UNIVERSITÀ
DI PAVIA

FACOLTÀ DI INGEGNERIA DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE.

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA.

TESI DI LAUREA

TITOLO

MACHINE LEARNING APPROACHES TO PREDICT CARDIOVASCULAR RISK IN PREDIABETES USING DANISH NATIONAL REGISTRY DATA

Candidato: Lorenzo Calvagna

Relatore: Prof. Riccardo Bellazzi
Correlatore: Prof. Morten Hasselstrøm Jensen

A.A.2024/2025

INDICE

SOMMARIO

Le malattie cardiovascolari rappresentano una delle principali complicazioni nei pazienti che presentano una condizione di prediabete. La possibilità di prevedere precocemente tali eventi riveste un ruolo essenziale nelle strategie di prevenzione, al fine di consentire l'individuazione immediata di soggetti a maggior rischio, e anche l'adozione di interventi mirati.
L'obiettivo del presente studio è quello di valutare l'efficacia dell'utilizzo di algoritmi di Intelligenza Artificiale applicati a dati amministrativi e clinici estratti dai registri sanitari danesi, al fine di esaminare il rischio cardiovascolare in una coorte di pazienti con prediabete.

Il presente studio si basa sull'analisi dei dati estratti dai registri sanitari danesi nel periodo compreso tra gli anni 2010 – 2015. Considerata la grande quantità e complessità dei dati da analizzare, è stata implementata una pipeline di Data Engineering. Tale fase preliminare è stata necessaria al fine di scremare i dati, risolvere le problematiche legate all'utilizzo di database differenti e creare un dataset strutturato. Successivamente, sono stati sviluppati, addestrati e confrontati quattro modelli di Machine Learning, ponendo particolare attenzione alla questione relativa allo sbilanciamento delle classi, al fine di ottenere delle predizioni attendibili.

I modelli sviluppati hanno dimostrato simile capacità predittiva, validata da metriche quali AUC e Recall. Dall'analisi della Feature Importance è emerso che l'età e i livelli di emoglobina glicata (HbA1c) sono i predittori principali.

I risultati ottenuti confermano che il Machine Learning, applicato ai dati sanitari, rappresenta un ottimo strumento di screening di primo livello, capace di individuare i pazienti con prediabete ad alto rischio cardiovascolare.
Nonostante le limitazioni dei registri sanitari, come l'assenza di variabili legate allo stile di vita, il presente studio fornisce delle basi solide su cui fondare studi e

approfondimenti futuri volti all'integrazione con dati di trial clinici e ulteriori analisi mediante approcci di Machine Learning.

1. INTRODUZIONE

1.1 L'importanza dei registri nella raccolta dei dati sanitari

Nell'ultimo decennio si è progressivamente sviluppato l'utilizzo di registri clinici come fonte di dati per la ricerca scientifica, rappresentando un'alternativa molto valida rispetto ai classici trial clinici. Questo perché i registri consentono di raccogliere informazioni su ampie popolazioni di pazienti, permettendo così di avere una panoramica più completa e significativa del decorso delle malattie e dell'efficacia dei trattamenti al di fuori delle sole condizioni tipiche e controllate dei trial clinici.

Questa caratteristica rende i registri molto utili, perché permettono di ottenere risultati che possono essere applicati alla popolazione reale. Inoltre, grazie alla struttura già esistente, è possibile svolgere analisi retrospettive in modo più rapido e con minori risorse, senza dover reclutare nuovi pazienti o organizzare uno studio di follow-up.

Tuttavia, l'utilizzo dei registri come fonte primaria di dati presenta alcune limitazioni metodologiche che devono essere rigorosamente attenzionate. Trattandosi di dati osservazionali, i registri non consentono di stabilire relazioni causali tra esposizione e outcome, ma permettono unicamente di evidenziare associazioni. L'assenza di randomizzazione può inoltre introdurre diversi tipi di bias, in particolare di selezione e di informazione, che possono compromettere la validità interna dello studio.

Un'ulteriore criticità riguarda la qualità dei dati raccolti, che non è sempre uniforme: possono infatti essere presenti dati mancanti, errori di registrazione o differenze nei protocolli di raccolta tra i vari centri partecipanti. Un altro grande limite risiede nella disponibilità delle variabili utili all'analisi, poiché i registri

sono spesso progettati per scopi specifici e possono quindi non includere tutti i dettagli clinici necessari per condurre analisi più approfondite. [1]
Per ovviare a queste limitazioni, è possibile adottare delle strategie metodologiche e analitiche: per controllare i fattori di distorsione e ridurre i bias dovuti alla mancanza di randomizzazione, è possibile applicare tecniche statistiche avanzate come la regressione multivariata o i modelli a effetti misti.

Inoltre, l'integrazione di approcci basati su modelli di Machine Learning (ML) può contribuire a migliorare la qualità delle analisi, permettendo di individuare relazioni non lineari tra le variabili, gestire dati incompleti e riconoscere pattern complessi nei dataset di grandi dimensioni.

Il principale punto di distinzione tra il Machine Learning e i metodi statistici tradizionali riguarda l'obiettivo dell'analisi: i metodi di ML si concentrano sulla predizione, cercando di identificare pattern e relazioni nei dati in modo da massimizzare la capacità del modello di prevedere un determinato outcome.

Gli approcci statistici tradizionali, quali regressione e modelli a effetti misti, si basano invece su ipotesi esplicite — come linearità, distribuzioni note, indipendenza o correlazione regolata tra le osservazioni — che permettono di stimare parametri interpretabili e valutare la significatività dei risultati ottenuti in seguito a uno studio. [2]

1.2 Danish National Patient Registry (DNPR)

Il Danish National Patient Registry (DNPR) rappresenta uno dei più antichi e completi registri ospedalieri nazionali al mondo e costituisce una fonte fondamentale per la ricerca epidemiologica e clinica. Il registro raccoglie informazioni sanitarie dettagliate relative ai contatti ospedalieri dei cittadini

danesi, incluse diagnosi, procedure e ricoveri, consentendo una ricostruzione accurata dei percorsi clinici individuali.

Un elemento centrale del sistema sanitario danese è il Civil Registration System, che assegna a ogni residente un identificativo personale univoco (CPR). Tale identificativo permette il collegamento puntuale e affidabile tra il DNPR e altri registri sanitari e amministrativi nazionali. In particolare, il CPR consente l'integrazione dei dati clinici con le informazioni contenute in Statistics Denmark (DST), ampiamente utilizzate in ambito epidemiologico.

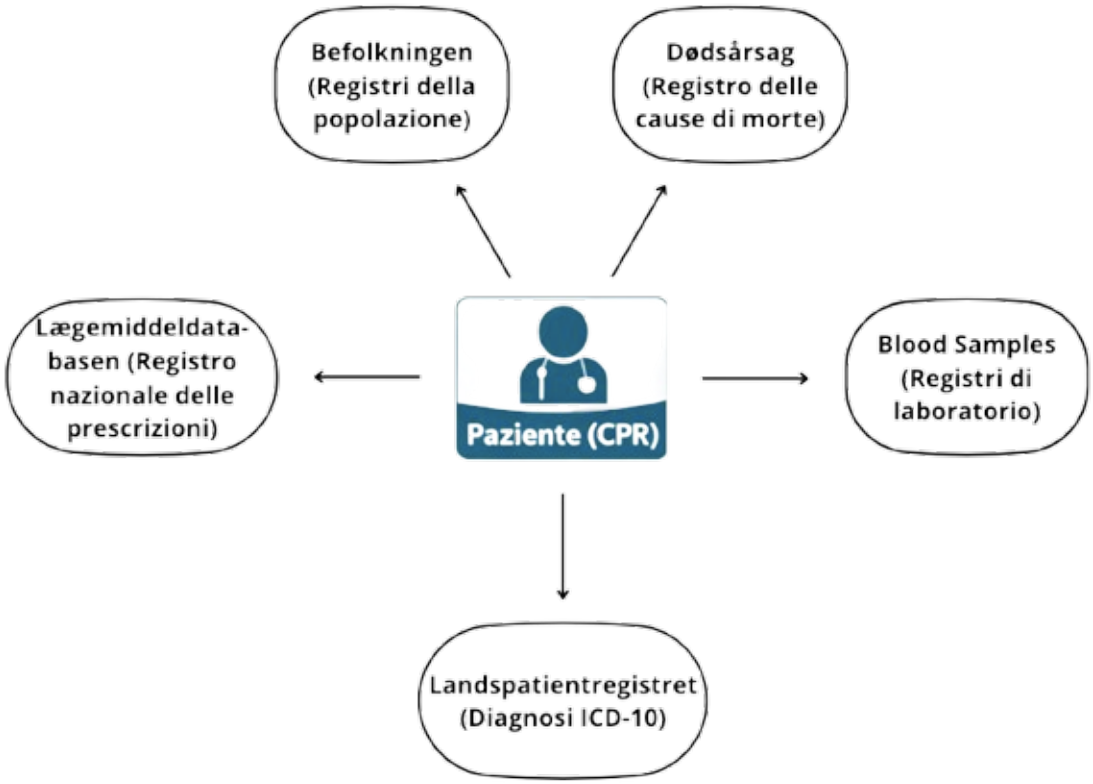

Figura 1: Architettura di integrazione dei registri sanitari danesi. L'identificativo personale univoco (CPR) funge da chiave primaria relazionale, consentendo il linkage tra i vari database nazionali.

Questa integrazione rende possibile un'analisi multidimensionale: mentre il DNPR consente di identificare chi si è ammalato e quali trattamenti ha ricevuto, i registri gestiti da Statistics Denmark forniscono informazioni relative alle caratteristiche demografiche, sociali ed economiche degli individui. La combinazione di queste

fonti rappresenta uno dei principali punti di forza del sistema danese dei registri sanitari.

Il DNPR è stato istituito nel 1977 e ha raggiunto una copertura nazionale completa a partire dal 1978. Nel corso degli anni, l'ampiezza e la qualità dei dati raccolti sono progressivamente aumentate; nel periodo compreso tra il 1977 e il 2012, il registro ha incluso informazioni relative a circa 8 milioni di individui distinti, coprendo di fatto l'intera popolazione danese nel lungo periodo. [3]

1.3 Machine Learning in medicina

Negli ultimi anni, il Machine Learning (ML) ha mostrato un grande potenziale nel campo della medicina, affermandosi come una delle risorse più promettenti per l'analisi di grandi quantità di dati clinici.
Il ML consente di identificare pattern complessi e relazioni non lineari all'interno dei dati, offrendo strumenti avanzati per la predizione di eventi clinici, la diagnosi automatizzata e la personalizzazione dei trattamenti.
Rispetto ai metodi statistici tradizionali, che richiedono ipotesi esplicite sul modello e sulla distribuzione dei dati, gli algoritmi di ML sono in grado di apprendere direttamente dalle informazioni disponibili, adattandosi anche a contesti in cui le relazioni tra le variabili sono altamente complesse o sconosciute.

Il ML trova applicazione in numerosi ambiti della medicina, tra cui la predizione del rischio di sviluppare una malattia, la classificazione di immagini mediche, l'analisi di dati genomici e l'estrazione automatica di informazioni cliniche dalle cartelle sanitarie elettroniche.
Queste applicazioni dimostrano la versatilità di questa tecnologia, che può essere impiegata sia come strumento di supporto alla diagnosi e alla prognosi, sia come mezzo per comprendere i meccanismi biologici che stanno alla base delle patologie.

L’adozione del ML in ambito sanitario comporta vantaggi significativi, come la possibilità di analizzare grandi quantità di dati in tempi ridotti e di individuare relazioni nascoste tra le variabili. Tuttavia, sono presenti anche alcune limitazioni, tra cui la necessità di dataset di elevata qualità e dimensione, il rischio di overfitting e la difficoltà di interpretazione dei modelli più complessi, che può limitarne l’applicazione diretta nella pratica clinica. [4,5]

Recentemente, l’impiego di modelli di ML si è rivelato uno strumento promettente per valutare la correlazione tra prediabete e complicanze cardiovascolari. Tali modelli, integrando dati clinici, biochimici e demografici, consentono di identificare soggetti a rischio elevato e di prevedere l’insorgenza di eventi cardiovascolari, supportando così decisioni cliniche più mirate e personalizzate. [6]

1.4 Prediabete: definizione, criteri diagnostici e parametri di valutazione

L’esame della glicemia (concentrazione di glucosio nel sangue) serve a stimare la quantità di glucosio presente nel sangue e, in alcuni casi, nelle urine.
Il glucosio rappresenta la principale fonte di energia per le cellule dell’organismo ed è l’unica fonte di energia per il cervello e per il sistema nervoso.
La misurazione della glicemia è quindi fondamentale per chi desidera mantenere sotto controllo i propri livelli glicemici.

Il prediabete è una condizione in cui i livelli di glucosio nel sangue risultano più alti del normale, ma non abbastanza elevati da soddisfare i criteri diagnostici del diabete mellito di tipo 2 (glicemia più elevata di 7 mmol/L o 126 ml/dl a digiuno per due giorni consecutivi).
La valutazione del prediabete si basa sull’analisi di diversi parametri glicemici, che riflettono il metabolismo del glucosio in condizioni basali e dopo uno stimolo. Questi indicatori consentono di identificare precocemente le alterazioni del

controllo glicemico e di classificare i soggetti a rischio di progressione verso il diabete di tipo 2.

La curva glicemica, o test da carico orale di glucosio (OGTT), è un esame che valuta la capacità dell'organismo di metabolizzare il glucosio.
Il test prevede una misurazione iniziale della glicemia a digiuno (FPG), che rappresenta la concentrazione di glucosio nel sangue dopo almeno 8 ore di digiuno, seguita dalla somministrazione orale di una soluzione contenente una quantità standard di glucosio (circa 75 g).
Successivamente vengono effettuati prelievi di sangue a intervalli regolari, generalmente dopo 1 e 2 ore, per osservare l'andamento della glicemia nel tempo.

Un ulteriore parametro clinico molto utilizzato è l'emoglobina glicata.
L'emoglobina glicata (HbA1c) è un esame del sangue che misura l'andamento medio delle glicemie negli ultimi tre mesi.
Si forma dalla reazione del glucosio con l'emoglobina, una proteina contenuta nei globuli rossi che trasporta l'ossigeno.
Quanto più elevata è la concentrazione di glucosio nel sangue, tanto maggiore sarà la quantità di emoglobina glicata.
Ne esistono due forme, HbA1 e HbA1c, ma per uso clinico si utilizza la seconda poiché è più stabile: una volta formata, resta nel sangue per circa tre mesi e permette di stimare i valori medi della glicemia negli ultimi 90 giorni. [7]

Il prediabete comprende due condizioni principali che rappresentano uno stadio intermedio tra la normalità metabolica e il diabete mellito di tipo 2:
-Impaired Fasting Glucose (IFG).
-Impaired Glucose Tolerance (IGT).
Entrambe le condizioni indicano una difficoltà dell'organismo nel mantenere normali livelli di glucosio nel sangue, ma si manifestano in momenti diversi del metabolismo glucidico.

L’IFG viene diagnosticata quando la glicemia a digiuno, misurata dopo almeno 8 ore senza assumere cibo, risulta compresa tra 5.6 e 6.9 mmol/L.
Questa alterazione indica che, durante il digiuno, il fegato produce e rilascia una quantità di glucosio superiore al necessario, anche quando i livelli di glucosio nel sangue sono già adeguati.

L’IGT, invece, si riscontra quando la glicemia due ore dopo OGTT è compresa tra 7.8 e 11 mmol/L.
Questa condizione indica un difetto nella risposta insulinica postprandiale, cioè una ridotta capacità dei muscoli e dei tessuti periferici di assorbire il glucosio dopo un pasto.
Di conseguenza, il glucosio resta più a lungo nel sangue, anche se i valori a digiuno possono apparire normali. [8]
In condizioni di normoglicemia, la glicemia a digiuno (FPG) è inferiore a 5.6 mmol/L, la glicemia a due ore dal test da carico orale di glucosio (OGTT) è inferiore a 7.8 mmol/L, e i valori di emoglobina glicata (HbA1c) sono inferiori a 5.7% (39 mmol/mol).

Nel prediabete, invece, la glicemia a digiuno si colloca tra 5.6 e 6.9 mmol/L, la glicemia post-carico (OGTT) tra 7.8 e 11.0 mmol/L, e i valori di HbA1c tra 5.7% e 6.4% (39 – 47 mmol/mol).

Considerato come uno stato intermedio tra normoglicemia e diabete mellito di tipo 2, il prediabete è caratterizzato da un equilibrio glicemico instabile e da un rischio aumentato di sviluppare disfunzioni metaboliche e complicanze cardiovascolari. [9]

1.5 Epidemiologia e importanza clinica del prediabete

Studiare il prediabete oggi è di cruciale importanza per vari motivi, tra questi il prediabete interessa una quota molto ampia della popolazione adulta a livello

globale: recenti stime indicano che la prevalenza di Impaired Glucose Tolerance (IGT) nel 2021 era di circa il 9.1% della popolazione adulta (circa 464 milioni di persone) e si prevede un aumento fino al 10% (circa 638 milioni) entro il 2054, mentre l'Impaired Fasting Glucose (IFG) nel 2021 era di circa 5.8% (circa 298 milioni) con proiezioni in aumento. [10]

Da evidenze epidemiologiche è emerso che le persone con prediabete presentano un aumento significativo del rischio di mortalità per tutte le cause e una maggiore incidenza di malattie cardiovascolari. Una meta-analisi ha evidenziato un incremento del rischio di mortalità globale pari al 13% e di eventi cardiovascolari del 15% rispetto a soggetti con normoglicemia. Inoltre, il prediabete è associato a complicanze metaboliche e microvascolari precoci, come microalbuminuria, neuropatie delle piccole fibre e alterazioni retiniche, che rappresentano l'inizio di un danno d'organo già nelle fasi iniziali della disfunzione glicemica. [11]

Le cause del prediabete non sono ancora del tutto note, ma la predisposizione genetica e la familiarità con il diabete svolgono un ruolo determinante, ma non unico: sovrappeso, obesità, sedentarietà e dieta ad alto contenuto di zuccheri semplici contribuiscono in modo significativo all'alterazione del metabolismo glucidico.

Riconoscere precocemente il prediabete riveste un'importanza cruciale, poiché rappresenta un campanello d'allarme clinico: l'intervento mediante strategie di prevenzione quali la modifica dello stile di vita e del regime alimentare potrebbe ridurre la probabilità di evoluzione verso il diabete di tipo 2 fino al 58%. [12]

Negli ultimi anni, l'attenzione verso l'identificazione precoce del prediabete è cresciuta notevolmente, grazie allo sviluppo di tecniche diagnostiche sempre più sensibili e di facile applicazione. Oltre ai classici test di laboratorio, come la glicemia a digiuno, il test da carico orale di glucosio e la determinazione

dell'emoglobina glicata, la ricerca ha introdotto strumenti di valutazione più sofisticati, quali la misurazione della resistenza insulinica tramite HOMA-IR e il monitoraggio continuo della glicemia (CGM).
In quest'ottica il prediabete non rappresenta soltanto un oggetto di studio, ma una vera e propria sfida clinica e preventiva di rilevante importanza per la medicina moderna. [13,14]

1.6 Eziologia e fisiopatologia del prediabete

Ad oggi, nonostante la causa esatta del prediabete non sia ancora del tutto nota, la familiarità e la predisposizione genetica svolgono un ruolo fondamentale nello sviluppo della condizione.
Quello che è noto è che le persone con prediabete non metabolizzano più correttamente il glucosio. Di conseguenza, il glucosio si accumula nel sangue invece di essere immagazzinato nelle cellule muscolari e in altri tessuti. [15]

Nel prediabete, il metabolismo del glucosio risulta compromesso a causa di una ridotta secrezione di insulina o di una diminuita sensibilità dei tessuti periferici alla sua azione. L'insulina è un ormone prodotto dalle cellule β delle isole di Langerhans pancreatiche che regola i livelli di glucosio nel sangue favorendone l'assorbimento da parte delle cellule. Quando questo meccanismo si altera, come accade nel prediabete, il pancreas non riesce a compensare adeguatamente la resistenza insulinica e il glucosio tende ad accumularsi nel circolo sanguigno, determinando una condizione di iperglicemia lieve. [16]

1.7 Epidemiologia in Danimarca e possibilità che il prediabete diventi diabete di tipo 2

In Danimarca, numerosi studi hanno approfondito il fenomeno stimando che tra 271000 e 292000 adulti danesi, pari al 6.6 - 6.9% della popolazione tra i 20 e gli

85 anni, presentano valori di HbA1c indicativi di prediabete. Inoltre, circa 57000-61000 persone (1.4% della popolazione adulta) risultano affette da diabete di tipo 2 non diagnosticato, rappresentando circa un quarto del totale dei casi di diabete.
Questi dati suggeriscono che, nonostante la rete sanitaria danese garantisca un'elevata accessibilità ai controlli, non tutti i soggetti a rischio vengono identificati precocemente. La scelta del parametro diagnostico influenza in modo significativo le stime: l'impiego del criterio basato su HbA1c, rispetto all'OGTT, tende infatti a ridurre la percentuale di soggetti classificati come prediabetici, per cui è necessario interpretare i dati tenendo conto della metodologia utilizzata. [17]

In altri studi basati su dati di laboratori nazionali raccolti tra il 2012 e il 2018, hanno riportato una prevalenza di prediabete del 7.1%, con un'incidenza di 14.2 nuovi casi ogni 1000 persone-anno. L'età media alla diagnosi era di circa 67 anni, e la copertura dei test HbA1c raggiungeva il 70.8% della popolazione adulta, a conferma dell'elevato livello di monitoraggio sanitario del paese.
Inoltre è stato osservato che circa il 21.3% delle persone danesi con prediabete sviluppa diabete di tipo 2 entro cinque anni, considerando anche la mortalità come possibile evento concorrente. Il rischio di progressione aumenta con i valori iniziali di HbA1c: passa infatti da circa l'11.7% nei soggetti con HbA1c di 42 mmol/mol (6%) fino a quasi il 60% nei soggetti con HbA1c di 47 mmol/mol (6.4%). Allo stesso tempo, la mortalità a cinque anni tra le persone con prediabete è risultata pari al 17.5%, indicando che, soprattutto nelle fasce di età più avanzate, una parte dei soggetti può andare incontro a morte prima di sviluppare il diabete conclamato.[18]

Questi risultati evidenziano l'importanza di riconoscere e gestire il prediabete come una condizione clinicamente significativa, attuando strategie di prevenzione efficaci per rallentare la progressione verso il diabete di tipo 2 e ridurre il rischio di complicanze correlate.

1.8 Associazione tra prediabete e complicanze cardiovascolari e microvascolari

Numerose evidenze scientifiche dimostrano che il prediabete si associa non solo a un rischio aumentato di diabete di tipo 2, ma anche a complicanze cardiovascolari e microvascolari già nelle fasi iniziali della disfunzione glicemica.
Per quanto riguarda il rischio cardiovascolare, diversi studi epidemiologici hanno evidenziato che il prediabete è correlato a un'aumentata incidenza di malattie coronariche, ictus e mortalità cardiovascolare.

È stata osservata un'elevata prevalenza di microalbuminuria e macroalbuminuria nei soggetti con prediabete, segno di una nefropatia che può progredire nel tempo. Inoltre sono stati documentati nelle fasi prediabetiche anche alterazioni del filtrato glomerulare e fenomeni di iperfiltrazione renale. [19]
Anche il sistema nervoso periferico e autonomo può essere interessato dal prediabete. In questa condizione si possono osservare una riduzione della variabilità della frequenza cardiaca, alterazioni della sudorazione e neuropatie delle piccole fibre. Questi disturbi, un tempo ritenuti tipici solo del diabete, indicano che il danno ai nervi può comparire già nelle fasi iniziali di alterazione del metabolismo glucidico. [20]

Anche il microcircolo retinico può mostrare segni di alterazione nei soggetti con prediabete. In studi condotti su popolazioni prediabetiche si è evidenziata una maggiore frequenza di retinopatia già in questa fase precoce. Tra le modificazioni osservate vi sono un aumento del calibro delle arteriole e una riduzione del rapporto tra arteriole e venule, segni che indicano un iniziale coinvolgimento vascolare anche in assenza di diabete conclamato. [21]

Dunque il prediabete emerge come una condizione clinicamente significativa e multifattoriale, che non solo predispone al diabete di tipo 2, ma può già

determinare danni d'organo precoci. Per questo motivo, il riconoscimento e la gestione tempestiva, mediante modifiche dello stile di vita, monitoraggio periodico e, in casi estremi, interventi farmacologici, risultano fondamentali per ridurre il rischio di progressione e l'insorgenza delle sue complicanze metaboliche, cardiovascolari e microvascolari.

2. MATERIALI E METODI

2.1 Panoramica dello studio

Ho svolto il seguente lavoro di tesi utilizzando i dati real-world provenienti dal Danish National Patient Registry (DNPR).

Considerata l'elevata sensibilità delle informazioni trattate, l'intero progetto è stato sviluppato esclusivamente all'interno dell'infrastruttura protetta di Statistics Denmark, che garantisce un ambiente di analisi dati sicuro e rispetta le normative sulla protezione dei dati. I dati sono stati analizzati in modo da non rendere le persone identificabili e senza la possibilità di scaricare informazioni personali.

Ho condotto uno studio osservazionale retrospettivo, identificando una coorte di pazienti nel periodo compreso tra il 2010 e il 2015; tale scelta è stata concordata per garantire un adeguato periodo di follow-up e, contemporaneamente, contenere la numerosità della popolazione in studio (*Figura 2*), semplificando la gestione computazionale e l'elaborazione dei dati.

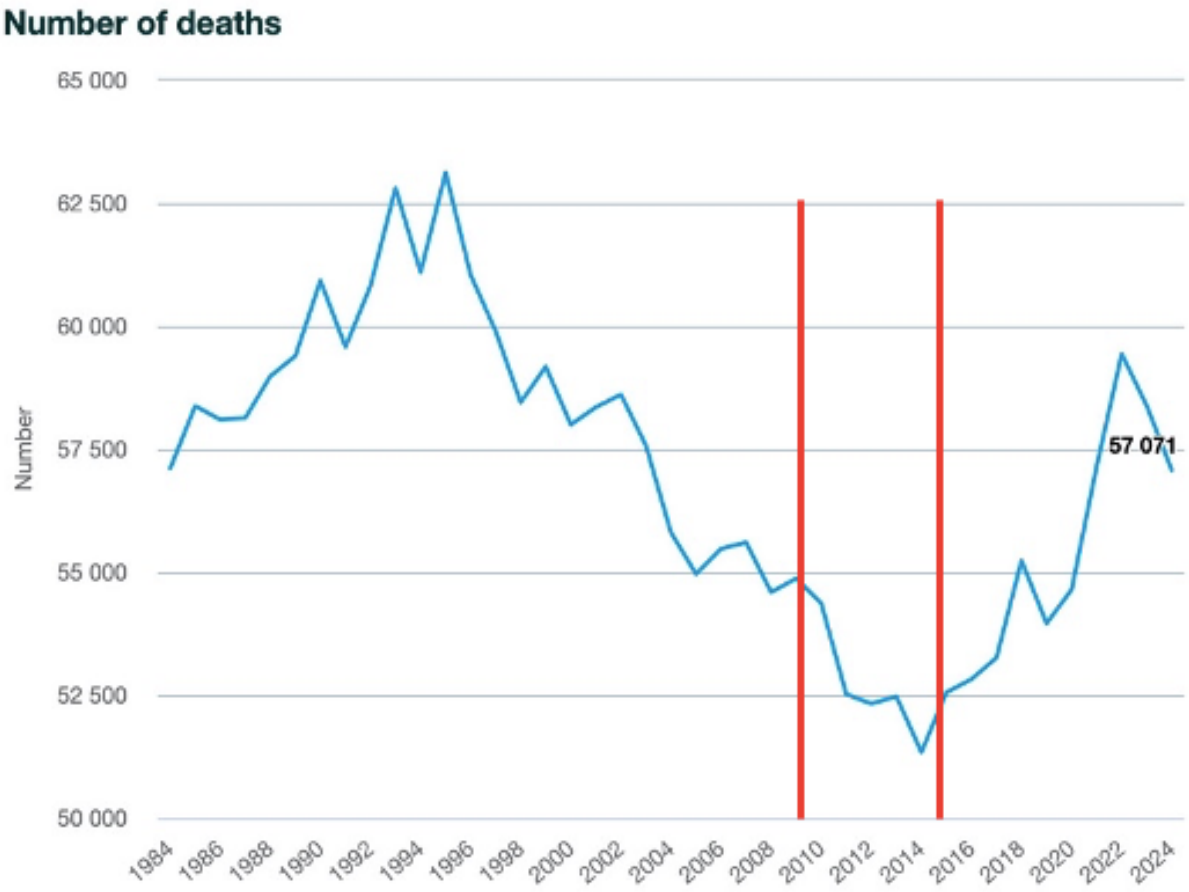

Figura 2: Andamento storico del numero totale di decessi registrato nell'arco di quattro decenni, dal 1984 al 2024. Si possono notare le variazioni della mortalità nel corso degli anni.

(Grafico proveniente da: https://www.dst.dk/en/Statistik/emner/borgere/befolkning/doedsfald)

Il quadro sperimentale è stato progettato con l'obiettivo di sviluppare, analizzare e validare modelli di Machine Learning (ML) per la previsione del rischio di eventi cardiovascolari (CVD) a partire da una popolazione di pazienti con prediabete.

L'intero processo metodologico, dall'acquisizione dei dati grezzi fino alla validazione dei modelli predittivi, è schematizzato in *Figura 3*. Come illustrato nel diagramma di flusso, il piano di lavoro è stato strutturato in cinque macrofasi, necessarie per trasformare le informazioni amministrative provenienti dai registri danesi in uno strumento predittivo affidabile per il rischio cardiovascolare.

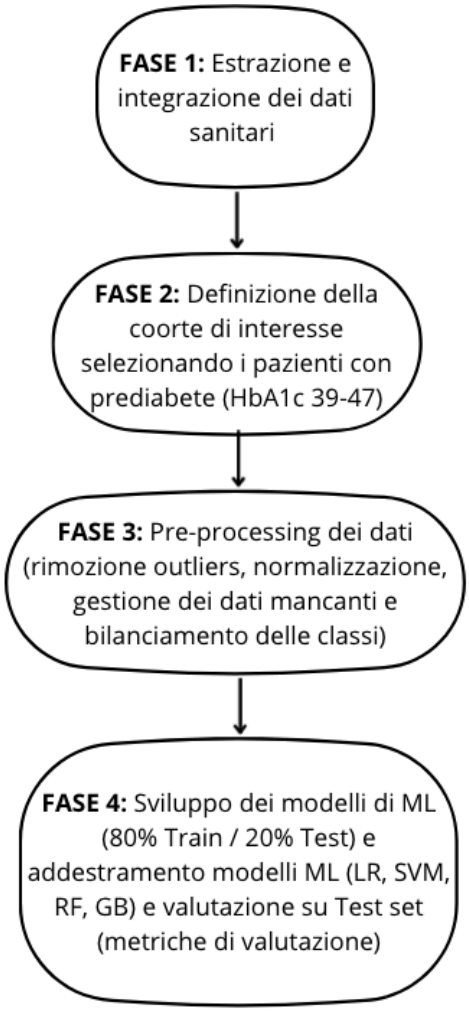

Figura 3: Diagramma di flusso della pipeline metodologica dello studio.

2.2 Fonte dei dati

Per la costruzione della coorte di studio e per l'arricchimento delle variabili cliniche e demografiche, ho effettuato un processo di linkage tra diversi registri nazionali, sfruttando l'identificativo personale univoco (CPR).
In particolare, ho utilizzato i seguenti database:

-Befolkningen (Registri della popolazione):
Archivio amministrativo nazionale contenente informazioni demografiche di base su tutti i residenti in Danimarca, tra cui sesso, data di nascita, stato civile e stato di

residenza. Questo registro è stato utilizzato per ottenere le variabili anagrafiche necessarie alla definizione della coorte.

-Dødsårsag (Registro delle cause di morte):
Registro nazionale che documenta tutti i decessi avvenuti in Danimarca. Include la causa primaria e le eventuali cause secondarie di morte, codificate secondo la classificazione ICD-10. È stato utilizzato per identificare i pazienti che hanno sviluppato l'evento cardiovascolare e per identificare la data di decesso.

-Blood Samples (Registri di laboratorio):
Database contenente i risultati delle analisi di laboratorio effettuate all'interno del sistema sanitario danese. In questo studio è stato utilizzato per identificare i valori di HbA1c necessari alla selezione dei pazienti con prediabete e alla definizione della data di diagnosi del prediabete.

-Lægemiddeldatabasen (Registro nazionale delle prescrizioni):
Registro nazionale delle dispensazioni farmaceutiche effettuate presso le farmacie territoriali. Contiene informazioni sui farmaci erogati, classificati secondo il sistema ATC. È stato impiegato per l'identificazione dei trattamenti farmacologici assunti nell'anno precedente all'ingresso nello studio.

-Landspatientregistret:
Registro che raccoglie informazioni relative ai pazienti trattati negli ospedali danesi, incluse diagnosi e procedure codificate secondo la classificazione ICD-10. È stato utilizzato per l'estrazione delle comorbilità pregresse rilevanti ai fini dell'analisi.
L'integrazione di questi registri ha consentito la costruzione di un dataset strutturato e multidimensionale, comprendente variabili demografiche, cliniche, laboratoristiche, farmacologiche e di outcome, fondamentale per lo sviluppo dei modelli predittivi descritti nelle sezioni successive.

2.3 Pipeline analitica e strumenti computazionali

I database originali, forniti in formato SAS, sono stati inizialmente supervisionati utilizzando il software SAS Enterprise Guide. Le query implementate in tale ambiente hanno guidato la fase preliminare di ispezione dell'intera mole di dati.

Data l'elevata mole computazionale, le successive fasi di analisi, pre-processing e modellazione sono state eseguite operando su macchine remote fornite dal DST (Statistics Denmark), utilizzando il linguaggio di programmazione Python, precisamente con l'ambiente di sviluppo Spyder.

I dati sono stati importati ed elaborati tramite le librerie Pandas e NumPy. Nello specifico, la struttura a dataframe di Pandas ha permesso di gestire le fasi di pre-processing, quali l'integrazione delle diverse tabelle mediante le operazioni di merging, il filtraggio della coorte di studio e la gestione dei valori mancanti. Parallelamente NumPy ha consentito di trasformare i dati tabellari strutturati in matrici e vettori numerici.

Successivamente, è stata condotta una fase di validazione per accertare la coerenza epidemiologica dei dati estratti. Al fine di verificare la correttezza delle query e delle operazioni di merging, le statistiche descrittive della coorte (quali il numero dei decessi e le frequenze di prescrizione dei farmaci) sono state confrontate con i dati forniti dai registri ufficiali di Statistics Denmark e dal database nazionale medstat.dk.
Questo passaggio è stato fondamentale per controllare la qualità dei dati estratti, garantendo l'assenza di bias di estrazione, certificando l'affidabilità del dataset prima di procedere alla modellazione.
Per lo sviluppo, l'addestramento e l'ottimizzazione degli algoritmi di Machine Learning è stata impiegata la libreria Scikit-learn.

Quest'ultima è stata utilizzata sia per la normalizzazione delle features tramite Min-Max Scaling, sia per l'ottimizzazione dei modelli: la ricerca degli iperparametri è stata condotta mediante GridSearchCV, supportata da una validazione incrociata di tipo Stratified K-Fold per mantenere la reale distribuzione della variabile target in ogni iterazione.
Le tecniche di ricampionamento per la gestione dello sbilanciamento delle classi sono state implementate tramite il pacchetto Imbalanced-learn, applicando tecniche di Undersampling e come ulteriore approccio sono state impiegato le tecniche di Balanced Bagging.
Infine, per la generazione dei grafici e la visualizzazione delle metriche di performance ho utilizzato le librerie Matplotlib e Seaborn.

2.4 Costruzione della coorte di studio

La selezione della coorte di studio è iniziata dall'analisi del dataset "lab_dm_forsker.sas7bdat", derivato dal database Blood Samples precedentemente descritto. All'interno di questo database, la classificazione delle analisi del sangue è strutturata mediante una specifica variabile denominata "ANALYSIS_CODE". Tale variabile è stata importantissima per isolare, tra le innumerevoli tipologie di test, i valori dell'emoglobina glicata (HbA1c), il cui monitoraggio rappresenta il parametro clinico fondamentale per l'inclusione dei soggetti con prediabete.

L'identificazione dei soggetti di interesse è avvenuta applicando i seguenti passaggi:

-Filtraggio del biomarcatore: ho isolato i record relativi all'emoglobina glicata (HbA1c) utilizzando il codice analisi "NPU27300".
-Criteri di esclusione: ho implementato un filtro per l'esclusione dei pazienti con diagnosi di diabete. Nello specifico, i soggetti con valori di HbA1c > 47 mmol/mol

(soglia diagnostica per il diabete di tipo 2) sono stati rimossi per garantire che la coorte finale non includesse pazienti con diabete di tipo 2.

-Criteri di inclusione: tra i soggetti candidati, ho selezionato coloro che presentavano valori di HbA1c compresi nell'intervallo 39 – 47 mmol/mol.
-Finestra temporale e data di diagnosi: ho considerato valide le misurazioni effettuate nel periodo compreso tra il 01/01/2010 e il 31/12/2015. Per i pazienti con misurazioni multiple nel range prediabetico, la prima rilevazione utile è stata definita come data di ingresso nello studio (INDEX_DATE).

Una volta definita la popolazione target, ho recuperato le informazioni demografiche (sesso e data di nascita) mediante un'operazione di linkage con i registri della popolazione (bef files) relativi agli anni 2010 - 2015. L'analisi dei file anagrafici è stata iterata annualmente applicando il filtro "STATSB == 5100", al fine di includere nello studio esclusivamente i cittadini danesi. Per ogni CPR (Civil Personal Registration number) identificato, ho estratto le variabili (KOEN - Sesso e FOED_DAG - Data di nascita).

Al fine di arricchire il dataset e caratterizzare la coorte, ho recuperato le informazioni cliniche pregresse. Per ciascun paziente ho definito una finestra di osservazione temporale di 365 giorni antecedenti la data di inclusione.

Le comorbidità sono state identificate mediante codici ICD-10, focalizzandosi sulle seguenti condizioni di interesse cardiovascolare e metabolico [22]:

-Ipertensione e obesità: rappresentano i classici fattori di rischio primari. L'ipertensione causa un danno meccanico continuo alle pareti dei vasi sanguigni e affatica il cuore, mentre l'obesità porta problemi di natura metabolica, cardiocircolatoria e strutturale all'interno dell'organismo.

-Malattia renale cronica (CKD): il deterioramento della funzione renale causa calcificazioni vascolari, mantenimento dei liquidi e ipertensione, aumentando esponenzialmente il rischio cardiovascolare.

-Fibrillazione atriale: aumenta severamente il rischio di ictus e scompensi cardiaci.

-Cardiopatia ischemica (IHD): è una malattia cardiaca che colpisce le coronarie, incapaci di apportare sangue ed ossigeno al cuore a causa di un restringimento progressivo.

-Broncopneumatia cronica ostruttiva (BPCO): dal momento che il cuore e i polmoni lavorano in serie e la BPCO riduce l'ossigeno nel sangue, ciò causa affaticamento del lato destro del cuore e un danneggiamento irreversibile degli alveoli polmonari.

Analogamente, è stata effettuata una mappatura dei trattamenti farmacologici. Ho identificato le prescrizioni dispensate nell'anno precedente l'ingresso nello studio, raggruppandole in sette classi principali tramite codici ATC:

-Inibitori RAAS, beta-bloccanti, calcio-antagonisti e diuretici: dall'assunzione di questi farmaci si ha la conferma che il paziente è un soggetto iperteso e quindi indica condizioni cardiache ad altissimo rischio.

-Ipolipemizzanti (statine): indicano la presenza di displipidemia (colesterolo alto) in un soggetto.

-Antitrombotici: l'assunzione di questi farmaci classifica il paziente come soggetto ad alto rischio trombotico, magari a causa della fibrillazione atriale.

-Ipoglicemizzanti: classe di farmaci specifici per pazienti diabetici, poiché consentono all'ormone insulina di lavorare meglio.

Tali variabili sono state definite come variabili binarie (1 = presenza, 0 = assenza).

2.5 Definizione della variabile target

La definizione della variabile dipendente dello studio (EVENT_CVD) è stata effettuata per identificare la mortalità specifica per cause cardiovascolari all'interno della coorte.

Per la costruzione di tale variabile, il dataset è stato elaborato come segue:

-Ai soggetti deceduti ho associato le informazioni patologiche, nello specifico la causa primaria di decesso (C_DOD_1A) e la data dell'evento (D_DODSDATO).

-Per i soggetti viventi (o non presenti nel registro di mortalità), tali campi sono stati valorizzati come dati mancanti (NaN), indicando l'assenza dell'evento fatale al termine del follow-up.

La variabile target binaria è stata derivata applicando un algoritmo decisionale basato sulla classificazione internazionale delle malattie (ICD-10). È stata definita una lista di codici target (https://icd.who.int/browse10/2019/en#/IX) identificativi degli eventi cardiovascolari di interesse.
L'algoritmo di classificazione ha seguito la seguente logica:

-Outcome = 1 (Evento CVD): il paziente è presente nel registro dei decessi e la causa primaria di morte (C_DOD_1A) corrisponde a uno dei codici ICD-10 inclusi nella lista predefinita.

-Outcome = 0 (Nessun evento / Altro evento): il paziente risulta vivo al termine dello studio, oppure è deceduto per cause non cardiovascolari (codice di morte non presente nella lista target).

Tabella 1: Qualche riga di esempio del dataset ottenuto

(Per privacy non posso riportare il CPR, ma è una stringa numerica del tipo: 123456789)

PATIENT_CPR	SEX	AGE_AT_INDEX	HBA1C	INDEX_DATE	C_OD_1A	D_DODSDATO	EVENT_CVD
xxxxxxxxxx	2.0	68	39.0	2014-02-24			0
yyyyyyyyyy	1.0	81	42.0	2011-07-12	I619	2015-03-10	1
zzzzzzzzzz	2.0	66	40.0	2013-12-11	R990	2018-09-01	0

2.6 Pre-processing dei dati

Il pre-processing costituisce una fase cruciale nella pipeline del machine learning per garantire che i classificatori possano estrarre dei pattern validi ed evitare alterazioni predittive. [23]

Nello specifico, le operazioni di pre-processing sono state suddivise nelle seguenti fasi, mirate a risolvere uno specifico problema.

2.6.1 Gestione degli outliers

In prima istanza, ho affrontato la gestione degli outliers, in quanto la loro presenza può inficiare la robustezza degli algoritmi di apprendimento e distorcere i risultati delle analisi statistiche. In questo studio ho optato per la rimozione delle istanze anomale ritenendo che tale operazione non comportasse una significativa perdita

di contenuto informativo. La procedura di identificazione ed esclusione utilizzata è il metodo dell'intervallo interquartile (IQR).

Il metodo IQR rappresenta una tecnica robusta per l'identificazione degli outliers. Un'osservazione viene classificata come anomala se ricade al di fuori del seguente intervallo di accettabilità:

-Limite inferiore: Q1 – 1.5 * IQR

-Limite superiore: Q3 + 1.5 * IQR

2.6.2 Normalizzazione dei dati

Considerata l'eterogeneità delle scale di misura e delle unità delle variabili, si è proceduto alla normalizzazione dei dati. Tale trasformazione è fondamentale per riportare le feature su una scala comune, evitando che variabili con range numerici più ampi dominino la funzione obiettivo del modello o influenzino negativamente algoritmi basati sulle distanze. Nello specifico, ho applicato una normalizzazione di tipo Min-Max Scaling, mappando tutti i valori nell'intervallo [0,1].

Questa tecnica, a differenza di altri metodi quali Z-score o decimal scaling, permette di aumentare l'accuratezza degli algoritmi, minimizzando il costo computazionale e i tempi di training. [24]

2.6.3 Trattamento dei missing values

Il controllo della completezza del dataset è stato effettuato trasversalmente durante l'intera fase di costruzione e integrazione dei dati. Diverse strategie per il trattamento dei valori mancanti esistono [25], tra cui:

-Mantenimento del dato mancante (se l'algoritmo lo supporta).

-Imputazione statistica (sostituzione con media, moda o mediana).

-Imputazione stocastica (utilizzo di valori casuali rispettando la distribuzione).

-Imputazione model-based (utilizzo di modelli predittivi, come alberi decisionali o k-NN, per stimare il valore mancante).

Tuttavia, l'analisi preliminare ha evidenziato una scarsa incidenza di valori nulli nelle variabili di interesse. Pertanto, ho scelto di procedere alla rimozione delle istanze incomplete. Un controllo di integrità finale ha confermato l'assenza di dati mancanti nel dataset definitivo destinato all'analisi.

2.6.4 Gestione dello sbilanciamento tra le classi

L'analisi esplorativa della variabile target ha rivelato un marcato sbilanciamento delle classi, come mostrato in *Figura 4:*

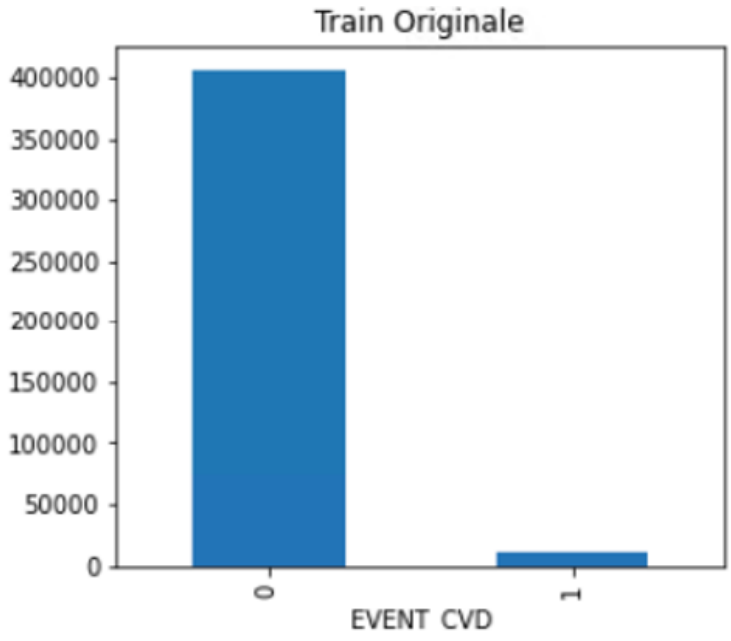

Figura 4: Distribuzione delle frequenze assolute nella variabile target (EVENT_CVD) nel training set originale. L'istogramma evidenzia lo sbilanciamento tra la classe 0 (soggetti sani, classe maggioritaria) e la classe 1 (eventi cardiovascolari, classe minoritaria).

Il problema delle classi sbilanciate è una criticità nota nell'apprendimento supervisionato che tende a sviluppare un bias a favore della classe maggioritaria, massimizzando l'accuratezza globale a discapito della capacità di identificare

correttamente la classe minoritaria. Per mitigare tale problematica, sono state valutate le principali tecniche di ricampionamento:

1.Random Undersampling: consiste nel bilanciare il dataset riducendo casualmente il numero di osservazioni appartenenti alla classe maggioritaria.

2.Random Oversampling: prevede la replicazione casuale delle osservazioni della classe minoritaria fino al raggiungimento dell'equilibrio.

3.SMOTE (Synthetic Minority Over-sampling Technique): una tecnica avanzata di sovracampionamento che genera dati sintetici. A differenza della semplice duplicazione, l'algoritmo seleziona un campione della classe minoritaria, identifica i suoi k vicini più prossimi (k Nearest Neighbors) nello spazio delle feature e genera nuovi punti interpolando linearmente tra l'esempio selezionato e i suoi vicini. Questo approccio arricchisce la varietà del dataset evitando l'overfitting tipico della duplicazione esatta.

Nello specifico, la strategia adottata in questo progetto di tesi è stata il Random Undersampling, perché il dataset in esame era molto grande e la classe maggioritaria era enormemente sovrarappresentata. È importante precisare come la tecnica sia stata applicata esclusivamente al set di addestramento (training set). Sebbene questa tecnica comporti l'eliminazione di una porzione di dati della classe maggioritaria — con una potenziale perdita di informazione — essa ha permesso di ottenere una distribuzione bilanciata, riducendo al contempo la complessità computazionale del training.

Al fine di mitigare la grande perdita di istanze a causa dell'Undersampling, si è optato per un approccio basato sui modelli Ensemble. Questa strategia consente di preservare e sfruttare appieno il contenuto informativo della classe maggioritaria. Lo schema, *Figura 5*, prevede l'addestramento in parallelo di molteplici classificatori, fornendo a ciascuno di essi la totalità dei campioni della classe minoritaria combinata con un sottoinsieme casuale, di pari numerosità, campionato

dalla classe maggioritaria. Tale logica iterativa garantisce un bilanciamento perfetto in fase di training per ogni singolo modello, massimizzando allo stesso tempo l'utilizzo di tutti i dati a disposizione.

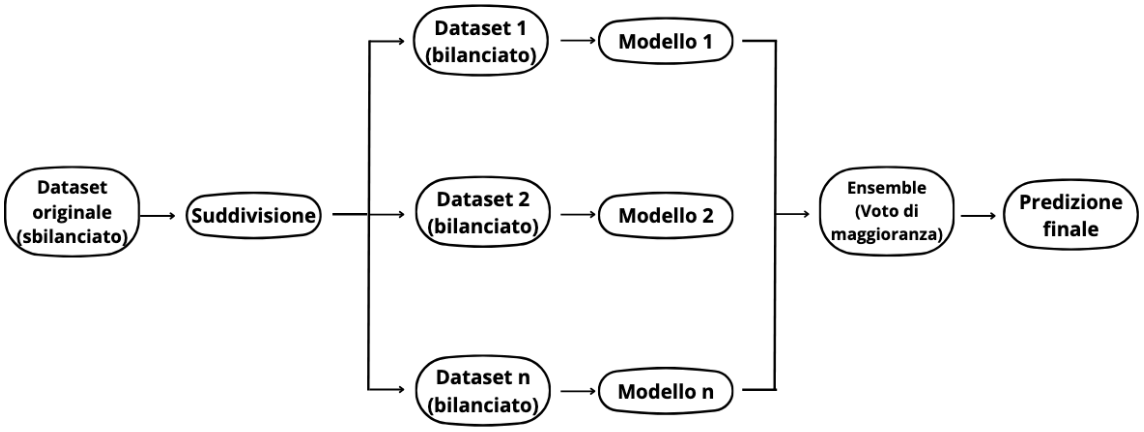

Figura 5: Rappresentazione schematica dell'architettura di Balanced Bagging. Il training set originale, caratterizzato da un forte sbilanciamento, viene sottoposto a diverse iterazioni di random undersampling. Vengono così generati n sotto-campioni perfettamente bilanciati (nel presente studio, n= 20), ciascuno dei quali viene impiegato per addestrare un classificatore indipendente (Random Forest). Le predizioni dei singoli modelli vengono infine aggregate tramite voto a maggioranza per produrre un output finale più robusto e privo di bias verso la classe maggioritaria.

2.7 Sviluppo dei modelli di Machine Learning

L'obiettivo è stato quello di sviluppare e confrontare algoritmi di Machine Learning supervisionato (*Figura 6*) per predire il rischio di eventi cardiovascolari (CVD).

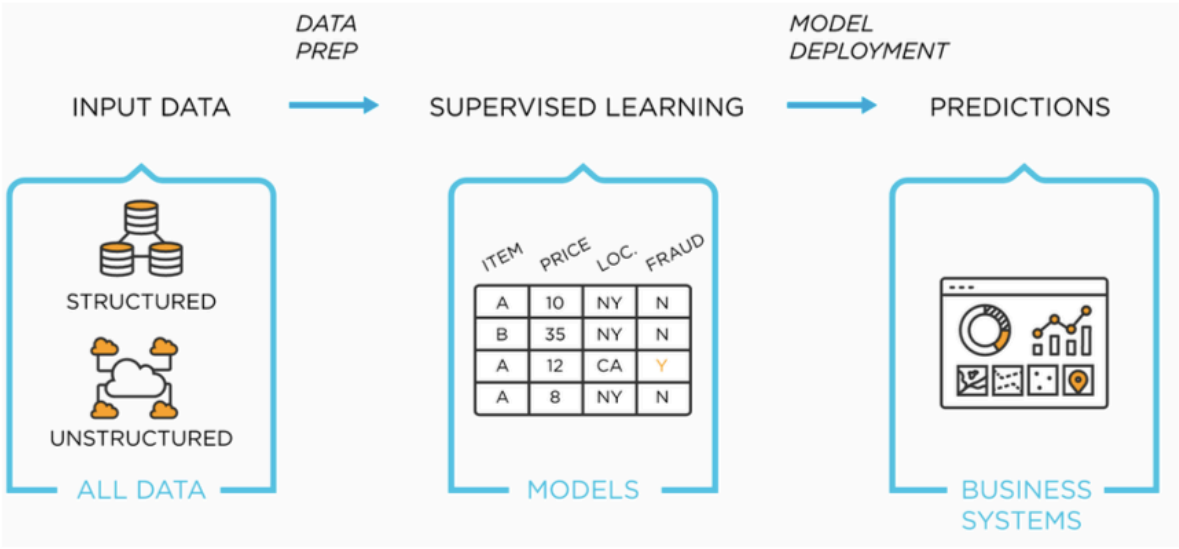

Figura 6: Schema di un sistema di apprendimento supervisionato. Il processo si articola in tre macrofasi: 1.Acquisizione e preparazione; 2.Addestramento; 3.Generazione delle predizioni.

L'obiettivo dell'apprendimento è apprendere una regola di classificazione che permetta di riconoscere una classe osservando i valori degli attributi (*Figura 7*).
In questo lavoro di tesi la classe target è stata riconosciuta attraverso la combinazione di attributi farmacologici e anamnestici del paziente.

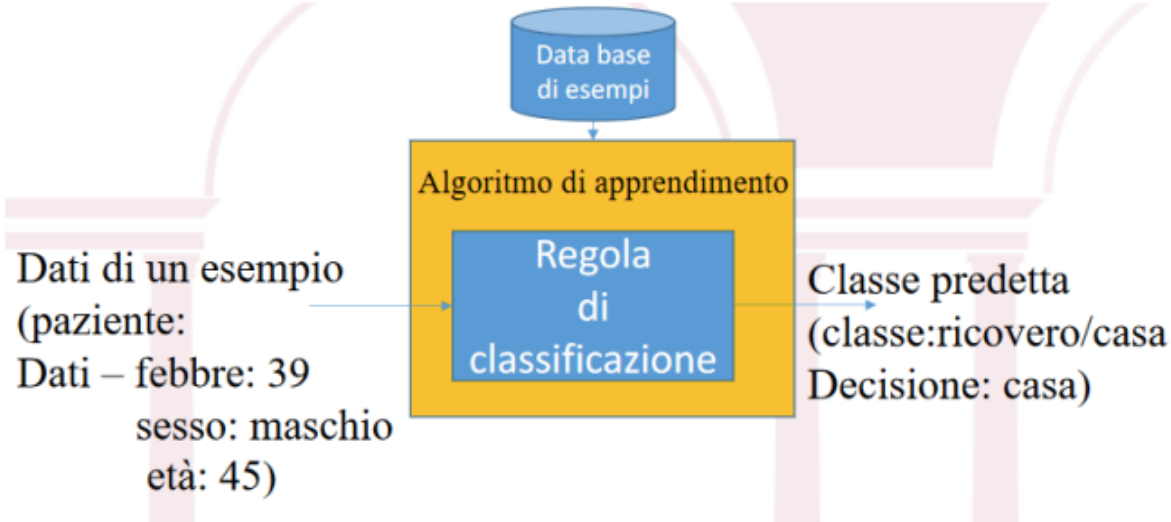

Figura 7: Viene illustrata la logica che sta dietro al funzionamento di un algoritmo di classificazione applicata a un caso clinico. Il sistema riceve in ingresso le feature di un paziente e le elabora attraverso una regola di classificazione che permette di mappare i dati in ingresso verso uno specifico output. Il risultato finale è la predizione della classe di appartenenza del paziente.

L'obiettivo della modellazione è quello di sviluppare algoritmi con un'elevata capacità di generalizzazione, cioè un modello deve essere in grado di effettuare stime accurate e precise su dati nuovi e mai osservati dall'algoritmo, evitando il fenomeno del sovradattamento (overfitting). Per garantire una valutazione oggettiva delle performance, il dataset finale è stato diviso in due sottoinsiemi indipendenti:

-Training set (80%): volto esclusivamente alla fase di apprendimento e addestramento dei modelli.

-Test set (20%): utilizzato unicamente per la validazione finale.

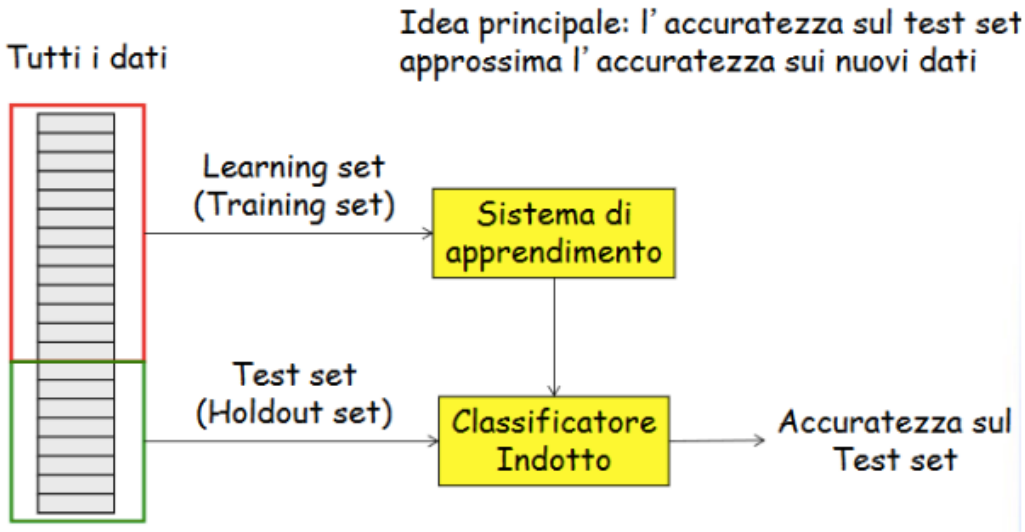

Figura 8: Schema concettuale del partizionamento dei dati. Il dataset originale viene diviso in due sottoinsiemi indipendenti: il training set, impiegato per addestrare l'algoritmo di apprendimento, e il test set, utilizzato esclusivamente per valutare il modello finale. Tale separazione permette di misurare in modo oggettivo la capacità di generalizzazione del classificatore sui dati non osservati in fase di addestramento.

2.8 Scelta dei modelli

Per la fase di classificazione predittiva sono stati implementati e confrontati i seguenti quattro modelli di classificazione:

-La Regressione Logistica, metodo molto utilizzato per analisi multivariata e costruire classificatori.
In ambito epidemiologico costituisce un importante algoritmo per la stima del rischio clinico. Essa è ideata per modellare la probabilità che una popolazione sviluppi una determinata malattia sulla base di specifici fattori di rischio.
Un ulteriore punto di forza è la flessibilità del modello nel gestire dati di diversa natura (continui, binari e categorici) permettendo l'inclusione di importanti variabili per questo studio come l'età, il sesso e comorbilità.

-Le Support Vector Machines (SVM) sono modelli di apprendimento supervisionato appartenenti alla classe dei modelli discriminativi.
A differenza della Regressione Logistica, le SVM affrontano il problema della classificazione cercando un iperpiano ottimale che separi le classi massimizzando il margine tra di esse.

L'idea è quella di trovare il piano che massimizza il margine di separazione, ovvero l'obiettivo principale di una SVM è trovare il miglior iperpiano che separi i dati in classi differenti in modo corretto. I punti che stanno sul margine sono detti Support Vector.

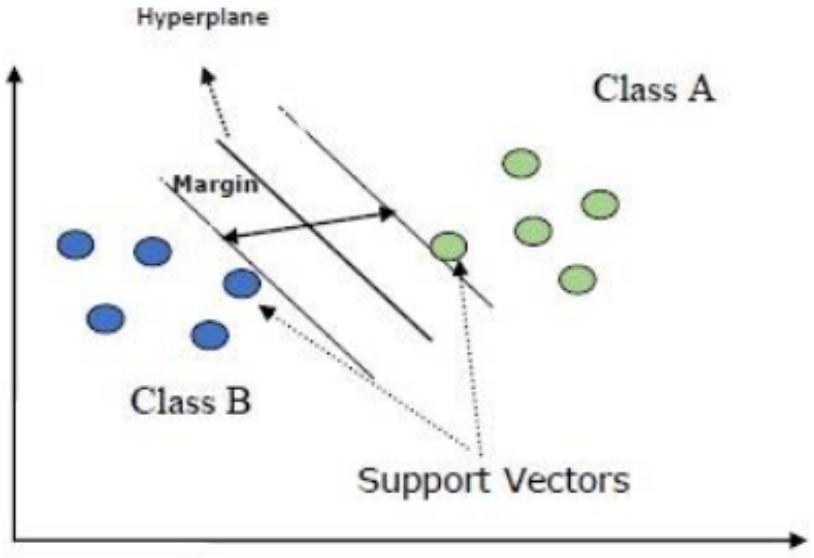

Figura 9: Principio di funzionamento concettuale di una Support Vector Machine (SVM). L'algoritmo individua l'iperpiano ottimale di classificazione massimizzando il margine di separazione tra le due classi. Le singole osservazioni che giacciono esattamente sui confini di tale margine prendono il nome di Support Vectors.

La classificazione può essere lineare e non lineare:

Lineare cioè quando i dati sono linearmente separabili, SVM può trovare facilmente un iperpiano che divide le classi, invece non lineare è quando i dati non sono linearmente separabili, SVM utilizza una tecnica chiamata "kernel trick" per mappare i dati in uno spazio ad alta dimensione dove possono essere separati linearmente.

La scelta è ricaduta sul kernel gaussiano (RBF) in quanto gestisce le relazioni non lineari, e viene considerato universale poiché è in grado di modellare relazioni molto complesse e si adatta a strutture sconosciute dei dati.

-Le Random Forest sono un algoritmo utilizzato per problemi di classificazione e regressione.
È un esempio di tecnica di "ensemble learning", dove l'idea è di combinare le previsioni di diversi modelli deboli per ottenere un modello più forte e accurato. Quindi è un algoritmo di ensemble che migliora la precisione e la robustezza delle previsioni combinando molti alberi di decisione. Utilizzando campionamento casuale sia sui dati che sulle caratteristiche, crea modelli che generalizzano bene ai dati non visti, rendendoli una scelta valida per il nostro problema.

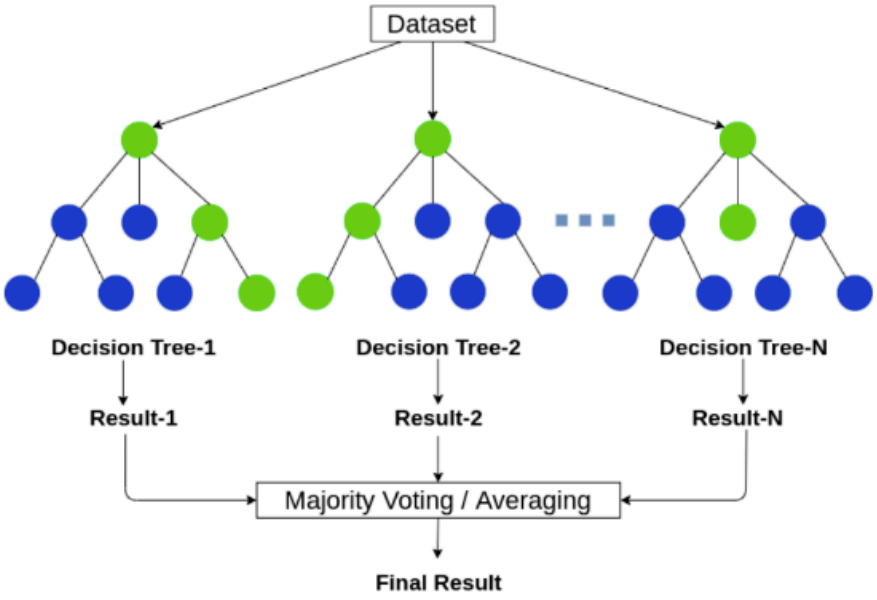

Figura 10: Architettura concettuale dell'algoritmo Random Forest. Il modello, addestra in parallelo una molteplicità di alberi decisionali indipendenti (Decision Tree) sfruttando sotto-campioni del dataset originale. La classificazione finale viene generata aggregando le singole predizioni tramite un sistema di voto a maggioranza, garantendo in tal modo una maggiore robustezza predittiva e una sensibile riduzione al rischio di overfitting.

- Il Gradient Boosting è un algoritmo molto potente che si basa sul combinare predittori deboli, per generare un modello predittivo finale molto robusto e accurato. L'algoritmo è in grado di gestire bene in molti casi sperimentali il compromesso tra bias e varianza, garantendo delle previsioni molto precise ed è in grado di catturare relazioni non lineari e interazioni complesse tra i vari fattori di rischio. [26]

2.9 Metriche di valutazione

La validazione quantitativa dei modelli si è basata sul calcolo di specifiche metriche di valutazione:

-AUC: rappresenta l'area sottesa alla curva ROC e quantifica le performance del modello assumendo un valore compreso tra 0 e 1. Un'AUC pari a 0.5 indica un classificatore casuale, privo di alcuna capacità discriminativa tra le classi, mentre

un valore pari a 1 corrisponde a un modello in grado di separare perfettamente i pazienti a rischio da quelli sani.

-Accuratezza: è il numero di previsioni corrette sul totale delle osservazioni. Viene calcolata come:

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

-Recall (sensibilità): esprime la capacità del classificatore di individuare il totale dei pazienti a rischio. In un problema di predizione in ambito medico caratterizzato da un forte sbilanciamento delle classi, la Recall è molto importante perché un valore elevato indica una forte capacità dell'algoritmo di individuare i pazienti a rischio.

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

-Precision: rappresenta la proporzione di predizioni positive che risultano effettivamente corrette. Esprime quindi la capacità del modello di non generare falsi allarmi (falsi positivi).

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

-F1-Score: è una metrica definita come la media armonica tra Precision e Recall. Viene utilizzata per avere un unico indicatore che bilancia entrambe le misure, risultando utile quando le classi del dataset sono fortemente sbilanciate.

$$F1 = 2\frac{Precision * Recall}{Precision + Recall}$$

-MCC (Matthews Correlation Coefficient): è un indicatore che indica la qualità di un classificatore binario che prende in considerazione tutti i quattro valori della matrice di confusione. Il suo valore è sempre compreso tra -1 e +1:

- +1 indica una classificazione perfetta e concorde con la realtà.

- 0 corrisponde a una classificazione casuale.

- -1 rappresenta una predizione totalmente errata.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(FN + TN)(TP + FN)(FP + TN)}}$$

3. RISULTATI

3.1 Descrizione della coorte finale

L'integrazione di tali dati mi ha permesso di generare il dataset finale utilizzato per le analisi, le cui caratteristiche sono descritte nella seguente tabella.

Tabella 2: N.B. La somma di training set e test set non coincide con il totale perché prima di dividere il dataset in training set e test set, ho rimosso gli outliers e di conseguenza ho circa 10000 dati in meno sul totale.

Caratteristiche	**Totale(n=520408)**	**Train(n=406548)**	**Test(n=104082)**
Males	231334 (44.5%)	180925 (44.5%)	46105 (44.3%)
Females	289074 (55.5%)	225623 (55.5%)	57977 (55.7%)
Age_at_index	64.58 ± 14.73	65.36 ± 13.49	64.58 ± 14.74
HbA1c	40.63 ± 1.77	40.60 ± 1.71	40.63 ± 1.77
Obesity	2754 (0.5%)	2031 (0.5%)	565 (0.5%)
Hypertension	14780 (2.8%)	11723 (2.9%)	2906 (2.8%)
Dyslipidemia	4947 (1.0%)	3935 (1.0%)	967 (0.9%)
COPD	5264 (1.0%)	4119 (1.0%)	1061 (1.0%)
CKD	1856 (0.4%)	1473 (0.4%)	341 (0.3%)
Atrial_fib	8607 (1.7%)	6817 (1.7%)	1692 (1.6%)
IHD	10906 (2.1%)	8657 (2.1%)	2139 (2.1%)
Drug_raas	94425 (18.1%)	75068 (18.5%)	18709 (18.0%)
Drug_beta_blockers	55969 (10.8%)	44402 (10.9%)	11056 (10.6%)
Drug_calcium_blockers	54335 (10.4%)	43108 (10.6%)	10835 (10.4%)
Drug_diuretics	69852 (13.4%)	55395 (13.6%)	13782 (13.2%)
Drug_lipid_modifyng	89855 (17.3%)	71427 (17.6%)	17824 (17.1%)
Drug_antithrombotic	79199 (15.2%)	62770 (15.4%)	15782 (15.2%)
Drug_glucose_lowering	6447 (1.2%)	4898 (1.2%)	1298 (1.2%)

A seguito dell'analisi dell'importanza delle features, eseguita estraendo la Gini Importance (o Mean Decrease in Impurity) dall'algoritmo Random Forest [27], la variabile relativa all'età è risultata avere il maggiore potere predittivo in assoluto per la classificazione del rischio cardiovascolare.

Tabella 3: Top 5 features

Age_at_index	0.586581
HbA1c	0.127712
Sex	0.034536
Drug_antithrombotic	0.034345
Drug_diuretics	0.033903

In *Figura 11* viene mostrata la distribuzione dell'età della popolazione in studio, permettendo di valutare come sono fatti i dati che indirizzano le decisioni dei modelli.

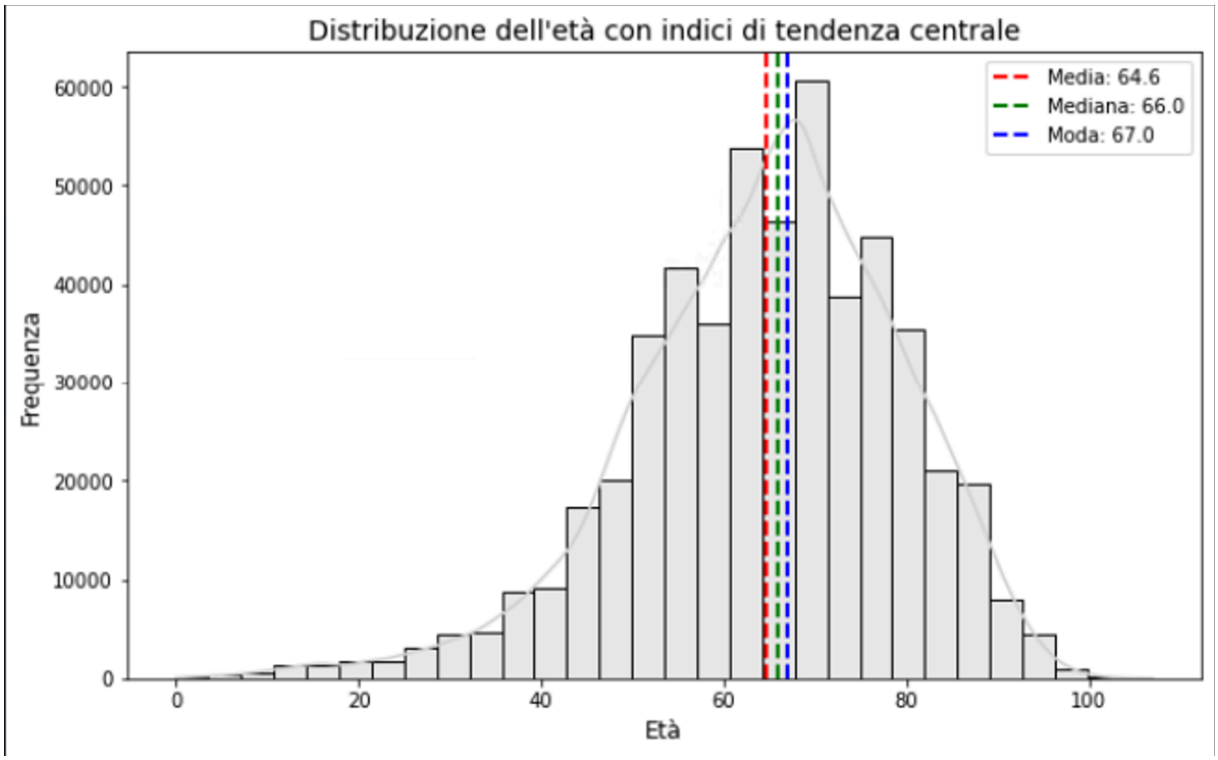

Figura 11: L'istogramma mostra la distribuzione dell'età della coorte in esame con i relativi indici di tendenza centrale: media, mediana e moda.

Dall'osservazione del grafico è stato possibile trarre i seguenti aspetti chiave:

-Forma della distribuzione: la curva rappresenta un andamento unimodale, caratterizzato da un'asimmetria negativa, ovvero il picco della distribuzione è spostato verso destra ad indicare che il peso maggiore del campione sta nelle fasce di età più alte.

-Indici di tendenza centrale: $media < mediana < moda$, la coda si allunga verso sinistra. Significa che la maggior parte dei pazienti è anziana, con pochi casi di pazienti molto giovani che portano la media verso il basso.

-Code della distribuzione: la coda sinistra è più lunga ma molto schiacciata verso l'asse delle ascisse, questo perché i soggetti giovani (al di sotto dei 40 anni) rappresentano una porzione irrilevante del dataset; la coda destra mostra invece un declino rapido dopo i 70 anni, coerentemente con il normale invecchiamento della popolazione.

In *Figura 12* è riportata la distribuzione dei valori dell'emoglobina glicata (HbA1c) all'interno della coorte di studio, contenuta all'intervallo diagnostico del prediabete (39 – 47 mmol/mol). Dall'osservazione dell'istogramma e dei relativi indici di tendenza centrale, emergono le seguenti considerazioni:

-Asimmetria positiva: la distribuzione non è gaussiana, ma presenta un'evidente asimmetria verso destra. Il picco massimo delle osservazioni si concentra sul limite inferiore dell'intervallo, evidenziando come la grande maggioranza dei pazienti prediabetici del dataset si trovi nelle fasi iniziali dell'alterazione glicemica.

-Indici di tendenza centrale: la relazione tra gli indici di tendenza centrale $moda < mediana < media$ conferma la forma della distribuzione. La moda si aggira sul

valore di 39 mmol/mol, indicando che è il valore più frequente. La mediana, pari a 40 mmol/mol, rivela che il 50% dell'intera coorte presenta valori di HbA1c compresi tra 39 e 40 mmol/mol. La media risulta leggermente tirata verso destra (40.6 mmol/mol) a causa della lunga coda della distribuzione.

-Interpretazione clinica: all'avvicinarsi della soglia diagnosticata per il diabete di tipo 2 (> 47 mmol/mol), la frequenza dei pazienti diminuisce notevolmente. Questo andamento decrescente indica che la gran parte della popolazione monitorata viene intercettata nelle primissime fasi del prediabete, non appena i valori superano la soglia di normoglicemia.

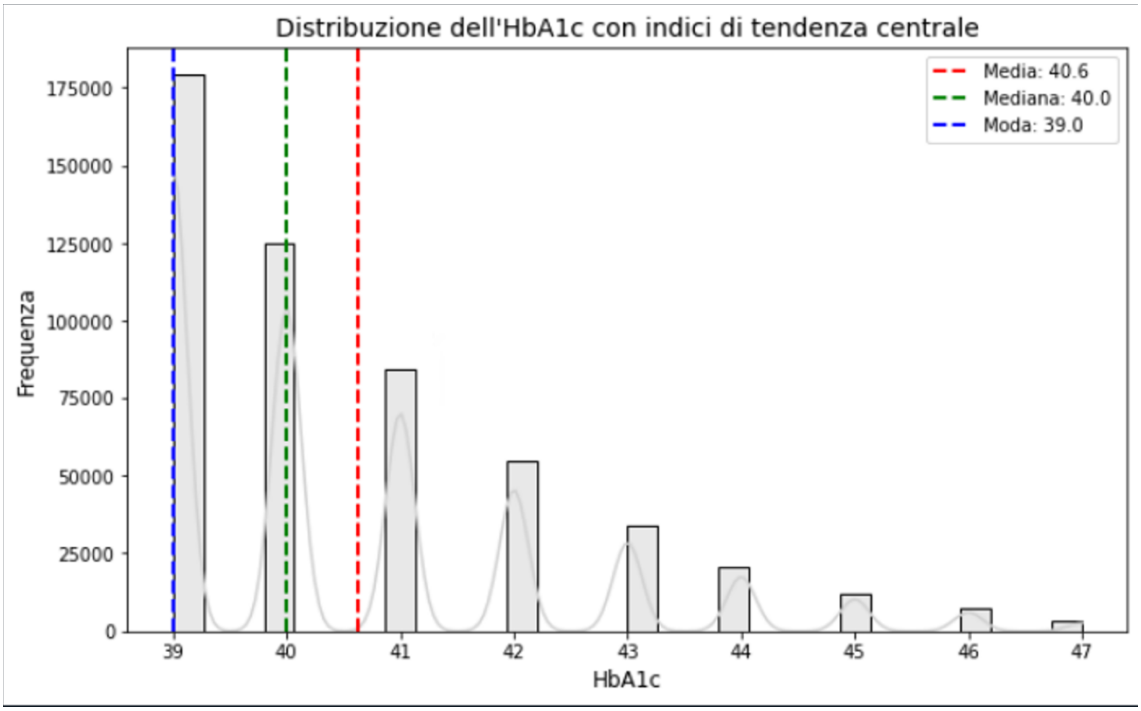

Figura 12: Distribuzione delle frequenze dei valori di emoglobina glicata (HbA1c) nella coorte di studio, con indici di tendenza centrale. Il grafico mostra una forte asimmetria positiva, con una concentrazione dei pazienti nei valori minimi dell'intervallo prediabetico.

Al fine di valutare le relazioni lineari tra le variabili cliniche e la variabile target, ho generato una matrice di correlazione illustrata in *Figura 13*:

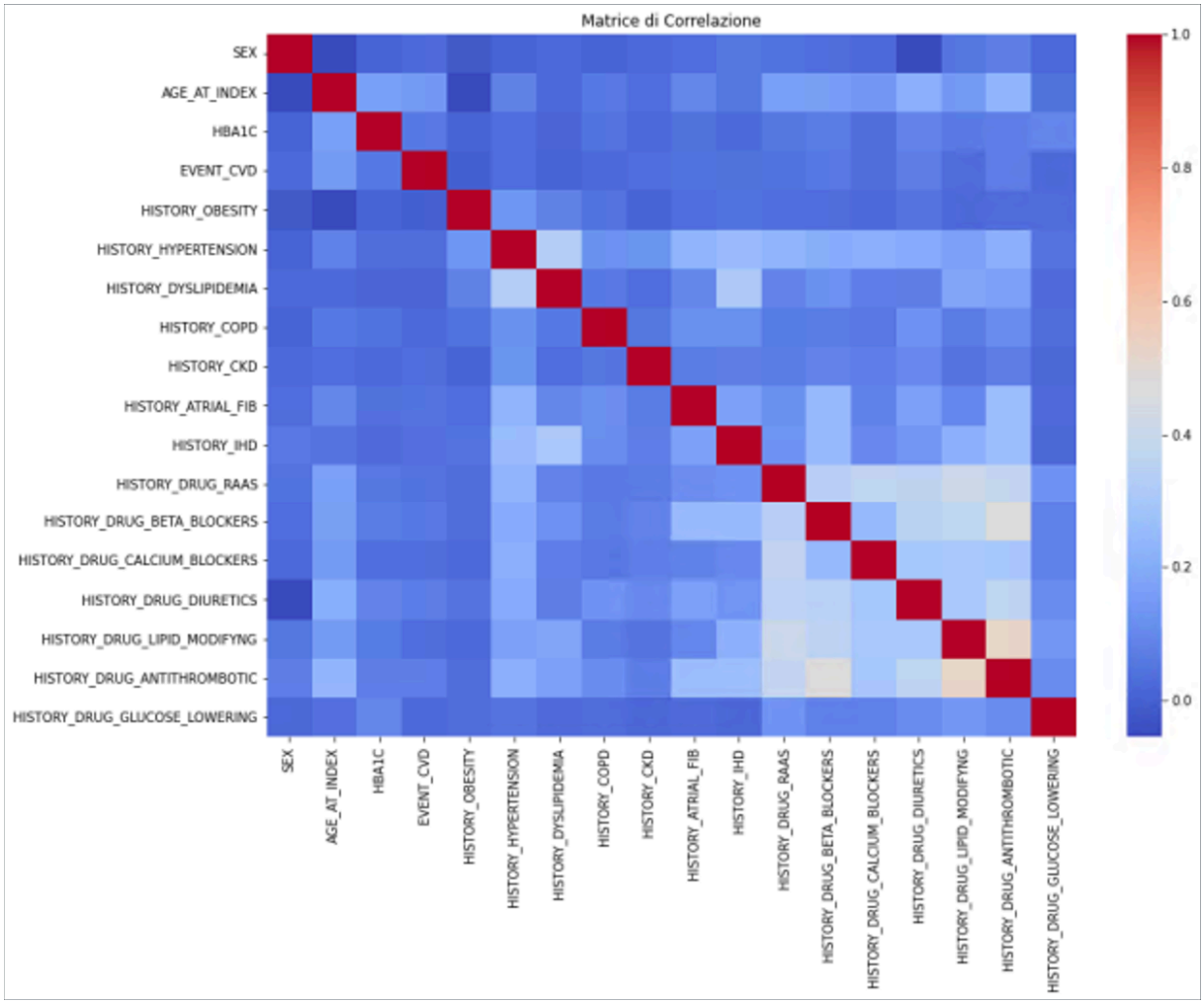

Figura 13: La matrice di correlazione è stata calcolata sulle variabili in ingresso, ed è utile per analizzare le relazioni lineari tra i fattori anagrafici, clinici e farmacologici.

Sulla destra è rappresentata la legenda in cui il colore rosso indica una correlazione positiva forte, mentre il colore blu indica un'assenza di correlazione. Dalla heatmap si evincono delle considerazioni interessanti:

-Bassa multicollinearità: il primo dato evidente è la forte assenza di correlazione, indicando che la maggior parte delle variabili non è correlata.
Dal punto di vista dell'analisi questo è un ottimo risultato in quanto le features in esame portano informazioni uniche e indipendenti.

-Coerenza tra patologie e terapie: le aree più chiare indicano delle leggere correlazioni positive. Entrando più nello specifico, si osserva una correlazione tra le diagnosi e i relativi farmaci (blu / celeste), ad esempio, si nota una correlazione tra “HISTORY_DYSLIPIDEMIA” e “HISTORY_DRUG_LIPID_MODYFING” (statine), tra “HISTORY_HYPERTENSION” e i farmaci antipertensivi, quali “DRUG_RAAS e DRUG_BETA_BLOCKERS”, e tra
“HISTORY_ATRIAL_FIB” e “HISTORY_DRUG_ANTITHROMBOTIC”. Si sottolinea anche una lieve correlazione positiva tra la variabile età e la maggior parte delle comorbilità e dei farmaci, confermando la teoria epidemiologica secondo cui con l’avanzare dell’età si rischia di avere un accumulo di patologie e terapie.
-Relazione con la variabile target: osservando la colonna (o la riga) relativa alla variabile target (EVENT_CVD), è evidente la quasi totale assenza di correlazione, quindi nessuna variabile presenta una forte correlazione lineare con l’evento cardiovascolare.
Ciò significa che non si può cercare di prevedere il rischio cardiovascolare con una semplice analisi univariata o con un biomarcatore clinico.

L’assenza di correlazioni dirette con la variabile target giustifica l’utilizzo di algoritmi di Machine Learning, che si basano su pattern e regole decisionali complesse.

3.2 Analisi dello sbilanciamento

L’analisi dello sbilanciamento tra le classi nella variabile target (EVENT_CVD) è stata condotta solo ed esclusivamente sul training set perché nel test set deve essere mantenuta inalterata la distribuzione a priori dei dati e dunque se il dataset originale è sbilanciato, il test set deve riflettere questa realtà per vedere se i modelli riescono a gestire o meno la classe minoritaria.

Addestrare modelli di Machine Learning su questa distribuzione dei dati porterebbe gli algoritmi a ignorare i casi patologici, classificando tutti i pazienti come sani, rendendo i modelli inutili ai fini diagnostici. Per ovviare a questa problematica ho applicato la tecnica di Undersampling che ha permesso di ottenere così un training set perfettamente bilanciato.

Il metodo dell'Undersampling causa però una perdita eccessiva di campioni. A tal proposito, per risolvere la perdita di informazione, nelle fasi successive allo studio ho optato per delle tecniche di Ensamble bilanciate, in grado di ripetere il processo di campionamento su più sotto-modelli. Nel caso specifico ho impiegato una tecnica nota come Balanced Bagging, che nello specifico caso ha i seguenti passi:

-Numero di soggetti con evento (classe minoritaria).

-Numeri di soggetti senza evento (classe maggioritaria).

Sono stati costruiti 20 modelli indipendenti e per ciascun modello:

1. Sono stati inclusi tutti i soggetti con evento CVD.

2. È stato selezionato un sottoinsieme casuale di soggetti senza evento con uguale numerosità.
3. Il modello è stato addestrato su questo dataset bilanciato (50% casi, 50% controlli).

Tale procedura è stata ripetuta 20 volte, generando 20 classificatori distinti, ciascuno addestrato su un diverso sottoinsieme della classe maggioritaria. Le predizioni finali sono state ottenute combinando i modelli mediante un voto di maggioranza.

Rispetto al Random Undersampling, il Balanced Bagging presenta diversi vantaggi:

- Non elimina definitivamente osservazioni della classe di maggioranza.
- Viene ridotta la varianza del modello grazie all'aggregazione.
- Ha maggiore capacità di generalizzare.

3.3 Performance dei modelli

In *Tabella 4* vengono riassunte le metriche di valutazione ottenute dai quattro modelli di Machine Learning addestrati.

Tabella 4: Performance dei modelli

Modello	**AUC**	**Accuratezza**	**Recall**	**Precision**	**F1-Score**	**MCC**
Reg. Log.	0.783	0.701	0.737	0.064	0.117	0.153
Ran. For.	0.785	0.706	0.732	0.064	0.118	0.153
Gr. Boos.	0.786	0.704	0.737	0.064	0.118	0.154
SVM	0.774	0.680	0.756	0.061	0.113	0.149

Comparando i risultati emergono le seguenti caratteristiche:

-Capacità discriminativa: l'area sotto la curva (AUC) rappresenta la metrica più robusta per valutare la bontà dei modelli. Tutti i modelli mostrano delle performance solide e circa equivalenti, con valori compresi tra 0.77 e 0.79.
In particolare, il modello Gradient Boosting si distingue per la performance superiore (AUC = 0.786), confermando l'efficacia degli algoritmi ensemble nel catturare relazioni complesse.

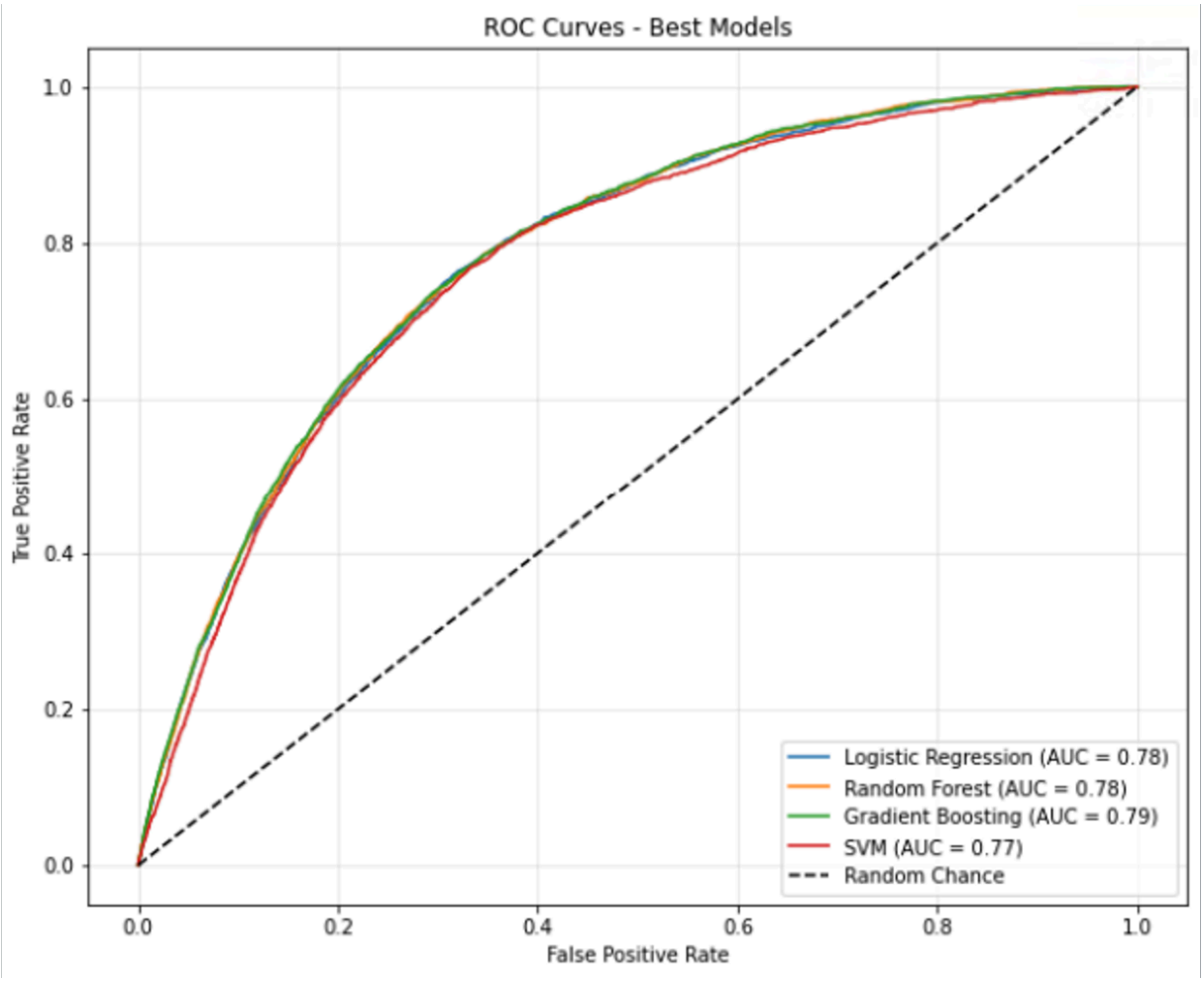

Figura 14: Curve ROC e valori di AUC per i quattro modelli di classificazione, confrontati con la classificazione casuale.

In *Figura 14* vengono messe in evidenza le curve "Receiver Operating Characteristic" (ROC) [28] prodotte dai quattro modelli utilizzati.

Quello che emerge è l'evidente distacco di tutte le curve dalla linea tratteggiata che rappresenta il "Random Chance (AUC = 0.5)", ovvero le performance di un modello privo di intelligenza che assegna le probabilità 50/50.
Dal momento in cui le quattro curve hanno superato ampiamente questa linea, l'intero set di modelli ha estrapolato positivamente un reale segnale clinico dai dati di addestramento.

Nonostante il marcato sbilanciamento delle classi dovuto alla bassa prevalenza della patologia, le forme assunte dalle curve ROC dimostrano che le probabilità di rischio cardiovascolare non sono casuali in quanto i modelli sono in grado di classificare i soggetti con evento da quelli senza evento.

-Recall (sensibilità) vs Precision: tutti i modelli mostrano una buona sensibilità, compresa tra il 73% e il 75%, mentre tutti e quattro presentano una bassa Precision che si aggira intorno al 6%.

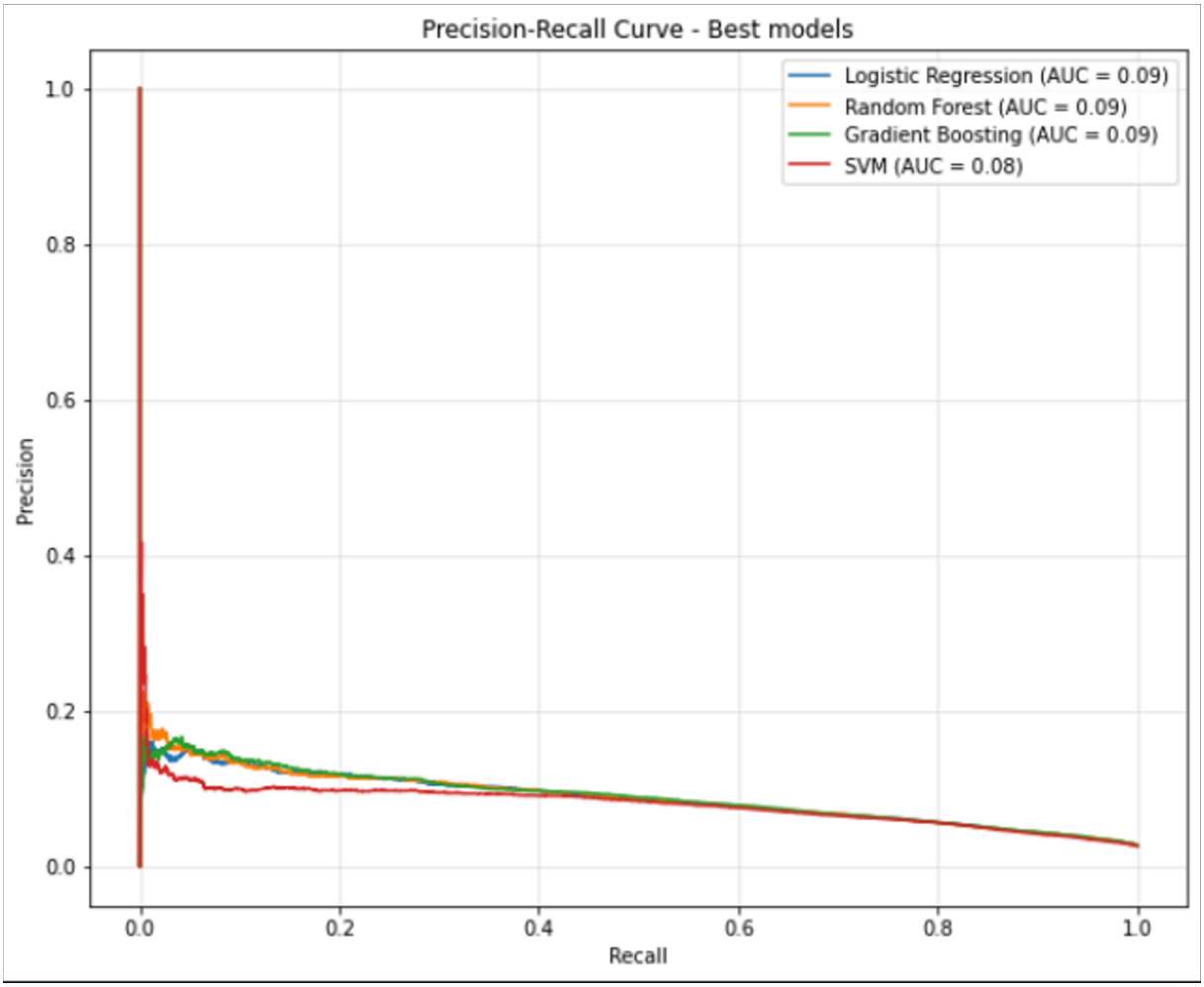

Figura 15: Curve Precision-Recall[29] relative ai modelli predittivi addestrati.

Dall'analisi in *Figura 15* si può notare che tutte le curve mostrano un picco iniziale di alta precisione, seguito da un crollo molto importante che successivamente si

assesta quando la Recall arriva a circa 0.15 – 0.20. In quel punto la Precision è crollata stabilizzandosi su un valore di circa 0.10.
Tale crollo è la conseguenza dello sbilanciamento del dataset, perché i modelli per poter aumentare la Recall sono costretti ad abbassare la soglia decisionale, includono una grande quantità di falsi positivi che abbattono la Precision.

Questo fenomeno può essere attenzionato numericamente mediante le seguenti matrici di confusione:

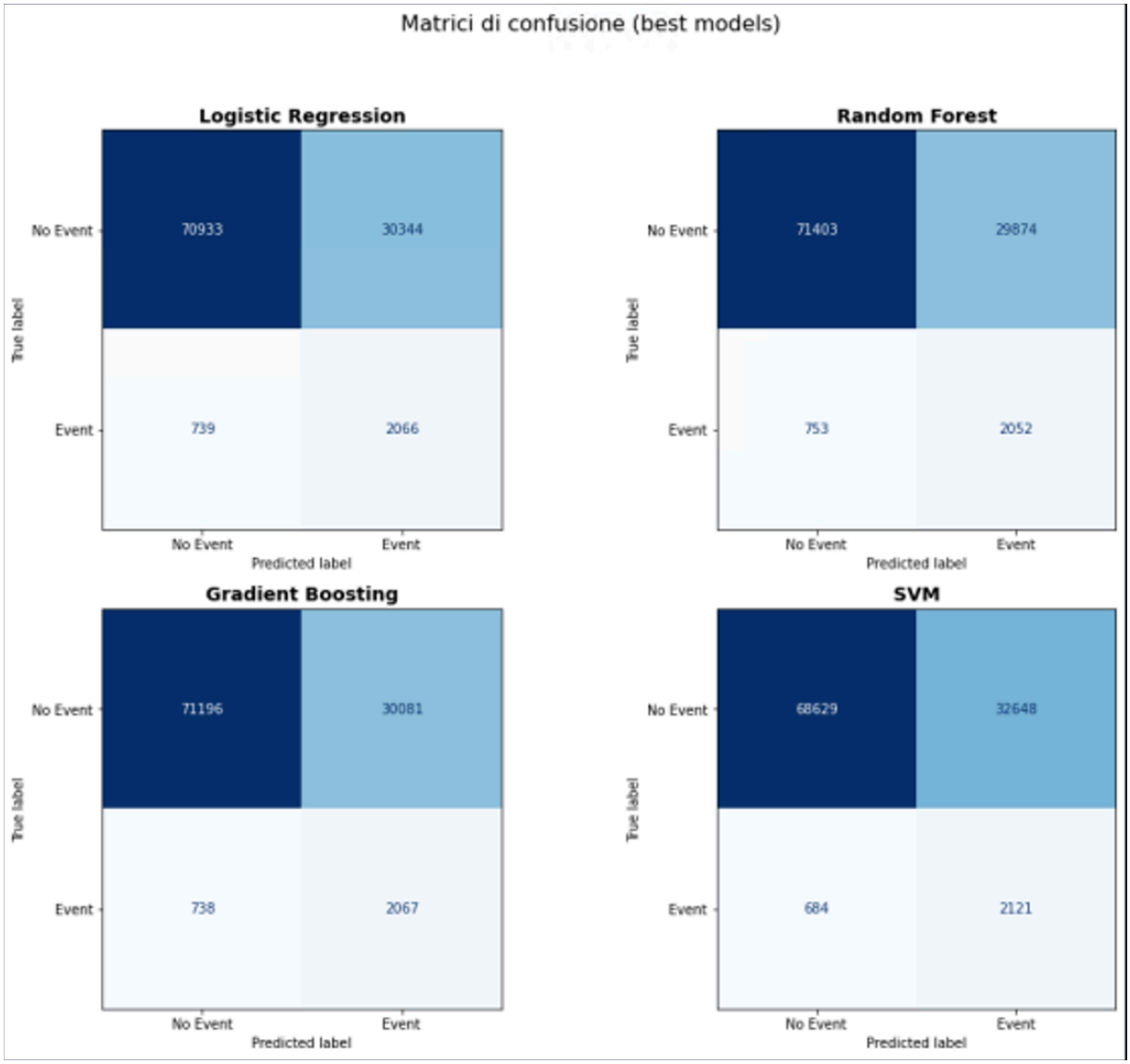

Figura 16: Matrici di confusione relative ai quattro modelli predittivi valutati sul test set. I grafici mettono in relazione le predizioni degli algoritmi (asse delle ascisse) con il reale stato clinico del paziente (asse delle ordinate). Questa

rappresentazione permette di quantificare il numero esatto di veri positivi (TP), veri negativi (TN), falsi positivi (FP) e falsi negativi (FN).

I valori delle quattro matrici spiegano in modo concreto il crollo della metrica Precision.

Dunque, una buona Recall associata a una bassa Precision si traduce in un modello altamente sensibile ma poco specifico. Quindi l'algoritmo identifica con successo i casi patologici, accettando però il compromesso di classificare erroneamente come malati numerosi pazienti sani, generando un grande numero di falsi positivi.

In ambito medico, questo risultato può essere accettabile se in presenza di uno strumento a valle dell'impiego del calcolatore del rischio che consenta uno screening a costo relativamente basso: in questo caso è meglio avere un falso positivo da verificare con ulteriori esami, piuttosto che mancare la diagnosi di un paziente a rischio (falso negativo).

-Affidabilità delle probabilità predette: si osserva in *Figura 17* che tutte le curve di tutti e quattro i modelli stanno nettamente e costantemente al di sotto della linea di perfetta calibrazione.
Questo significa che i modelli sono "overconfident", quindi la probabilità predetta dai modelli non rispecchia la probabilità reale.
Si tratta di una conseguenza del bilanciamento del dataset, che spinge il modello ad essere più prudente e allarmista. In ambito clinico è preferibile avere un sistema che massimizzi il rischio piuttosto che un sistema che rischia di sottovalutare i soggetti realmente a rischio.

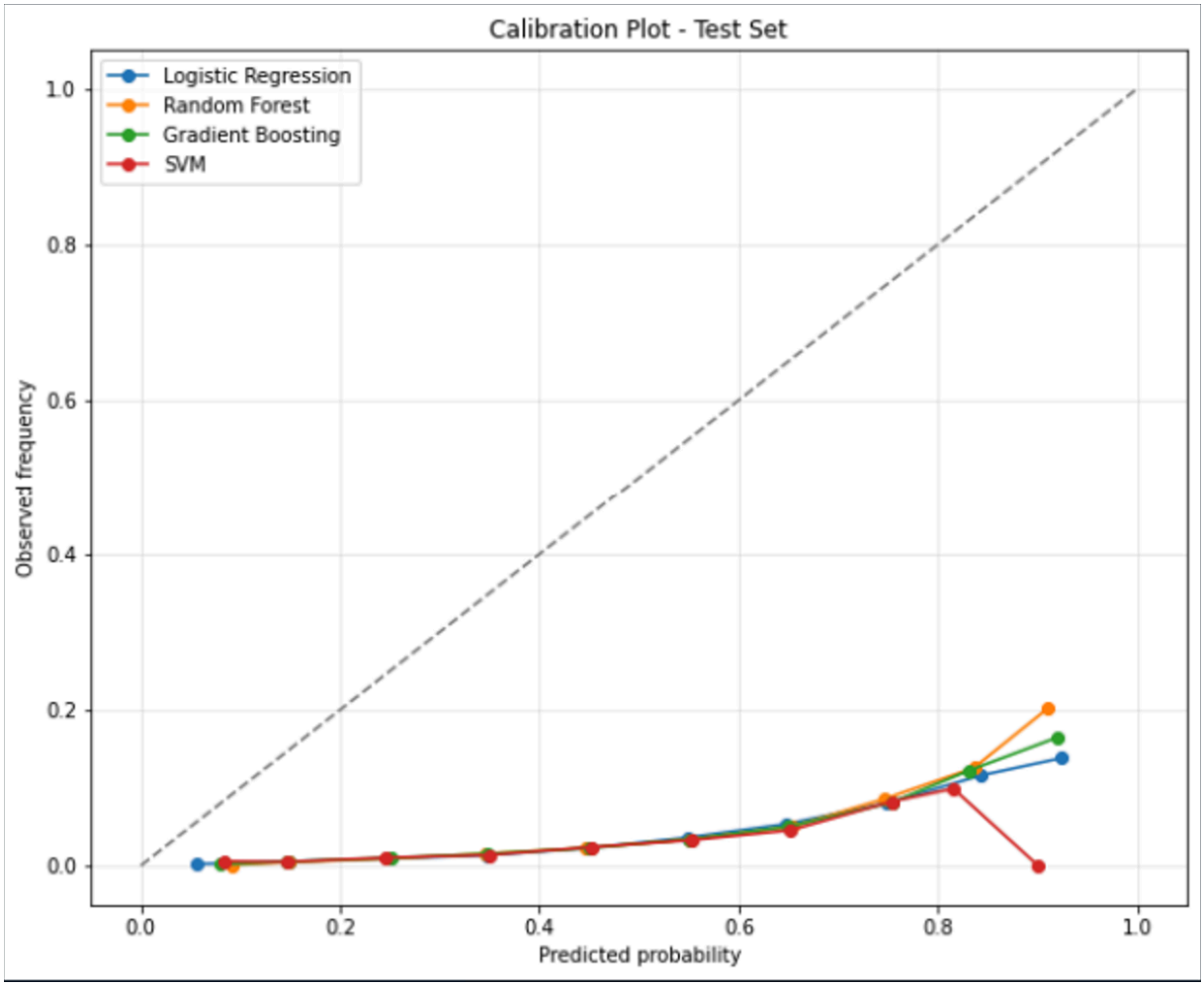

Figura 17: Calibration plot calcolato sul test set per i quattro modelli predittivi.

4. DISCUSSIONE E CONCLUSIONI

In conclusione, con il presente studio, è stato analizzato il rischio cardiovascolare in una coorte di pazienti con prediabete, utilizzando esclusivamente dati amministrativi e clinici di routine contenuti nei registri nazionali danesi.

Mediante un processo definito, che ha compreso la costruzione del dataset, una fase di preprocessing e l'addestramento di diversi modelli di Machine Learning, è stata dimostrata l'efficienza dello studio nel generare delle predizioni accurate, confermate da importanti metriche nell'ambito di screening preventivo, come AUC e Recall.

Come dimostrato dalle analisi dell'importanza delle features, *Figura 18*, è emerso che l'età è il principale fattore di rischio. Tale risultato risulta coerente con le evidenze riportate in letteratura medica, dalle quali emerge che l'invecchiamento costituisce un fattore di rischio non modificabile, in grado di incidere significativamente sulle malattie cardiovascolari. [30]

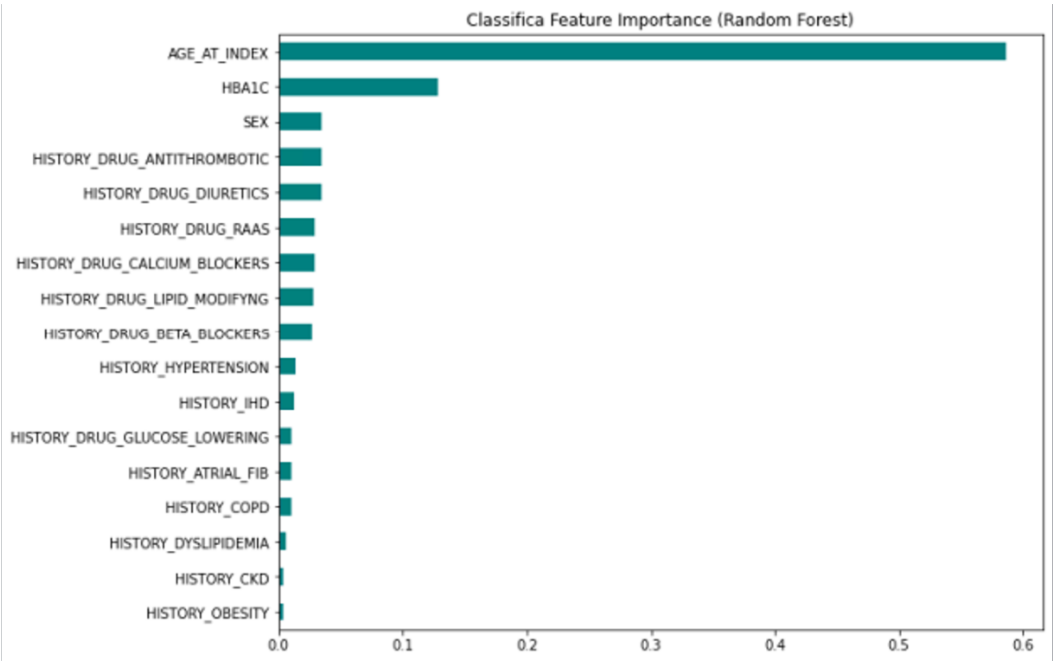

Figura 18: Analisi dell'importanza delle features tramite l'algoritmo Random Forest

Tra i fattori di rischio individuati, oltre al già menzionato fattore dell'età, è stata individuata un'altra feature rilevante: l'emoglobina glicata (HbA1c). Tale indice, che riflette il livello medio di glucosio nel sangue nel lungo periodo, si è rivelato particolarmente rilevante, in quanto dimostra l'importanza dei dati biochimici di laboratorio nell'analisi del rischio cardiovascolare.

Procedendo con l'analisi delle features, oltre ai fattori demografici e biochimici, si rinvengono quelle legate ai farmaci. Tali farmaci costituiscono degli indicatori indiretti di condizioni quali l'ipertensione o problemi di coagulazione del sangue, che approfondiscono la storia clinica dei pazienti.
L'integrazione di questi dati ha permesso ai modelli di apprendere pattern più complessi, migliorando la predizione sugli eventi cardiovascolari.

Uno dei principali punti di forza di tale studio è l'utilizzo di dati provenienti dai registri sanitari nazionali danesi. A differenza dei trial clinici, tali dati mettono in evidenza un'analisi basata sulla pratica clinica quotidiana e sul reale percorso dei pazienti.
La struttura di questi registri consente un tracciamento completo, capace di seguire l'intero iter dei pazienti dalle diagnosi ospedaliere alla prescrizione dei farmaci fino agli eventi di decesso, garantendo una continuità informativa completa nell'arco temporale analizzato.

L'utilizzo di questi registri è stato fondamentale per analizzare il rischio cardiovascolare su una coorte di pazienti con prediabete sul territorio nazionale danese. Si evidenzia, tuttavia, che il processo di Data Engineering è stato un valore aggiunto, necessario per la trasformazione dei dati grezzi e complessi in dati puliti, utili per l'addestramento dei modelli di Machine Learning. [31]

Va ulteriormente sottolineato che una prima limitazione dello studio riguarda l'arco temporale analizzato, che attiene al periodo compreso tra l'anno 2010 e il

2015. Nonostante in tale arco temporale si registrino dati consistenti per numero di coorte, si rileva l'opportunità di addestrare i modelli con dati recenti, considerato che, nel corso del tempo, le terapie e le linee guida si evolvono e l'utilizzo di dati moderni consentirebbe di ottenere delle stime più precise. [32]

Un'ulteriore limitazione dello studio svolto riguarda l'assenza, all'intero dei registri analizzati, di variabili fondamentali per garantire risultati affidabili, quali quelle legate allo stile di vita dei pazienti, come il fumo, l'attività fisica e l'alimentazione. Tali fattori hanno un impatto importante sul rischio cardiovascolare, ma non sono presenti nei database ospedalieri o farmacologici.
Va rilevata, altresì, l'assenza di altri tipi di dati, come l'indice di massa corporea (BMI). [33]

In conclusione, questo studio ha dimostrato che le tecniche di Machine Learning, applicate ai registri sanitari, rappresentano un potente strumento di screening di primo livello. Ed invero, tale fase iniziale dello studio, basata sui dati provenienti dai registri sanitari, permette di monitorare l'intera popolazione in modo rapido, consentendo la facile individuazione di pazienti che presentano una condizione di prediabete con un rischio cardiovascolare elevato.
Tale studio, infatti, può essere considerato il punto di partenza per analisi cliniche più accurate. Gli sviluppi futuri potrebbero prevedere l'integrazione di tali risultati con veri e propri trial clinici sui pazienti, così arricchendo il dataset con le variabili prima citate, al fine di migliorare e validare le predizioni dei modelli.
Infine, guardando alle prospettive future, sarebbe auspicabile l'utilizzo di tecniche di Deep Learning, quali l'implementazione di reti neurali, capaci di analizzare i dati in modo più accurato e preciso e di individuare pattern più complessi.

5. BIBLIOGRAFIA

1. Psoter KJ, Rosenfeld M. Opportunities and pitfalls of registry data for clinical research. *Paediatr Respir Rev*. 2013;14(3):141-145. doi:10.1016/j.prrv.2013.04.004

2. Kim KJ, Tagkopoulos I. Application of machine learning in rheumatic disease research. *Korean J Intern Med*. 2019;34(4):708-722. doi:10.3904/kjim.2018.349

3. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin Epidemiol*. 2015;7:449-490. doi:10.2147/CLEP.S91125

4. Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. *Medicina (Mex)*. 2020;56(9):455. doi:10.3390/medicina56090455

5. Deo RC. Machine Learning in Medicine. *Circulation*. 2015;132(20):1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593

6. Schallmoser S, Zueger T, Kraus M, Saar-Tsechansky M, Stettler C, Feuerriegel S. Machine Learning for Predicting Micro- and Macrovascular Complications in Individuals With Prediabetes or Diabetes: Retrospective Cohort Study. *J Med Internet Res*. 2023;25:e42181. doi:10.2196/42181

7. Kim MS, Jo DS, Lee DY. Comparison of HbA1c and OGTT for the diagnosis of type 2 diabetes in children at risk of diabetes. *Pediatr Neonatol*. 2019;60(4):428-434. doi:10.1016/j.pedneo.2018.11.002

8. Rao SS. Impaired Glucose Tolerance and Impaired Fasting Glucose.

9. 2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2024. *Diabetes Care*. 2023;47(Suppl 1):S20-S42. doi:10.2337/dc24-S002

10. Rooney MR, Fang M, Ogurtsova K, et al. Global Prevalence of Prediabetes. *Diabetes Care*. 2023;46(7):1388-1394. doi:10.2337/dc22-2376

11. Cai X, Zhang Y, Li M, et al. Association between prediabetes and risk of all cause mortality and cardiovascular disease: updated meta-analysis. *The BMJ*. 2020;370:m2297. doi:10.1136/bmj.m2297

12. Tabák AG, Herder C, Rathmann W, Brunner EJ, Kivimäki M. Prediabetes: A high-risk state for developing diabetes. *Lancet*. 2012;379(9833):2279-2290. doi:10.1016/S0140-6736(12)60283-9

13. Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC. Homeostasis model assessment: insulin resistance and β-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia*. 1985;28(7):412-419. doi:10.1007/BF00280883

14. Vigersky RA, McMahon C. The Relationship of Hemoglobin A1C to Time-in-Range in Patients with Diabetes. *Diabetes Technol Ther*. 2019;21(2):81-85. doi:10.1089/dia.2018.0310

15. Razavi LN, Ebenibo S, Edeoga C, Wan J, Dagogo-Jack S. Five-Year Glycemic Trajectories Among Healthy African-American and European-American Offspring of Parents With Type 2 Diabetes. *Am J Med Sci*. 2020;359(5):266-270. doi:10.1016/j.amjms.2020.03.005

16. Galicia-Garcia U, Benito-Vicente A, Jebari S, et al. Pathophysiology of Type 2 Diabetes Mellitus. *Int J Mol Sci*. 2020;21(17):6275. doi:10.3390/ijms21176275

17. Jørgensen ME, Ellervik C, Ekholm O, Johansen NB, Carstensen B. Estimates of prediabetes and undiagnosed type 2 diabetes in Denmark: The end

of an epidemic or a diagnostic artefact? *Scand J Public Health*. 2020;48(1):106-112. doi:10.1177/1403494818799606

18. Nicolaisen SK, Pedersen L, Witte DR, Sørensen HT, Thomsen RW. HbA1c-defined prediabetes and progression to type 2 diabetes in Denmark: A population-based study based on routine clinical care laboratory data. *Diabetes Res Clin Pract*. 2023;203:110829. doi:10.1016/j.diabres.2023.110829

19. Vazzana A, Giannatiempo V, Dei Cas A. Prediabete e suoi fenotipi. *L'Endocrinologo*. 2025;26(3):326-332. doi:10.1007/s40619-025-01623-y

20. Santhanalakshmi D, Gautam S, Gandhi A, Chaudhury D, Goswami B, Mondal S. Heart Rate Variability (HRV) in Prediabetics – A Cross Sectional Comparative Study in North India.

21. Peng RP, Zhu ZQ, Shen HY, et al. Retinal Nerve and Vascular Changes in Prediabetes. *Front Med*. 2022;9. doi:10.3389/fmed.2022.777646

22. Sundbøll J, Adelborg K, Munch T, et al. Positive predictive value of cardiovascular diagnoses in the Danish National Patient Registry: a validation study. *BMJ Open*. 2016;6(11):e012832. doi:10.1136/bmjopen-2016-012832

23. Malley B, Ramazzotti D, Wu JT yu. Data Pre-processing. In: *Secondary Analysis of Electronic Health Records*. Springer International Publishing; 2016:115-141. doi:10.1007/978-3-319-43742-2_12

24. Shalabi LA, Shaaban Z, Kasasbeh B. Data Mining: A Preprocessing Engine. *J Comput Sci*. 2006;2(9):735-739. doi:10.3844/jcssp.2006.735.739

25. Di Lena P, Sala C, Prodi A, Nardini C. Missing value estimation methods for DNA methylation data. Wren J, ed. *Bioinformatics*. 2019;35(19):3786-3793. doi:10.1093/bioinformatics/btz134

26. Omar ED, Mat H, Abd Karim AZ, et al. Comparative Analysis of Logistic Regression, Gradient Boosted Trees, SVM, and Random Forest Algorithms for Prediction of Acute Kidney Injury Requiring Dialysis After Cardiac Surgery. *Int J Nephrol Renov Dis*. 2024;Volume 17:197-204. doi:10.2147/IJNRD.S461028

27. Nembrini S, König IR, Wright MN. The revival of the Gini importance? Valencia A, ed. *Bioinformatics*. 2018;34(21):3711-3718. doi:10.1093/bioinformatics/bty373

28. Roumeliotis S, Schurgers J, Tsalikakis DG, et al. ROC curve analysis: a useful statistic multi-tool in the research of nephrology. *Int Urol Nephrol*. 2024;56(8):2651-2658. doi:10.1007/s11255-024-04022-8

29. Zhou QM, Zhe L, Brooke RJ, Hudson MM, Yuan Y. A relationship between the incremental values of area under the ROC curve and of area under the precision-recall curve. *Diagn Progn Res*. 2021;5:13. doi:10.1186/s41512-021-00102-w

30. North BJ, Sinclair DA. The Intersection Between Aging and Cardiovascular Disease. *Circ Res*. 2012;110(8):1097-1108. doi:10.1161/CIRCRESAHA.111.246876

31. Cui C, Chou SHS, Brattain L, Lehman CD, Samir AE. Data Engineering for Machine Learning in Women's Imaging and Beyond. *Am J Roentgenol*. 2019;213(1):216-226. doi:10.2214/AJR.18.20464

32. Hageman SHJ, Kaptoge S, De Vries TI, et al. Prediction of individual lifetime cardiovascular risk and potential treatment benefit: development and recalibration of the LIFE-CVD2 model to four European risk regions. *Eur J Prev Cardiol*. 2024;31(14):1690-1699. doi:10.1093/eurjpc/zwae174

33. Chen Y, Yu W, Lv J, et al. Early adulthood BMI and cardiovascular disease: a prospective cohort study from the China Kadoorie Biobank. *Lancet Public Health*. 2024;9(12):e1005-e1013. doi:10.1016/S2468-2667(24)00043-4