carattere medico che di solito si ottengono da analisi di laboratorio ed è per questo che finora vi è stato solo un limitato numero di ricerche in questo ambito [22]. Tuttavia, esiste uno studio che ha esaminato i dati raccolti da tali dispositivi sull'esercizio fisico cercando di collegarli a diverse malattie, tra cui il diabete e l'ipertensione. Questo studio è noto come progetto di test di esercizio Henry Ford (FIT) [23]. Tutti i modelli hanno avuto una precisione che si colloca tra il 60 e il 70 percento, ma il metodo Random Tree Forest (RTF) ha raggiunto una precisione del 91%. Come per lo studio [15] esaminato in questo elaborato, il Random Forest e la sua variante Random Tree Forest, si confermano efficaci strumenti per determinare i veri positivi e, dunque, per rilevare efficacemente i casi di CVD in pazienti T2D.

## Bibliografia


[1] B. Tong and C. Stevenson, “Comorbidity of cardiovascular disease, diabetes and chronic kidney disease in Australia”. Deakin University, 01-Jan-2007.

[2] Centers for Disease Control and Prevention, “NHANES - About the National Health and Nutrition Examination Survey,” *Centers for Disease Control and Prevention*, 2019. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

[3] C. M. Italia, “Intelligenza artificiale in sanità: applicazioni, pro e contro,” *www.cgm.com*. https://www.cgm.com/ita_it/magazine/articles/intelligenza-artificiale-in-sanita.html (accessed Nov. 2, 2023).

[4] “Diabete e rischio cardiovascolare: le principali complicanze,” *www.grupposandonato.it*. https://www.grupposandonato.it/news/2022/novembre/diabete-e-rischio-cardiovascolare (accessed Nov. 5, 2023).

[5] “Diabete tipo 2,” *Humanitas*. https://www.humanitas.it/malattie/diabete-tipo-2/ (accessed Nov. 5, 2023)

[6] E. Ahlqvist *et al.*, “Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables,” *The Lancet Diabetes & Endocrinology*, vol. 6, no. 5, pp. 361–369, May 2018, doi: https://doi.org/10.1016/s2213-8587(18)30051-2.

[7] Evangelos Oikonomou and R. Khera, “Machine learning in precision diabetes care and cardiovascular risk prediction,” *Cardiovascular Diabetology*, vol. 22, no. 1, Sep. 2023, doi: https://doi.org/10.1186/s12933-023-01985-3.

[8] “Feature Selection Library,” *it.mathworks.com*, Nov. 14, 2023. https://it.mathworks.com/matlabcentral/fileexchange/56937-feature-selection-library (accessed Nov. 6, 2023).

[9] H. Quan *et al.*, “Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data,” *Medical care*, vol. 43, no. 11, pp. 1130–9, 2005, doi: https://doi.org/10.1097/01.mlr.0000182534.19832.83.

[10] J. R. Quinlan, *Induction of decision trees*. Broadway, N.S.W., Australia: New South Wales Institute Of Technology, School Of Computing Sciences, 1985.

[11] J. Yang, D. Yao, X. Zhan, X. Zhan, *Predicting disease risks using feature selection based on random forest and support vector machine*, Springer : 2014.

[12] K. Arun Bhavsar, J. Singla, Y. D. Al-Otaibi, O.-Y. Song, Y. Bin Zikriya, and A. Kashif Bashir, “Medical Diagnosis Using Machine Learning: A Statistical Review,” *Computers, Materials*

*& Continua*, vol. 67, no. 1, pp. 107–125, 2021, doi: https://doi.org/10.32604/cmc.2021.014604.

[13] L. Saba *et al.*, "The present and future of deep learning in radiology," *European journal of radiology*, vol. 114, pp. 14–24, 2019, doi: https://doi.org/10.1016/j.ejrad.2019.02.038.

[14] M. della Salute, "Diabete mellito tipo 2," *www.salute.gov.it*. https://www.salute.gov.it/portale/nutrizione/dettaglioContenutiNutrizione.jsp?lingua=italiano&id=5511&area=nutrizione&menu=croniche

[15] M. E. Hossain, S. Uddin, and A. Khan, "Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes," *Expert Systems with Applications*, vol. 164, p. 113918, Feb. 2021, doi: https://doi.org/10.1016/j.eswa.2020.113918.

[16] M. Juhola, H. Joutsijoki, K. Penttinen, and K. Aalto-Setälä, "Detection of genetic cardiac diseases by Ca 2+ transient profiles using machine learning methods," *Scientific Reports*, vol. 8, no. 1, p. 9355, Jun. 2018, doi: https://doi.org/10.1038/s41598-018-27695-5.

[17] "Pima Indians Diabetes Database," *www.kaggle.com*. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[18] R. K. Sevakula, W. M. Au-Yeung, J. P. Singh, E. K. Heist, E. M. Isselbacher, and A. A. Armoundas, "State-of-the-Art Machine Learning Techniques Aiming to Improve Patient Outcomes Pertaining to the Cardiovascular System," *Journal of the American Heart Association*, vol. 9, no. 4, Feb. 2020, doi: https://doi.org/10.1161/jaha.119.013924.

[19] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?", *PLOS ONE*, vol. 12, no. 4, p. e0174944, Apr. 2017, doi: https://doi.org/10.1371/journal.pone.0174944.

[20] S. Halim *et al.*, "Simultaneous Consideration of Multiple Candidate Protein Biomarkers for Long-Term Risk for Cardiovascular Events," *Circulation-cardiovascular Genetics*, vol. 8, no. 1, pp. 168–177, Feb. 2015, doi: https://doi.org/10.1161/circgenetics.113.000490.

[21] S. J. Al'Aref *et al.*, "Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the CONFIRM registry," *European Heart Journal*, vol. 41, no. 3, pp. 359–367, Sep. 2019, doi: https://doi.org/10.1093/eurheartj/ehz565.

[22] S. J. Al'Aref, G. Singh, L. Baskaran, and D. Metaxas, *Machine Learning in Cardiovascular Medicine*. Academic Press, 2020.

[23] S. Sakr *et al.*, "Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford ExercIse Testing (FIT) Project," *PLOS ONE*, vol. 13, no. 4, p. e0195344, Apr. 2018, doi: https://doi.org/10.1371/journal.pone.0195344.

[24] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, Dec. 2019, doi: https://doi.org/10.1186/s12911-019-1004-8.

[25] The Emerging Risk Factors Collaboration, "Diabetes mellitus, Fasting Blood Glucose concentration, and Risk of Vascular disease: a Collaborative meta-analysis of 102 Prospective Studies," *The Lancet*, vol. 375, no. 9733, pp. 2215–2222, Jun. 2010, doi: https://doi.org/10.1016/s0140-6736(10)60484-9.

[26] Thorsten Joachims, "Making large scale SVM learning practical," *Technical reports*, Oct. 1999, doi: https://doi.org/10.17877/de290r-14262.

[27] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Medical Informatics and Decision Making*, vol. 10, p. 16, Mar. 2010, doi: https://doi.org/10.1186/1472-6947-10-16.