Bayesian network analysis for diabete's prediction

Napolitano Margherita Maria, Nappi Severino

ARTICLE INFO

Keywords:
Bayesian Network
Diabete
Prediction
BRFSS
Hill-Climbing
K2Score

Sommario

L'obiettivo di questo lavoro è prevedere l'insorgenza del diabete e diagnosticare la causa, informando così lo sviluppo di strategie efficaci di prevenzione e controllo. Sono stati sviluppati modelli di rete bayesiana (BN) per esplorare le specifiche relazioni tra i fattori influenzanti e il diabete.

È stata svolta un'analisi preliminare del dataset eseguendo feature extraction e selection attraverso l'uso di algoritmi di Random Forest e analisi di correlazione per identificare i predittori più rilevanti tra le numerose variabili disponibili.

Successivamente, è stata addestrata e testata una Rete Bayesiana (BN) ed è stata valutata la sua performance predittiva tramite il calcolo di alcune metriche quali accuratezza, precisione, recall e F1-score.

Il nostro modello ha dimostrato una grande capacità predittiva, ottenendo buoni punteggi su una serie di parametri, tra cui l'accuratezza (71%). I risultati possono essere utilizzati per adattare gli interventi medici, migliorare l'efficacia dei programmi di trattamento e implementare modifiche allo stile di vita per i pazienti che hanno un potenziale per sviluppare il diabete.

1. Introduction

Con il continuo miglioramento del livello economico, l'accelerazione dell'invecchiamento della popolazione e il cambiamento dello stile di vita delle persone, le malattie croniche sono diventate un problema di salute pubblica di grande rilievo. Tra queste, il diabete mellito (DM) e la cardiopatia coronarica (CAD) sono due malattie croniche comuni che rappresentano una seria minaccia per la salute umana. Il diabete è una malattia metabolica cronica grave, che costituisce la base patologica di una varietà di complicanze croniche, specialmente vascolari. Il diabete di tipo 2 (T2DM) è il più comune, interessando oltre il 90% di tutti i pazienti diabetici [2].

L'analisi predittiva delle malattie croniche, in particolare del diabete, richiede metodologie avanzate per affrontare la complessità e l'interazione tra molteplici fattori di rischio. Questo studio, partendo da un database del Behavioral Risk Factor Surveillance System (BRFSS) del 2022, utilizza un sistema di reti bayesiane per prevedere la probabilità che un individuo possa sviluppare il diabete.

In una prima fase, è stata effettuata una rigorosa selezione delle variabili (features) presenti all'interno del database attraverso l'uso di algoritmi di Random Forest e analisi di correlazione per identificare i predittori più rilevanti per i nostri scopi. La selezione delle features ha permesso di ridurre la dimensionalità del dataset, migliorando l'accuratezza e l'efficienza del modello predittivo.

Successivamente, sono state utilizzate le reti bayesiane per costruire un modello in grado di descrivere le relazioni dirette e indirette tra le variabili, semplificando il processo di ragionamento probabilistico. Le reti bayesiane, con la loro capacità di gestire sia dati completi che incompleti, offrono un vantaggio significativo rispetto ai metodi tradizionali, come la regressione logistica, e ad altri algoritmi di machine learning, che spesso richiedono indipendenza tra le variabili o presentano difficoltà nella quantificazione e interpretazione dei risultati.

L'approccio proposto permette non solo di prevedere se un individuo svilupperà il diabete, ma anche di identificare le relazioni causali tra i vari fattori di rischio, fornendo così preziose informazioni per lo sviluppo di strategie di prevenzione e controllo della malattia.

2. Dataset

Il dataset utilizzato all' interno di questa ricerca per effettuare analisi e predizioni è il risultato dell'indagine del Sistema di Sorveglianza dei Fattori di Rischio Comportamentali (BRFSS) [1]. Il BRFSS è una delle più grandi e longeve indagini di salute pubblica al mondo, gestita dai Centri per il Controllo e la Prevenzione delle Malattie (CDC) degli Stati Uniti. L'obiettivo primario del BRFSS è quello di raccogliere dati uniformi e specifici per ogni Stato sui comportamenti a rischio per la salute, le malattie e le condizioni croniche, l'accesso all'assistenza sanitaria e l'utilizzo dei servizi sanitari di prevenzione relativi alle principali cause di morte e disabilità negli Stati Uniti. Questi dati sono fondamentali per il monitoraggio delle condizioni di salute della popolazione e per l'elaborazione di politiche sanitarie mirate.

Il questionario BRFSS è composto da tre parti principali:

- Componente centrale: comprende domande standard poste da tutti gli Stati riguardanti lo stato di salute generale, l'accesso all'assistenza sanitaria, il consumo di alcol, l'uso di tabacco, i rischi di HIV/AIDS e le informazioni demografiche. Queste domande forniscono un quadro coerente delle condizioni di salute e dei comportamenti a rischio su scala nazionale.

- Moduli opzionali: trattano argomenti specifici come il prediabete, il diabete e la sopravvivenza al cancro, che ciascuno Stato può decidere di includere o meno nel questionario. Questi moduli permettono una maggiore flessibilità nella raccolta dei dati, consentendo di approfondire aspetti particolari della salute pubblica.

- Domande aggiunte dallo Stato: domande uniche sviluppate o acquisite dai singoli Stati per rispondere a esigenze specifiche. Queste domande riflettono le priorità sanitarie locali e permettono di affrontare problematiche particolari di ogni Stato.

Per la nostra analisi, abbiamo utilizzato i dati del BRFSS 2022, che comprendono 445.132 soggetti intervistati per un totale di 326 features per 54 Stati degli Stati Uniti. Questa vasta raccolta di dati ci ha permesso di avere una panoramica dettagliata delle condizioni di salute della popolazione e dei comportamenti a rischio, fornendo una solida base per le nostre analisi di selezione delle caratteristiche e predizione. Una panoramica più approfondita del database può essere trovata in https://www.cdc.gov/brfss/annual_data/2022/pdf/2022-calculated-variables-version4-508.pdf.

2.1. Analisi demografica

Effettuiamo un'analisi demografica del nostro dataset iniziale per capire la distribuzione dei soggetti che hanno partecipato alle interviste. In particolare, analizziamo la distribuzione di uomini e donne in base alla presenza o meno del diabete. I valori riportati nella tabella 1 rappresentano: Per la variabile _SEX:

- 1: maschio;
- 2: femmina.

Per la variabile DIABETE4:

- 1: si, ha il diabete;
- 2: si, ma è stato detto quando era incinta;
- 3: no, non ha il diabete;
- 4: no, pre-diabete o diabete borderline;
- 7: non si sa;
- 9: rifiutato.

Tale distribuzione può essere visualizzata nel grafico a barre impilate in figura 1. Possiamo facilmente affermare che hanno partecipato più donne allo studio, per un totale di 235.891 su 445.132 soggetti totali. In entrambi i casi, inoltre, abbiamo una maggioranza di soggetti che non presentano il diabete (DIABETE4 = 3).

Osserviamo ora la distribuzione del tipo di diabete, sempre in relazione al sesso, mediante la variabile DIABTYPE che presenta le seguenti etichette:

- 1: diabete di tipo 1;
- 2: diabete di tipo 2;
- 7: non si sa;
- 9: rifiutato;
- BLANK: non chiesto o manca.

Sesso	Diabete	Count
1	1	30.211
1	2	28
1	3	173.802
1	4	4.585
1	7	442
1	9	170
2	1	30.947
2	2	3.808
2	3	194.920
2	4	5.744
2	7	321
2	9	151

Tabella 1
Distribuzione del Diabete per Sesso

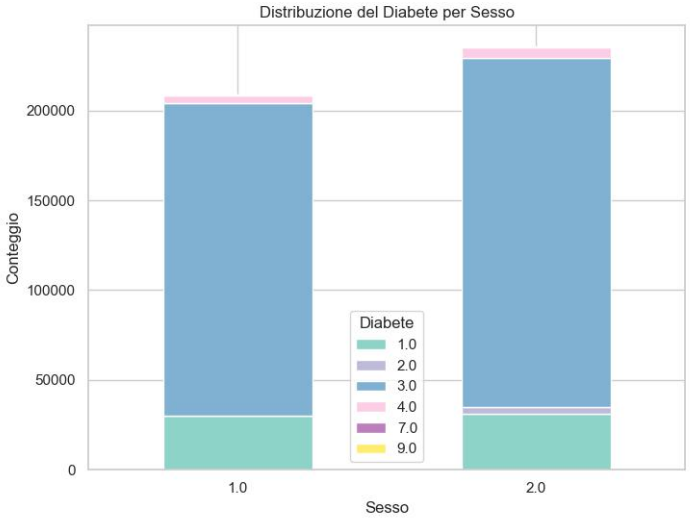

Figura 1: Distribuzione del diabete per sesso

Sesso	Tipo di Diabete	Count
1	1	554
1	2	5.031
1	7	506
1	9	12
2	1	496
2	2	5.528
2	7	460
2	9	13

Tabella 2
Distribuzione del tipo di Diabete per Sesso

Nel nostro dataset sono solo 12.600 i soggetti che presentano questa informazione e sono distribuiti come riportato nella tabella 2.

Possiamo dunque affermare che, nonostante il numero ristretto di soggetti a cui possiamo fare riferimento, è più diffuso il diabete di tipo 2 e lo possiamo visualizzare nella

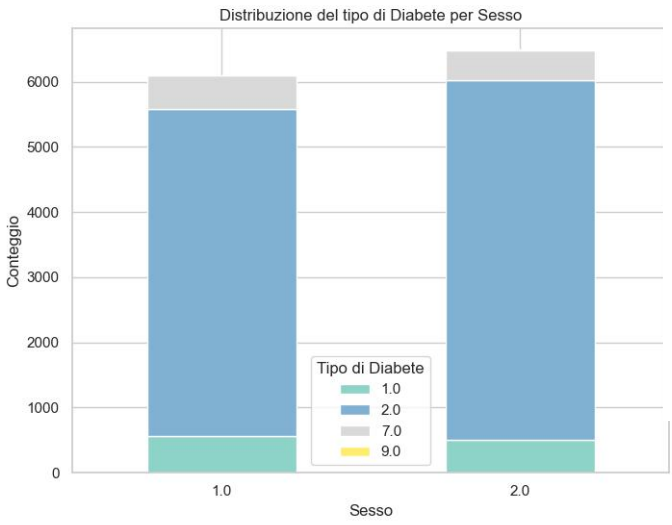

Figura 2: Distribuzione del tipo di diabete per sesso

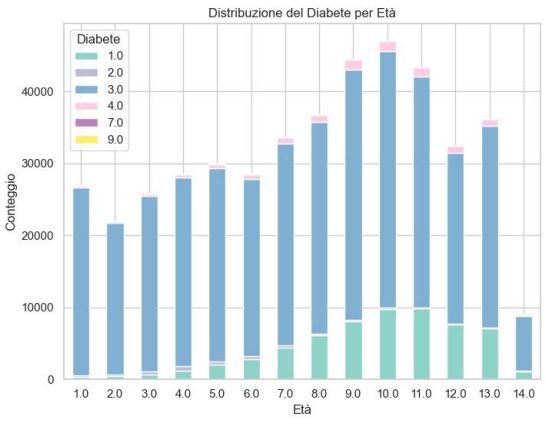

Figura 3: Distribuzione del diabete per età

figura 2. Tale tipo di diabete, infatti, è la forma più comune e rappresenta circa il 90% dei casi di questa malattia.

Vogliamo adesso analizzare la distribuzione della presenza di diabete in relazione all'età, riportata nella tabella 3. In particolare, utilizziamo la variabile _AGEG5YR in cui è riportata l'età dei soggetti in categorie di 5 anni:

- 1: età 18-24;
- 2: età 25-29;
- 3: età 30-34;
- 4: età 35-39;
- 5: età 40-44;
- 6: età 45-49;
- 7: età 50-54;
- 8: età 55-59;
- 9: età 60-64;
- 10: età 65-69;
- 11: età 70-74;
- 12: età 75-79;
- 13: età 80-99;
- 14: non si sa.

Nella figura 3 possiamo facilmente visualizzare che c'è una maggiore distribuzione dei soggetti nelle fasce di età compresa tra i 60 e i 74 anni (_AGEG5YR = 9, 10, 11), circa 130.000. In questo intervallo di età, inoltre, c'è anche una maggiore presenza di soggetti che hanno il diabete, circa 10.000 per ciascun gruppo (60-64, 65-69, 70-74).

3. Data preprocessing

Il dataset originale *'LLCP2022.XPT'* contiene 326 variabili. Per migliorare l'analisi, abbiamo deciso di eseguire diversi approcci di features selection. Quest'ultima, in particolare, consiste nel selezionare un sottoinsieme di dati riducendo così la dimensionalità del dataset originale.

3.1. Primo approccio

Inizialmente, abbiamo eliminato dal dataset le righe e le tre colonne contenenti esclusivamente valori NaN, poiché privi di contenuto rilevante, le colonne relative alla data di conduzione del sondaggio e tutti i record con valori NaN nella feature "DIABTYPE" 4, che indica il tipo di diabete (Tipo 1 o Tipo 2). Dopo tali modifiche, i soggetti presi in considerazioni per il nostro studio sono 12.600.

Facendo riferimento al paper *"RF-PCA: A New Solution for Rapid Identification of Breast Cancer Categorical Data Based on Attribute Selection and Feature Extraction"* [3], abbiamo deciso di utilizzare l'algoritmo Random Forest per la selezione delle variabili. Il paper, infatti, dimostra come l'uso della Random Forest, combinata con l'analisi delle componenti principali (PCA), possa migliorare significativamente l'accuratezza e la rapidità nella diagnosi del cancro al seno riducendo il numero di attributi e selezionando quelli più rilevanti. Pertanto, abbiamo adottato questa metodologia per ottimizzare il processo di selezione degli attributi. Tuttavia, nel nostro caso, non è stato possibile applicare la PCA (Principal Component Analysis) poiché la maggior parte delle variabili nel dataset sono categoriche e la PCA non è direttamente applicabile a esse.

Per tale motivo, abbiamo deciso di adottare **misure di correlazione multiple** in linea con il paper *"Feature Selection Based on Multiple Correlation Measures for Medical Examination Dataset"* [4]. Questo approccio è stato scelto per gestire efficacemente la ridondanza, l'irrilevanza e l'interazione tra le caratteristiche, garantendo un processo di selezione più robusto. Abbiamo analizzato la correlazione delle feature in relazione a "DIABTYPE" utilizzando l'*indice di correlazione di Pearson* e l'*indice di correlazione*

Età	Diabete	Count
1	1	367
1	2	146
1	3	26.140
1	4	201
1	7	82
1	9	5
2	1	385
2	2	233
2	3	21.121
2	4	207
2	7	43
2	9	1
3	1	661
3	2	448
3	3	24.320
3	4	331
3	7	43
3	9	4
4	1	1.222
4	2	562
4	3	26.264
4	4	426
4	7	51
4	9	1
5	1	1.992
5	2	503
5	3	26.859
5	4	524
5	7	55
5	9	9
6	1	2.816
6	2	366
6	3	24.635
6	4	659
6	7	45
6	9	10
7	1	4.401
7	2	355
7	3	27.954
7	4	867
7	7	58
7	9	9
8	1	6.051
8	2	281
8	3	29.376
8	4	1.053
8	7	47
8	9	13
9	1	8.023
9	2	242
9	3	34.802
9	4	1.376
9	7	62
9	9	15
10	1	9.708
10	2	206
10	3	35.719
10	4	1.376
10	7	68
10	9	22
11	1	9.821
11	2	173
11	3	32.109
11	4	1.288
11	7	64
11	9	17
12	1	7.610
12	2	116
12	3	23.788
12	4	948
12	7	49
12	9	7
13	1	7.013
13	2	141
13	3	28.061
13	4	955
13	7	64
13	9	17
14	1	1.088
14	2	64
14	3	7.574
14	4	127
14	7	32
14	9	191

Tabella 3
Distribuzione del Diabete per Età

di Spearman, considerando infine le feature risultanti comuni a entrambi gli approcci. Utilizzando *df.corr()* abbiamo calcolato la matrice di correlazione di Spearman e Pearson per tutte le colonne nel DataFrame. Specificando *['DIABTYPE'][:]* viene selezionata la colonna corrispondente a DIABTYPE dalla matrice di correlazione, che rappresenta la correlazione tra DIABTYPE e tutte le altre colonne del DataFrame. Questo permette di ottenere una serie di correlazioni tra DIABTYPE e tutte le altre variabili. Successivamente ordiniamo i valori di correlazione in ordine decrescente. L'obiettivo è quello di identificare rapidamente quali variabili sono più fortemente correlate (in senso positivo o negativo) con DIABTYPE. Viene creato un nuovo DataFrame che contiene solo le variabili che mostrano le 40 correlazioni più forti, sia positivamente che negativamente, con DIABTYPE. Calcoliamo la matrice di correlazione per le variabili selezionate nel DataFrame e mantieniamo solo le correlazioni comprese tra -0.1 e 0.1, cioè quelle considerate deboli o trascurabili. Una correlazione debole suggerisce che due variabili non hanno una forte relazione lineare, il che può essere utile per identificare variabili indipendenti tra loro. Infine, consideriamo solo le feature risultanti da entrambe le correlazioni. L'indice di Pearson misura la forza e la direzione di una relazione lineare tra due variabili continue. Presuppone che entrambe le variabili siano distribuite normalmente e che la relazione sia lineare. L'indice di Spearman è un

test non parametrico che misura la forza e la direzione di una relazione monotona, non necessariamente lineare tra due variabili ordinali o continue. Utilizzando entrambi i metodi, si può ottenere una visione più robusta della relazione tra le variabili. Se entrambi i coefficienti di correlazione mostrano una correlazione significativa, ciò rafforza l'affidabilità dei risultati. In questo modo, riusciamo a rappresentare tutto il dataset con un ristretto numero di features.

Da questa analisi abbiamo estratto 22 features:

- _AGEG5YR: Età riportata in categorie di 5 anni;
- BLDSTFIT: L'esame delle feci ematiche o del FIT è stato effettuato nell'ambito di un test Cologuard?;
- HPVADSHT: Quanti vaccini HPV ha ricevuto?;
- DIABAGE4: Hai il diabete?;
- CSRVINST: Queste istruzioni sono state scritte o stampate su carta?;
- PFPPRVN4: L'ultima volta che ha avuto un rapporto sessuale, lei o il suo partner avete fatto qualcosa per evitare una gravidanza?;
- NUMPHON4: Numero di telefono;
- DIABTYPE: Tipo di diabete;
- COVIDINT: Direbbe di aver già ricevuto tutte le dosi raccomandate, di avere in programma di ricevere tutte le dosi raccomandate o di non avere in programma di ricevere tutte le dosi raccomandate?;
- MARJOTHR: Ha fatto uso di marijuana o cannabis in altro modo?;
- DIABEDU1: Quando è stata l'ultima volta che ha frequentato un corso o una lezione su come gestire autonomamente il diabete?;
- HPVADVC4: Ha mai fatto una vaccinazione contro l'H.P.V.?;
- MARJEAT: Ha mangiato marijuana o cannabis?;
- CDHELP: Quando ha bisogno di aiuto per queste attività quotidiane, quanto spesso riesce a ottenere l'aiuto di cui ha bisogno?;
- CRGVLNG1: Per quanto tempo ha prestato assistenza a quella persona?;
- PSASUGST: Chi ha suggerito per primo questo test del PSA: lei, il suo medico o qualcun altro?;
- _PNEUMO3
- INSULIN1: Sta assumendo insulina?;
- ACEDRUGS: Ha vissuto con qualcuno che faceva uso di droghe illegali o che abusava di farmaci da prescrizione?;
- CNCRAGE: A che età è stata fatta la prima diagnosi di cancro?;
- NOBCUSE8: Quale è stata la ragione principale per cui non ha fatto nulla per prevenire la gravidanza l'ultima volta che ha avuto un rapporto sessuale?;
- CRGVHRS1: In una settimana media, quante ore fornisce cure o assistenza?.

Successivamente, abbiamo applicato la Random Forest al nostro dataset con 12.600 soggetti, di cui il 20% dei dati sarà utilizzato come set di test, mentre l'80% sarà utilizzato per l'addestramento del modello. Tale approccio è stato utilizzato per prevedere il tipo di diabete, esaminare l'importanza delle caratteristiche nel determinare il tipo di diabete e valutare la performance del modello mediante l'accuratezza. Per creare il classificatore Random Forest, come primo passo abbiamo scelto la variabile "DIABTYPE" come variabile target, ovvero, la colonna del DataFrame che contiene i valori che si vogliono prevedere. Successivamente, usiamo RandomForestClassifier per creare il classificatore Random Forest e settiamo i seguenti parametri:

- random_state=42: Garantisce la ripetibilità dei risultati.
- n_estimators=10: Usa 10 alberi nella foresta.
- max_depth=10: La profondità massima di ogni albero è 10.

Otteniamo un'accuratezza dell'86,51% per la predizione del tipo di diabete, e abbiamo estratto le seguenti feature più rilevanti:

- DIABAGE4: Hai il diabete?;
- CHKHEMO3: Quante volte negli ultimi 12 mesi un medico, un'infermiera o un altro operatore sanitario le ha fatto un controllo per l'A-one-C?;
- INSULIN1: Sta assumendo insulina?;
- _AGE80: Il valore dell'età imputata è crollato sopra gli 80 anni;
- WEIGHT2: Quanto pesa circa senza scarpe?;
- _BMI5: Indice di massa corporea (BMI);
- _AGE_G: Categoria di età imputata a sei livelli;
- WTKG3: Peso dichiarato in chilogrammi;
- _EDUCAG: Livello di istruzione completato;
- _AGEG5YR: Età dichiarata in categorie di età di cinque anni;
- INCOME3: Il suo reddito familiare annuale proviene da tutte le fonti?;
- HEIGHT3: Quanto sei alto senza scarpe?;

- HTM4: Altezza dichiarata in metri;
- DIABEDU1: Quando è stata l'ultima volta che ha frequentato un corso o una lezione su come gestire autonomamente il diabete?;
- _RFBMI5: Adulti che hanno un indice di massa corporea superiore a 25,00 (sovrappeso o obesi);
- _PSU: Unità di campionamento primaria (pari al numero di sequenza annuale);
- CHECKUP1: Da quanto tempo non si reca da un medico per un controllo di routine?;
- PHYSHLTH: Pensando alla sua salute fisica, che include malattie e lesioni fisiche, per quanti giorni negli ultimi 30 giorni la sua salute fisica non è stata buona?;
- SLEPTIM1: In media, quante ore di sonno dormite in un periodo di 24 ore?;
- _MRACE2: Categorizzazione della razza multirazziale calcolata;
- PNEUVAC4: Avete mai fatto un'iniezione di polmonite, nota anche come vaccino contro lo pneumococco?;
- MENTHLTH: Pensando alla vostra salute mentale, che comprende stress, depressione e problemi con le emozioni, per quanti giorni negli ultimi 30 giorni la vostra salute mentale non è stata buona?;
- EMTSUPRT: Quanto spesso ricevete il sostegno sociale ed emotivo di cui avete bisogno?;
- HTIN4: Altezza dichiarata in pollici;
- _INCOMG1: Categorie di reddito;
- POORHLTH: Negli ultimi 30 giorni, per quanti giorni circa la cattiva salute fisica o mentale le ha impedito di svolgere le sue attività abituali, come la cura di sé, il lavoro o lo svago?;
- DIABEYE1: Quando è stata l'ultima volta che un medico, un'infermiera o un altro operatore sanitario ha scattato una foto della parte posteriore dell'occhio con una macchina fotografica specializzata?;
- EDUCA: Livello di istruzione completato;
- COVIDSE1: In quale mese e anno ha ricevuto la seconda vaccinazione COVID-19?;
- EYEEXAM1: Quando è stata l'ultima volta che avete fatto una visita oculistica in cui le pupille sono state dilatate, rendendovi temporaneamente sensibili alla luce intensa?;
- RENTHOM1: Siete proprietari o affittuari della vostra casa?;
- FLSHTMY3: In quale mese e anno ha ricevuto il vaccino antinfluenzale più recente, spruzzato nel naso o iniettato nel braccio?;
- _BMI5CAT: Quattro categorie di Indice di Massa Corporea (IMC);
- _RACE1: Categorie di razza/etnia;
- EMPLOY1: Stato di occupazione.

La feature "DIABTYPE", su cui effettuiamo la nostra analisi, è descritta nella Figura 4.

Infine, abbiamo considerato solo le features risultanti dall'intersezione tra la correlazione e la Random Forest:

- AGEG5YR: Rappresenta le categorie di età in 14 intervalli di cinque anni a partire da 18 anni fino ad oltre 80, numerati da 1 a 14. La distribuzione di questa variabile ci aiuta a comprendere l'età dei partecipanti e a identificare eventuali gruppi di età che potrebbero essere più rappresentati nello studio.
- DIABEDU1: Indica se il partecipante ha mai frequentato un corso o una lezione su come gestire autonomamente il diabete. Le etichette forniscono quando è stato frequentato per l'ultima volta un corso:
  1. nell'ultimo anno;
  2. negli ultimi due anni;
  3. negli ultimi tre anni;
  4. negli ultimi cinque anni;
  5. negli ultimi dieci anni;
  6. più di dieci anni fa;
  7. non si sa;
  8. mai.

  La distribuzione di questa feature fornisce informazioni sull'educazione sanitaria ricevuta dai partecipanti.
- DIABAGE4: Rappresenta l'età alla diagnosi di diabete. Visualizzare questa distribuzione ci permette di osservare a quale età i partecipanti sono stati diagnosticati con il diabete, offrendo spunti su possibili fattori di rischio legati all'età.
- INSULIN1: Indica se il partecipante sta assumendo insulina. Analizzare la distribuzione di questa feature aiuta a comprendere la prevalenza dell'uso di insulina tra i partecipanti con diabete.

In figura 5 sono riportati i grafici delle distribuzioni delle feature in esame. Questi grafici rivelano importanti informazioni sulla popolazione in studio, ad esempio la distribuzione di _AGEG5YR mostra che la maggior parte dei partecipanti rientra nelle categorie di età più avanzata. Inoltre, facendo riferimento al paper *"DIABETIC MELLITUS PREDICTION WITH BRFSS DATASETS"* [5], abbiamo considerato le seguenti features:

- DIABETE3: Hai il diabete?;

Label: What type of diabetes do you have?
Section Name: Diabetes
Module Number: 2
Question Number: 1
Column: 271
Type of Variable: Num
SAS Variable Name: DIABTYPE
Question Prologue:
Question: According to your doctor or other health professional, what type of diabetes do you have?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Type 1	1,050	8.33	9.98
2	Type 2	10,559	83.80	81.48
7	Don't know/Not Sure	966	7.67	8.29
9	Refused	25	0.20	0.24
BLANK	Not asked or Missing Notes: Section 07.12, DIABETE4, is coded 2, 3, 4, 7, 9, or Missing	432,532	.	.

Figura 4: Descrizione variabile DIABTYPE

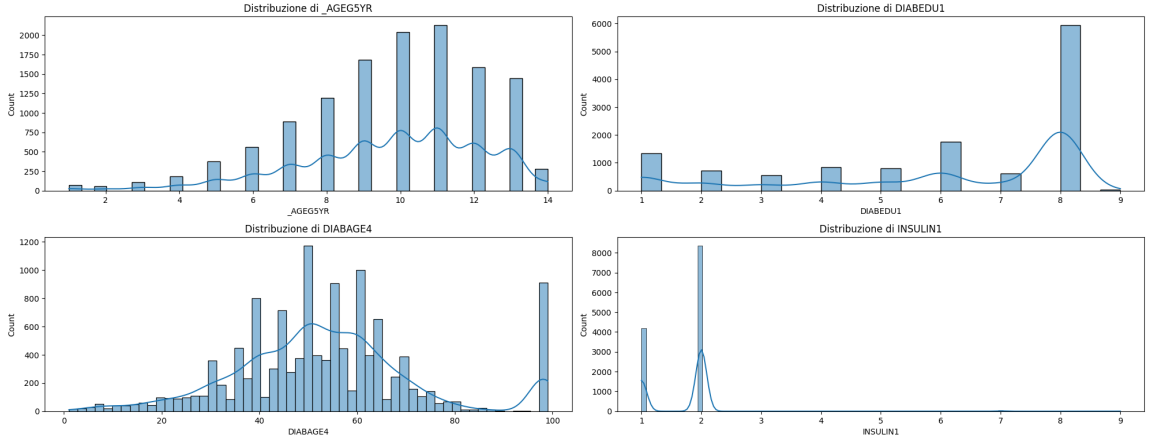

Figura 5: Istogrammi relativi alle features in esame

- _RFHYPE5: Adulti in buona o migliore salute;
- _BMI5: Indice di massa corporea (BMI);
- SMOKE100: Ha fumato almeno 100 sigarette in tutta la sua vita?;
- CVDSTRK3: Mai avuto una diagnosi di ictus;
- _MICHD: Intervistati che hanno dichiarato di aver avuto una malattia coronarica (CHD) o un infarto del miocardio (MI);
- _TOTINDA: Adulti che hanno riferito di aver svolto attività fisica o esercizio fisico negli ultimi 30 giorni al di fuori del loro lavoro abituale;
- MEDCOST: C'è stata una volta negli ultimi 12 mesi in cui ha avuto bisogno di un medico ma non ha potuto farlo perché non poteva permetterselo?;
- GENHLTH: Stato di salute generale;
- MENTHLTH: Pensando alla vostra salute mentale, che comprende stress, depressione e problemi con le emozioni, per quanti giorni negli ultimi 30 giorni la vostra salute mentale non è stata buona?;
- PHYSHLTH: Pensando alla sua salute fisica, che include malattie e lesioni fisiche, per quanti giorni negli ultimi 30 giorni la sua salute fisica non è stata buona?;
- DIFFWALK: Avete serie difficoltà a camminare o a salire le scale?;
- SEX: sesso;
- _AGEG5YR: Età dichiarata in categorie di età di cinque anni;
- EDUCA: Qual è il grado o l'anno di scuola più alto che ha completato?;

- INCOME2: È il vostro reddito familiare annuale da tutte le fonti;

come esaminato dagli autori del lavoro.
Successivamente vedremo che utilizzando questo dataset ridotto, ovvero formato dalle feature del paper di riferimento e dalle feature estratte dal nostro lavoro di preprocessing, le metriche risultanti dall'addestramento e dal test della rete bayesiana costruita non sono ottimi: Accuracy: 0.3937, Precision: 0.5, Recall: 0.1968, F1 Score: 0.2824. Abbiamo dunque deciso di lavorare nuovamente sul dataset iniziale ed eseguire una nuova analisi di feature selection.

3.2. Secondo approccio

Abbiamo analizzato nuovamente il dataset iniziale *'LLCP2022.XPT'* adottando un approccio diverso. In questo caso, non abbiamo applicato sul dataset modifiche preliminari quali eliminazione di valori NaN o di soggetti e variabili, ma abbiamo eseguito come primo passaggio la RandomForest con variabile target "DIABETE4" sull'intero dataset, ovvero completo di tutti i 445.132 soggetti che hanno partecipato allo studio e le 326 features, diviso sempre in set di train e set di test con la strategia 80/20. Abbiamo ottenuto un'accuratezza del 94.78% ed estratto le seguenti features:

- DIABAGE4: hai il diabete?;
- _BMI5: Indice di massa corporea (BMI);
- _AGEG5YR: Età dichiarata in categorie di età di cinque anni;
- WEIGHT2: Quanto pesa senza scarpe?;
- GENHLTH: salute generale;
- DIFFWALK: Avete serie difficoltà a camminare o a salire le scale?;
- INCOME3: reddito familiare annuo;
- _MICHD: Intervistati che hanno dichiarato di aver avuto una malattia coronarica (CHD) o un infarto del miocardio (MI);
- SLEPTIM1: In media, quante ore di sonno dormite in un periodo di 24 ore?;
- DIABEDU1: Quando è stata l'ultima volta che hai seguito un corso o una lezione su come gestire autonomamente il diabete?;
- DIABTYPE: Quale tipo di diabete hai?;
- INSULIN1: Stai prendendo insulina?;
- SMOKE100: Hai fumato almeno 100 sigarette in tutta la tua vita?;
- CVDSTRK3: Hai avuto un ictus?;
- _TOTINDA: Adulti che hanno dichiarato di aver svolto attività fisica o esercizio fisico negli ultimi 30 giorni, al di fuori del loro normale lavoro;
- MEDCOST1: C'è stato un momento negli ultimi 12 mesi in cui hai avuto bisogno di vedere un medico ma non hai potuto perché non te lo potevi permettere?;
- MENTHLTH: Pensando alla tua salute mentale, che comprende stress, depressione e problemi emotivi, per quanti giorni negli ultimi 30 giorni la tua salute mentale non è stata buona?;
- PHYSHLTH: Pensando alla tua salute fisica, che comprende malattie e infortuni, per quanti giorni negli ultimi 30 giorni la tua salute fisica non è stata buona?;
- LANDSEX1: Maschio o femmina?
- _AGEG5YR.1: Non specificata nel codeblock;
- _STSTR: Variabile di stratificazione del disegno del campione;
- _EDUCAG: Livello di educazione completo;
- DIABEYE1: Quando è stata l'ultima volta che un medico, un infermiere o un altro operatore sanitario ha scattato una foto della parte posteriore del tuo occhio con una macchina fotografica specializzata?;
- HEIGHT3: Altezza senza scarpe;
- COVIDINT: Diresti di aver già ricevuto tutte le dosi raccomandate, pensi di riceverle tutte o non pensi di riceverle tutte?;
- PFPPRVN4: L'ultima volta che hai avuto un rapporto sessuale, tu o il tuo partner avete fatto qualcosa per impedirti di rimanere incinta?;
- MARJEAT: Hai mangiato marijuana o cannabis?;
- HPVADSHT: Quante vaccinazioni contro l'HPV hai ricevuto?;
- BLDSTFIT: Il test del sangue nelle feci o il FIT sono stati eseguiti come parte di un test Cologuard?;
- CDHELP: Quando hai bisogno di aiuto per svolgere queste attività quotidiane, con quale frequenza riesci a ottenere l'aiuto di cui hai bisogno?;
- DIABETE4: Ti è mai stato detto di avere il diabete?;
- EDUCA: Qual è il voto o l'anno scolastico più alto che hai completato?;

Dopo aver selezionato le feature da analizzare, proseguiamo con l'analisi del dataset per ridurlo ulteriormente. In particolare, sostituiamo i valori 7, 9, 77, 99, 777, 999, 7777, 9999 e la stringa 'nan' con -1, in quanto considerati valori anomali come ad esempio risposte mancanti del tipo "non si sa" o " non è riportato". Successivamente, eliminiamo tutti questi valori anomali in quanto non ci danno informazioni rilevanti al fine della nostra analisi. Decidiamo poi di conteggiare i valori NaN per ciascuna colonna per

identificare le colonne con un numero significativo di valori mancanti, ovvero le colonne che presentano più di 120.000 valori NaN, che potrebbero influenzare la qualità dell'analisi o del modello predittivo ed eliminiamo le colonne, quindi le feature, che presentano più della metà di valori nulli. Infine, eliminiamo le righe, quindi i soggetti, in cui è presente almeno un valore nullo. È stata effettuata questa scelta per provare a ridurre il nostro dataset e addestrare al meglio la rete bayesiana che non può gestire direttamente i valori NaN. Alla fine di tale preprocessing, avremo un dataframe formato dal 57.495 righe (soggetti) e 16 colonne (features). In particolare sono state eliminate le seguenti feature: 'DIABEDU1', 'DIABAGE4', 'DIABTYPE', 'INSULIN1', 'SMOKE100','MEDCOST1', 'LANDSEX1', '_AGEG5YR.1', 'INCOME3', 'DIABEYE1','COVIDINT', 'PFPPRVN4', 'MARJEAT', 'HPVADSHT', 'BLDSTFIT', 'CDHELP','EDUCA'.

3.3. Terzo approccio

In un terzo approccio, abbiamo optato per affrontare i diversi tipi di diabete, tipo 1 e tipo 2, separatamente ma in modo parallelo. In primo luogo, a partire dal nostro database ripulito dai casi in cui non era stata registrata alcuna risposta alla domanda relativa alla feature **"DIABTYPE":** *"Secondo il suo medico o un altro professionista della salute, che tipo di diabete ha?"*, abbiamo suddiviso il database in due parti: una contenente 1.050 osservazioni relative al diabete di tipo 1 e l'altra con 10.559 osservazioni riguardanti il diabete di tipo 2. Inoltre, abbiamo selezionato un campione di soggetti 1.050 che hanno dichiarato di non essere affetti da diabete, utilizzando la feature **"DIABETE4":** *"Le hanno mai detto di avere il diabete?"*, come mostrato in figura 6. Poiché i tre dataset risultano significativamente sbilanciati in termini di numero di osservazioni disponibili, al fine di prevenire bias e distorsioni nei risultati, abbiamo posto in essere una procedura di campionamento stratificato. Quest'ultima ci ha permesso di equilibrare il numero di osservazioni tra i soggetti affetti da diabete di tipo 1 e quelli con diabete di tipo 2 ottenendo un nuovo dataset in cui la distribuzione delle variabili età, suddivisa in sei categorie, e sesso nei soggetti con diabete di tipo 2, che rappresenta il campione meno numeroso, sia comparabile a quella dei soggetti con diabete di tipo 1.

Il nostro approccio, pertanto, ha previsto una fase iniziale di analisi quantitativa e grafica della distribuzione combinata delle variabili **"SEXVAR":** *"Sesso dell'intervistato"* e **_AGE_G:** *"Categoria di età imputata a sei livelli"* per il gruppo di soggetti affetti da diabete di tipo 1, Fig.7.

Facendo lo stesso con i soggetti affetti da diabete di tipo 2, Fig.8.

Tale analisi ci ha permesso di comprendere sia visivamente che numericamente le differenze esistenti tra le due distribuzioni.

In seguito, abbiamo eseguito una procedura di sotto campionamento del gruppo con diabete di tipo 2, prelevando un campione che corrisponda a ciascuna combinazione di età e sesso del gruppo con diabete di tipo 1 così da rendere le due distribuzioni quanto più simili possibile. La procedura in sé prevede che se il gruppo di tipo 2 ha più soggetti di quelli necessari, selezioniamo casualmente un numero di individui uguale a quello presente nel gruppo tipo 1. Se invece il numero di soggetti è inferiore, includiamo tutti quelli disponibili. Una volta completato il sotto campionamento otteniamo 1033 soggetti affetti da diabete di tipo 2 la cui distribuzione della combinazione delle variabili sesso ed età si presenta nel seguente modo: Fig.9.

Questo processo ci permette di ottenere un nuovo dataset in cui i soggetti con diabete di tipo 2, originariamente 10.559, sono stati ridotti in modo da avere una distribuzione di età e sesso simile a quella del gruppo con diabete di tipo 1, che conta 1.050 soggetti. Grazie a questo bilanciamento, il dataset risultante può essere utilizzato per analisi o confronti tra i due gruppi, con la certezza che le variabili chiave disponibili dal punto di vista demografico, età e sesso, sono distribuite in modo simile.

In un'ulteriore fase del processo, abbiamo implementato una serie di procedure di data cleaning mirate a ottimizzare

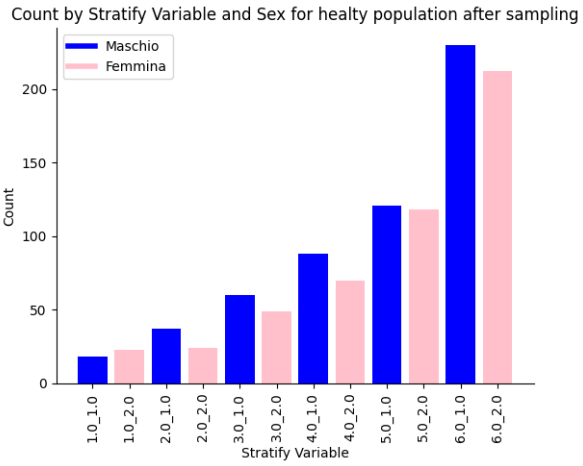

Figura 6: Distribuzione di soggetti sani per età e sesso.

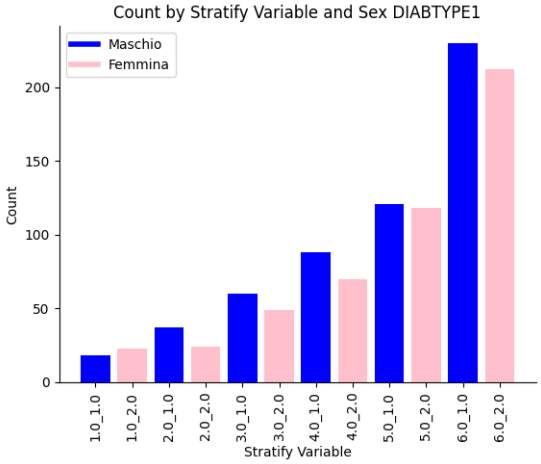

Figura 7: Distribuzione di intervistati affetti da diabete di tipo 1 per sesso ed età

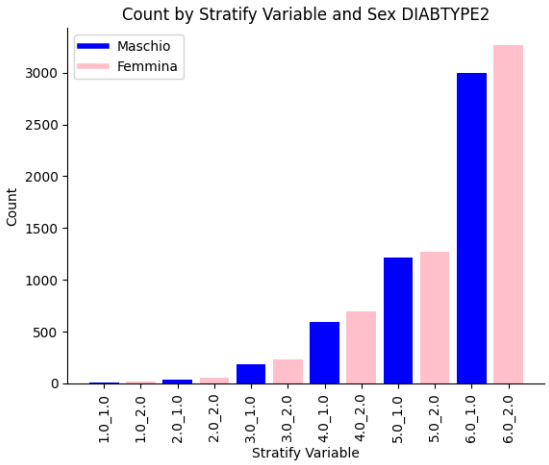

Figura 8: Distribuzione di intervistati affetti da diabete di tipo 2 per sesso ed età.

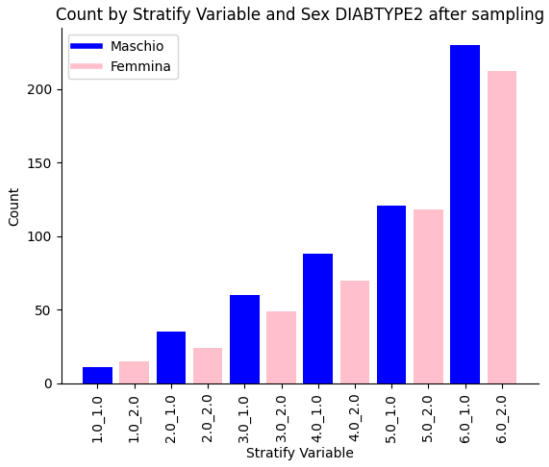

Figura 9: Distribuzione di intervistati affetti da diabete di tipo 2 per sesso ed età rispetto a quelli di tipo 1.

la fruibilità e la qualità del dataset. In particolare, ci siamo concentrati sulla rimozione di tutte quelle features che presentavano una percentuale di valori mancanti superiore al 70% rispetto al totale delle osservazioni, in quanto ritenute poco significative. Inoltre, abbiamo proceduto all'eliminazione di variabili che rappresentavano informazioni temporali considerate non pertinenti ai fini dell'analisi, al fine di ridurre la complessità del dataset e migliorarne la coerenza complessiva. Infine, al fine di individuare le features più rilevanti per i nostri obiettivi, abbiamo applicato due diversi metodi di correlazione su entrambi i dataset (diabete di tipo 1 e diabete di tipo 2). In primo luogo, abbiamo calcolato la correlazione di ciascuna variabile con la variabile target, rispettivamente *DIABTYPE1* e *DIABTYPE2*, utilizzando il metodo di Pearson per analizzare le relazioni lineari. Successivamente, abbiamo applicato il metodo di Spearman, che valuta le correlazioni monotone. Impostando una soglia di significatività con un p-value del 95%, abbiamo selezionato le variabili che risultavano significativamente correlate

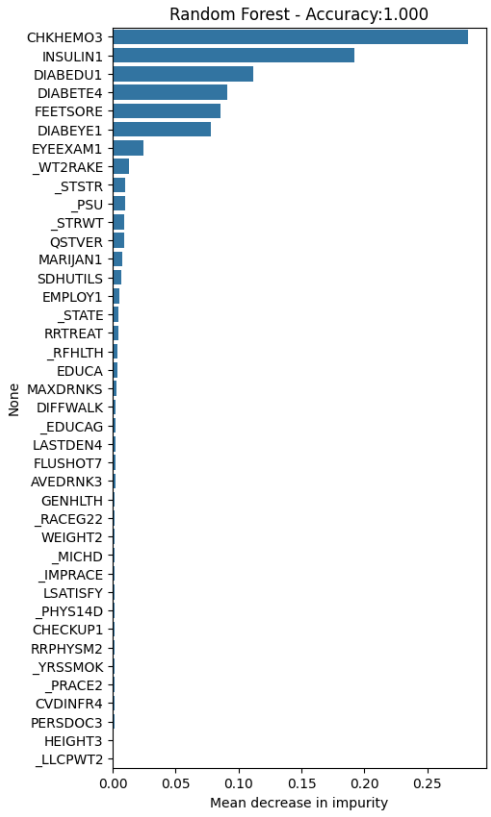

Figura 10: Random Forest diabete di tipo 1.

con le variabili target. In aggiunta, abbiamo eseguito una Random Forest per identificare le features più rilevanti in base all'importanza assegnata da questo algoritmo. Questa tecnica, mostrata nelle immagini 10 e 11, ci ha permesso di valutare ulteriormente la rilevanza delle variabili nel predire le classi di diabete (tipo 1 e tipo 2), integrando i risultati ottenuti dai metodi di correlazione e fornendo una selezione più robusta delle features da includere nelle analisi successive.

In questo modo, siamo riusciti a ottenere 85 variabili rilevanti per la correlazione di Pearson, 103 variabili per la correlazione di Spearman e 78 variabili per la random forest per il diabete di tipo 1. Ugualmente abbiamo ottenuto 84 variabili rilevanti per la correlazione di Pearson, 106 variabili per la correlazione di Spearman e 84 variabili per la random forest per il diabete di tipo 2. Per garantire una selezione delle features più accurata e completa, tenendo conto dell'importanza assegnata a ciascuna variabile da ogni metodo, abbiamo calcolato una media aritmetica basata sulla posizione relativa delle variabili nei risultati di ciascun approccio. Questa procedura ci ha permesso di ottenere un ranking più bilanciato e rappresentativo delle features da considerare per le successive analisi. A questo punto, abbiamo selezionato le prime 20 features con il ranking più elevato, risultanti dalla media aritmetica delle posizioni ottenute con i vari metodi utilizzati (Pearson, Spearman e Random Forest). Le features selezionate per i soggetti affetti da diabete di tipo 1 sono le seguenti:

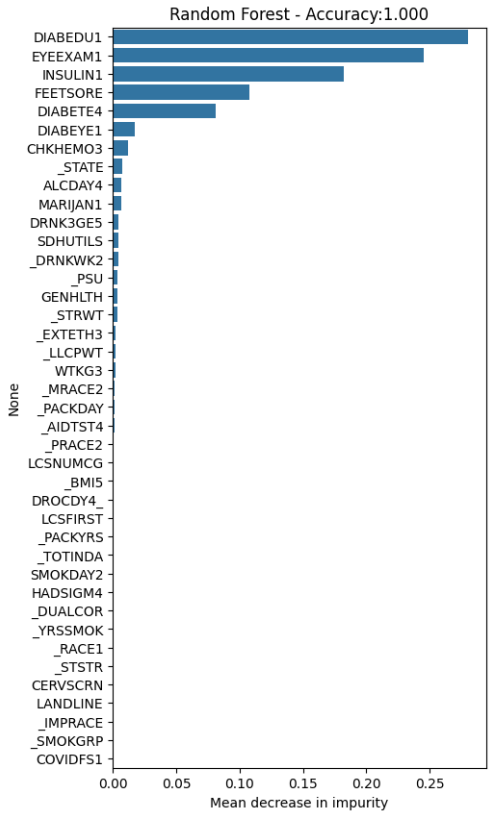

Figura 11: Random Forest diabete di tipo 2.

- DIABETE4: Ti è mai stato detto di avere il diabete?;
- GENHLTH: Stato di salute generale;
- INSULIN1: Stai prendendo insulina?;
- EYEEXAM1: uando è stata l'ultima volta che avete fatto una visita oculistica in cui le pupille sono state dilatate, rendendovi temporaneamente sensibili alla luce intensa?;
- CHKHEMO3: Quante volte negli ultimi 12 mesi un medico, un'infermiera o un altro operatore sanitario le ha fatto un controllo per l'A-one-C?;
- DIABEYE1: Quando è stata l'ultima volta che un medico, un infermiere o un altro operatore sanitario ha scattato una foto della parte posteriore del tuo occhio con una macchina fotografica specializzata?;
- DIABEDU1: Quando è stata l'ultima volta che hai seguito un corso o una lezione su come gestire autonomamente il diabete?;
- FEETSORE: Hai mai avuto piaghe o irritazioni ai piedi che hanno impiegato più di quattro settimane per guarire?;
- LSATISFY: In generale, quanto sei soddisfatto della tua vita?;
- MARIJAN1: Negli ultimi 30 giorni, quanti giorni hai usato marijuana o cannabis?;
- DIFFWALK: Avete serie difficoltà a camminare o a salire le scale?;
- EDUCA: Qual è il voto o l'anno scolastico più alto che hai completato?;
- CHECKUP1: Quanto tempo è passato dall'ultima volta che hai visitato un medico per un controllo di routine?;
- EXERANY2: Durante l'ultimo mese, oltre al tuo lavoro regolare, hai partecipato ad attività fisiche o esercizi come corsa, ginnastica ritmica, golf, giardinaggio o camminate per fare esercizio?;
- _RFHLTH: Adulti con salute buona o migliore;
- RRPHYSM2: Negli ultimi 30 giorni, hai avvertito qualche sintomo fisico, ad esempio mal di testa, mal di stomaco, tensione muscolare o battito cardiaco, a causa del modo in cui sei stato trattato in base alla tua razza?;
- RRTREAT: Negli ultimi 12 mesi, ritieni che in generale sei stato trattato peggio, uguale o migliore delle persone di altre razze?;
- RRHCARE4: Negli ultimi 12 mesi, quando hai richiesto assistenza sanitaria, ritieni che le tue esperienze siano state peggiori, uguali o migliori rispetto a quelle di persone di altre razze?;
- _WT2RAKE: Peso di progetto utilizzato nel rastrellamento;
- _STRWT: Peso strato;
- DIABTYPE1: Intervistati affetti da diabete di tipo 1.

Invece, le features selezionate per i soggetti affetti da diabete di tipo 2 sono le seguenti:

- DIABETE4: Ti è mai stato detto di avere il diabete?;
- GENHLTH: Stato di salute generale;
- INSULIN1: Stai prendendo insulina?;
- EYEEXAM1: uando è stata l'ultima volta che avete fatto una visita oculistica in cui le pupille sono state dilatate, rendendovi temporaneamente sensibili alla luce intensa?;
- CHKHEMO3: Quante volte negli ultimi 12 mesi un medico, un'infermiera o un altro operatore sanitario le ha fatto un controllo per l'A-one-C?;
- DIABEYE1: Quando è stata l'ultima volta che un medico, un infermiere o un altro operatore sanitario ha scattato una foto della parte posteriore del tuo occhio con una macchina fotografica specializzata?;

- DIABEDU1: Quando è stata l'ultima volta che hai seguito un corso o una lezione su come gestire autonomamente il diabete?;
- FEETSORE: Hai mai avuto piaghe o irritazioni ai piedi che hanno impiegato più di quattro settimane per guarire?;
- LSATISFY: In generale, quanto sei soddisfatto della tua vita?;
- MARIJAN1: Negli ultimi 30 giorni, quanti giorni hai usato marijuana o cannabis?;
- CHCKDNY2: Escludendo calcoli renali, infezioni della vescica o incontinenza, ti è mai stato detto che avevi una malattia renale?;
- ALCDAY4: Hai mai fatto una TAC o una TAC nella zona del torace?;
- SDHUTILS: Negli ultimi 12 mesi c'è stato un momento in cui un'azienda elettrica, del gas, del petrolio o dell'acqua ha minacciato di interrompere i servizi?;
- ADDEPEV3: (Hai mai detto) (di aver avuto) un disturbo depressivo?;
- _RFHLTH: Adulti con salute buona o migliore;
- DRNK3GE5: Considerando tutti i tipi di bevande alcoliche, quante volte negli ultimi 30 giorni hai bevuto 5 o più drink per gli uomini o 4 o più drink per le donne in un'occasione?;
- _BMI5: ndice di massa corporea (BMI);
- RRHCARE4: Negli ultimi 12 mesi, quando hai richiesto assistenza sanitaria, ritieni che le tue esperienze siano state peggiori, uguali o migliori rispetto a quelle di persone di altre razze?;
- LCSSCNCR: Qualcuna delle scansioni TC o CAT dell'area del torace è stata eseguita principalmente per controllare o diagnosticare il cancro ai polmoni?;
- _STRWT: Peso strato;
- DIABTYPE2: Intervistati affetti da diabete di tipo 2.

Queste 20 features, risultate tra le più rilevanti, saranno utilizzate nelle successive fasi di analisi per migliorare la predittività dei modelli.

4. Bayesian Network

Nel nostro studio, abbiamo sviluppato un modello predittivo per la presenza di diabete utilizzando reti bayesiane, sfruttando tecniche avanzate di machine learning per gestire e analizzare un ampio dataset.

4.1. Primo approccio

Il dataset preprocessato è stato suddiviso in due insiemi: un set di training e un set di test, utilizzando una divisione di 80/20 poichè, dal principio di Pareto (chiamato anche regola 80-20), si afferma che l'80% dell'effetto è guidato dal 20% delle cause e viceversa. La struttura della rete bayesiana è stata determinata utilizzando l'algoritmo Hill-Climbing, combinato con il metodo di punteggio K2 e un'indegree massimo di 2, per ottimizzare le dipendenze probabilistiche tra le variabili. Questo passaggio è cruciale per garantire che la rete possa catturare le complesse interazioni tra i vari fattori di rischio del diabete. Per valutare le prestazioni del modello, abbiamo deciso di voler predire i valori della variabile 'DIABETE4', descritta in Figura 14.

Una volta appresa la struttura ottimale, il modello di rete bayesiana è stato costruito e addestrato sui dati di training utilizzando l'estimatore di massima verosimiglianza (MLE). Questo processo ha consentito di calcolare le distribuzioni di probabilità condizionate per ciascun nodo della rete, basate sui dati disponibili. La rete risultante rappresenta una mappa probabilistica che può essere utilizzata per inferenze e previsioni accurate. Sono stati creati due diversi modelli di rete bayesiana nel nostro lavoro, uno per ciascun dataset ridotto ricavati dall'approccio uno e approccio due.

Il grafico fornito in Figura 12 è una visualizzazione della struttura della Rete Bayesiana costruita sul dataset da 12.600 soggetti. I nodi rappresentano variabili nel dataset e gli archi rappresentano dipendenze condizionali tra queste variabili. BMI5, WEIGHT2, HEIGHT3 e DIFFWALK sono nodi centrali, ovvero con il maggior numero di connessioni (archi). Questi nodi rappresentano variabili chiave che hanno un'influenza significativa su altre variabili o che sono fortemente influenzate da esse.

Il nodo **BMI5 (Body Mass Index)** ha numerose connessioni, in particolare è collegato direttamente a variabili come WEIGHT2 (peso) e HEIGHT3 (altezza), che, infatti, sono direttamente correlate al calcolo del BMI. Le connessioni con variabili come DIFFWALK (difficoltà a camminare) e PHYSHLTH (salute fisica) suggeriscono che il BMI potrebbe influenzare la mobilità e la salute fisica degli individui.

Nel grafico sono visibili dei cluster che rappresentano gruppi di variabili che sono strettamente correlate tra loro. Identificarli può aiutare a comprendere meglio le interrelazioni tra variabili nel dataset.

Il cluster composto dalle variabili *BMI, WEIGHT2, HEIGHT3, DIFFWALK, PHYSHLTH, e GENHLTH* indica una forte interrelazione tra le misure di peso, altezza, indice di massa corporea e i fattori di salute fisica e mobilità. Le variabili in questo cluster influenzano e sono influenzate dalla salute e dalla capacità di movimento.

Un secondo cluster include variabili come *SMOKE100 (fumatore), EXERANY2 (esercizio fisico), SLEPTIM1 (ore di sonno), e DIABTYPE (tipo di diabete)* e mostra la relazione tra comportamenti di stile di vita come il fumo, l'esercizio fisico e il sonno, e condizioni di salute come il diabete. Queste variabili possono avere un impatto significativo sulla

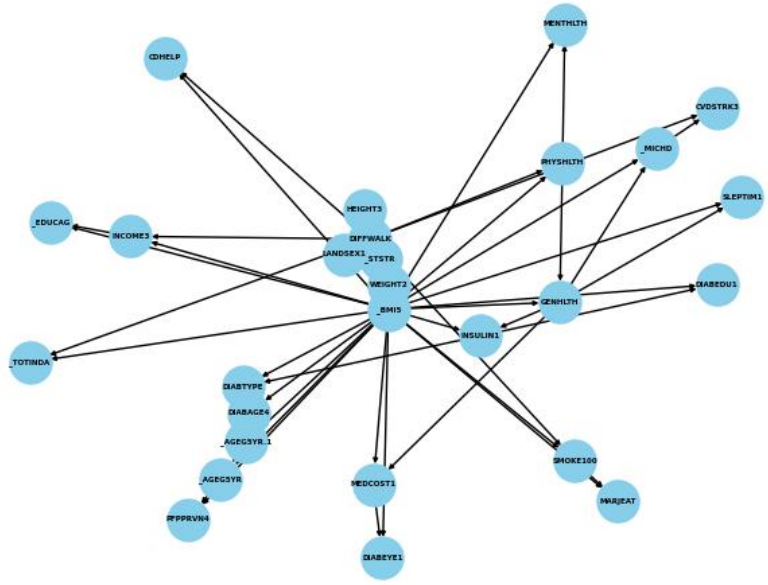

Figura 12: Struttura rete bayesiana (1)

salute generale degli individui.

Infine il cluster con variabili quali *INCOME3 (reddito) e EDUCA (educazione)* evidenzia come fattori socioeconomici come il reddito e l'educazione siano correlati e possano influenzare altri aspetti della vita e della salute degli individui.

Da tali connessioni possiamo dedurre che:

- il BMI emerge come un nodo cruciale con molte connessioni, sottolineando la sua importanza come indicatore di salute che influisce su e viene influenzato da molte altre variabili;
- come le variabili relative alla salute fisica, alla mobilità e allo stile di vita siano interconnesse, suggerendo che interventi in uno di questi ambiti potrebbero avere effetti a catena su altri;
- i fattori socioeconomici, anche se meno connessi direttamente, possono ancora avere un'influenza significativa sulle altre variabili della rete, mostrando l'importanza di considerare questi fattori nei programmi di salute pubblica.

4.2. Secondo approccio

Il grafico fornito in Figura 13 è una visualizzazione della struttura della Rete Bayesiana costruita sul secondo dataset analizzato in precedenza. Commentiamo le relazioni generate tra i nodi. Il nodo **EDUCAG** è connesso con **PHYSHLTH, GENHLTH, DISTRES**. Tali relazioni indicano che il livello di istruzione può influenzare la salute fisica, la salute generale e il livello di stress. Questo suggerisce che un'istruzione migliore potrebbe portare a una salute migliore e a un minor stress.

Il nodo **MICHNO**, numero di malattie croniche, è collegato con **GENHLTH, PHYSHLTH, MEDCOST** e assume che un maggior numero di malattie croniche influisce negativamente sulla salute generale e fisica, e aumenta i costi medici.

Il nodo **DISTRES** connesso con **MENTHLTH, PHYSHLTH, AGESGRY** indica che il livello di stress ha un impatto diretto sulla salute mentale, fisica e varia con l'età. Più stress può portare a problemi di salute mentale e fisica.

Il nodo **AGESGRY** collegato con **DIABETE4, WEIGHT2, GENHLTH** ci dice che l'età può influenzare la probabilità di avere il diabete, il peso e la salute generale. Con l'avanzare dell'età, le condizioni di salute possono deteriorarsi.

Il nodo **DIABETE4** è connesso con **GENHLTH, BMI5** e afferma che la presenza di diabete influisce negativamente sulla salute generale ed è correlata con l'indice di massa corporea.

Il nodo **MENTHLTH** collegato con **DISTRES, GENHLTH** suggerisce che la salute mentale è influenzata dallo stress e contribuisce alla salute generale.

La rete evidenzia come la *salute generale (GENHLTH)* sia un nodo centrale con diverse dipendenze. Questo è coerente con il concetto che la salute generale è influenzata da vari fattori come condizioni croniche, età, diabete, e stato fisico e mentale.

Inoltre, *la salute fisica (PHYSHLTH) e la salute mentale (MENTHLTH)* sono strettamente interconnesse e influenzano il benessere generale e i costi medici, evidenziando l'importanza di un approccio integrato alla salute.

Il *livello di istruzione (EDUCAG)* emerge come un fattore che può influenzare indirettamente la salute, suggerendo

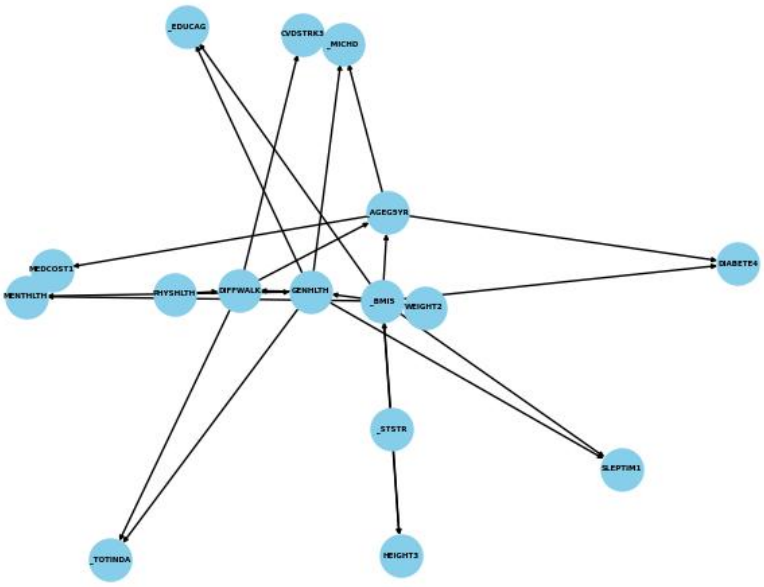

Figura 13: Struttura rete bayesiana (2)

che interventi educativi potrebbero migliorare la salute della popolazione riducendo lo stress e migliorando la salute generale.

È stato utilizzato il metodo di eliminazione variabile per eseguire previsioni sul set di test. Ogni previsione è stata fatta considerando le evidenze fornite dai dati di test, ed è stata confrontata con i valori reali di diabete. In particolare, le evidenze sono generate eliminando dal set di test la colonna che contiene i valori da predire, ovvero la variabile "DIABETE4" che rappresenta se un soggetto ha il diabete o meno. Le previsioni sono state raccolte e confrontate utilizzando diverse metriche di valutazione: accuratezza, precisione, recall e punteggio F1. Queste metriche sono state calcolate per quantificare l'efficacia del modello nel predire correttamente l'incidenza del diabete.

4.3. Terzo Approccio

Con questo ultimo approccio, la struttura della Rete Bayesiana risulta essere quella presente in figura 15 e 16, più nello specifico, nella rete del diabete di tipo 1, il nodo centrale è "INSULINA", che sottolinea l'importanza dei livelli di insulina nella patogenesi del diabete di tipo 1. Nella rete del diabete di tipo 2, il nodo centrale è "BMI5", che evidenzia l'importanza dell'indice di massa corporea come predittore di rischio. La rete del diabete di tipo 2 è più estesa e complessa, con un maggior numero di connessioni dirette dal nodo centrale ai nodi periferici, mentre la rete del diabete di tipo 1 ha una struttura leggermente più compatta con meno nodi centrali. La rete del diabete di tipo 1 sembra focalizzarsi maggiormente su fattori genetici e clinici, mentre quella del tipo 2 integra anche fattori comportamentali o legati allo stile di vita, come il BMI e l'alcol. Per trarre i nostri risultati, abbiamo suddiviso il dataset in due parti, destinando l'80% delle osservazioni al campione di addestramento e il restante 20% al campione di validazione. Sul campione di addestramento, abbiamo implementato una procedura di cross-validation, suddividendolo in 4 fold stratificati. Questo ci ha permesso di bilanciare adeguatamente la presenza di soggetti sani e malati in ciascun sotto-campione, garantendo che la distribuzione delle classi fosse rappresentativa in ogni fold. Successivamente, abbiamo utilizzato il restante 20% delle osservazioni per validare il modello. Le metriche di performance ottenute per i soggetti affetti da diabete di tipo 1 sono le seguenti:

- Accuratezza media (CV): 0.6690 vs Accuratezza sul set di test: 0.7162: L'accuratezza migliora nel set di test rispetto alla media ottenuta durante la cross-validation, indicando che il modello generalizza leggermente meglio fuori dal campione di addestramento.
- Precisione media (CV): 0.7893 vs Precisione sul set di test: 0.8204: La precisione, che misura la percentuale di veri positivi tra tutte le previsioni positive, è elevata, suggerendo che il modello fa poche false previsioni di positivi (pochi falsi positivi).
- Recall medio (CV): 0.6951 vs Recall sul set di test: 0.7124: Il recall, che indica la capacità del modello di identificare correttamente i casi di diabete di tipo 1, è buono, anche se potrebbe essere migliorato ulteriormente per catturare un numero maggiore di veri positivi.
- F1 Score medio (CV): 0.6463 vs F1 Score sul set di test: 0.6887: L'F1 score bilancia precisione e recall.

Label: (Ever told) you had diabetes
Section Name: Chronic Health Conditions
Core Section Number: 7
Question Number: 12
Column: 129
Type of Variable: Num
SAS Variable Name: DIABETE4
Question Prologue:
Question: (Ever told) (you had) diabetes? (If 'Yes' and respondent is female, ask 'Was this only when you were pregnant? '. If Respondent says pre-diabetes or borderline diabetes, use response code 4.)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	61,158	13.74	12.04
2	Yes, but female told only during pregnancy—Go to Section 08.01 AGE	3,836	0.86	1.01
3	No—Go to Section 08.01 AGE	368,722	82.83	84.34
4	No, pre-diabetes or borderline diabetes—Go to Section 08.01 AGE	10,329	2.32	2.27
7	Don't know/Not Sure—Go to Section 08.01 AGE	763	0.17	0.23
9	Refused—Go to Section 08.01 AGE	321	0.07	0.11
BLANK	Not asked or Missing	3	.	.

Figura 14: Descrizione variabile DIABETE4

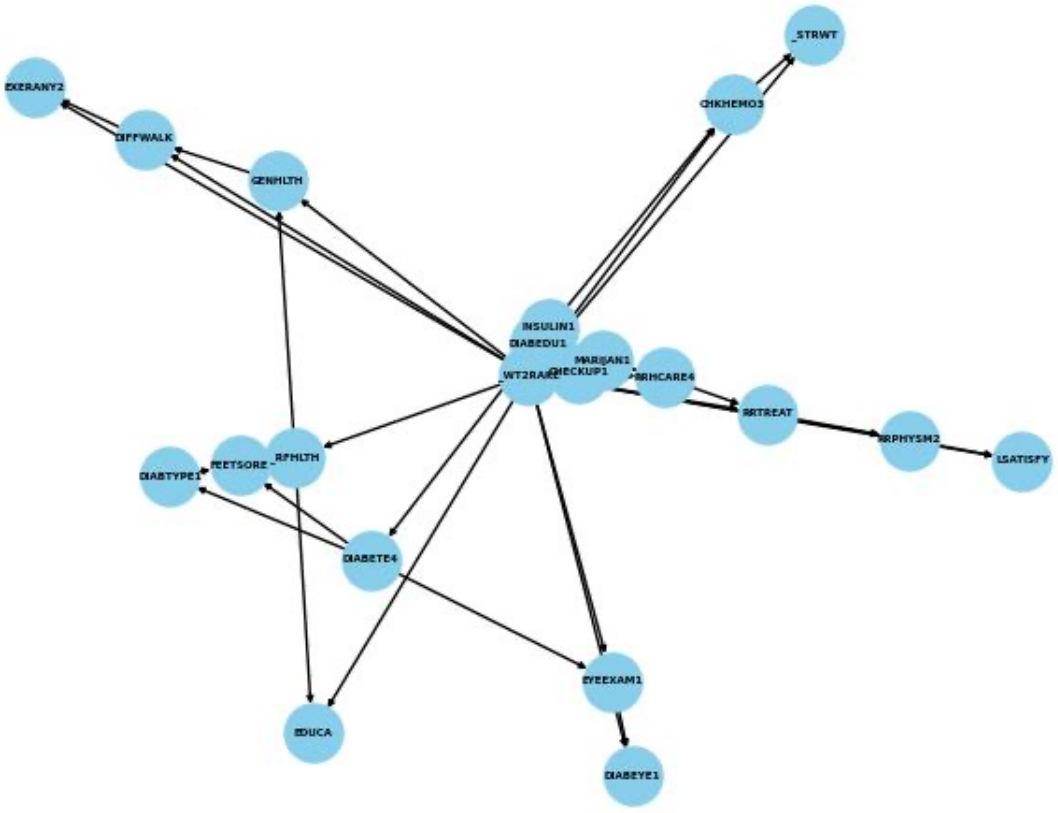

Figura 15: Struttura Rete Bayesiana DIABTYPE1.

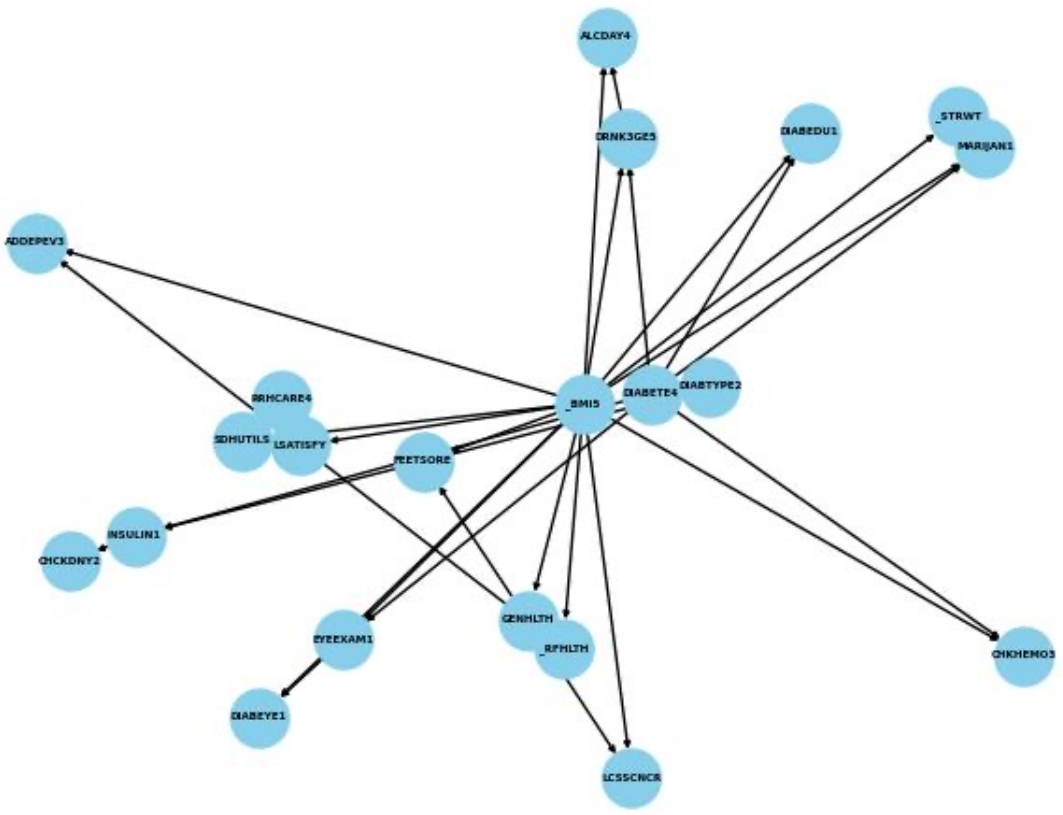

Figura 16: Struttura Rete Bayesiana DIABTYPE2.

Qui il modello dimostra prestazioni equilibrate e migliorate nel set di test rispetto alla cross-validation, il che indica una buona performance complessiva.

Specularmente, le metriche di performance ottenute per i soggetti affetti da diabete di tipo 1 sono le seguenti:

- Accuratezza media (CV): 0.6682 vs Accuratezza sul set di test: 0.6854: Simile al diabete di tipo 1, anche qui il modello mostra una leggera miglioria nell'accuratezza nel set di test rispetto alla media della cross-validation, pur rimanendo in una fascia moderata.
- Precisione media (CV): 0.7880 vs Precisione sul set di test: 0.7905: La precisione è stabile tra cross-validation e test, indicando una buona capacità del modello di minimizzare i falsi positivi anche per il diabete di tipo 2.
- Recall medio (CV): 0.6979 vs Recall sul set di test: 0.7210: Il recall è leggermente migliore nel test rispetto alla cross-validation, il che significa che il modello sta identificando correttamente una percentuale maggiore di casi di diabete di tipo 2.
- F1 Score medio (CV): 0.6486 vs F1 Score sul set di test: 0.6740: Anche qui l'F1 score migliora nel set di test, indicando che il modello ha trovato un buon equilibrio tra precisione e recall per il diabete di tipo 2.

In entrambi i casi, i risultati del test mostrano un miglioramento rispetto alla cross-validation, il che suggerisce che il modello si adatta bene ai dati nuovi. Tuttavia, ci sono margini di miglioramento, specialmente in termini di recall, per identificare una percentuale maggiore di casi di diabete. L'F1 score indica che il modello bilancia bene precisione e recall, con prestazioni leggermente migliori per il diabete di tipo 1 rispetto al tipo 2.

I risultati della cross-validation con 5 fold stratificati mostrano un notevole miglioramento rispetto alla versione precedente con 4 fold, soprattutto per il diabete di tipo 1. Nel dettaglio, per gli intervistati affetti da diabete di tipo 1 risultano essere:

- Accuratezza media: 0.9694: Il modello mostra un'eccellente accuratezza, indicando che quasi tutte le previsioni sono corrette.
- Precisione media: 0.9770: La precisione è molto alta, suggerendo che il modello fa pochissimi falsi positivi, ovvero quando predice diabete di tipo 1, ha quasi sempre ragione.

- Recall medio: 0.9639: Il modello riesce a identificare correttamente quasi tutti i soggetti affetti da diabete di tipo 1, dimostrando un'ottima capacità di rilevazione dei veri positivi.
- F1 Score medio: 0.9669: L'F1 score, che bilancia precisione e recall, è altrettanto elevato, confermando che il modello gestisce in modo molto efficace entrambi gli aspetti, con prestazioni eccezionali.

Invece, per gli intervistati affetti da diabete di tipo 2 le metriche risultano essere:

- Accuratezza media: 0.7496: Sebbene più bassa rispetto al diabete di tipo 1, l'accuratezza per il tipo 2 è ancora rispettabile, anche se il modello potrebbe migliorare ulteriormente in termini di previsione generale.
- Precisione media: 0.8270: La precisione è elevata, indicando che il modello ha buone performance nel limitare i falsi positivi per il diabete di tipo 2.
- Recall medio: 0.7611: Il recall è buono, ma c'è ancora spazio per migliorare la capacità di individuare un numero maggiore di casi di diabete di tipo 2 rispetto al tipo 1, dove il recall è molto più elevato.
- F1 Score medio: 0.7379: L'F1 score riflette una discreta bilanciatura tra precisione e recall, con buone prestazioni complessive per il diabete di tipo 2, anche se leggermente inferiori rispetto al diabete di tipo 1.

I risultati della cross-validation a 5 fold stratificati mostrano un'eccellente performance per il diabete di tipo 1, con metriche molto elevate su tutti i fronti. Per il diabete di tipo 2, pur essendo le prestazioni più modeste, il modello offre comunque buoni risultati, specialmente in termini di precisione. Il miglioramento del recall per il diabete di tipo 2 potrebbe aiutare a bilanciare ulteriormente il modello, permettendo una maggiore identificazione dei soggetti affetti.

5. Results

Le prestazioni del modello di inferenza sono state valutate utilizzando metriche standard: accuratezza, precisione, recall e F1 score. Tali metriche sono essenziali per comprendere come il modello si comporta nella classificazione dei dati. Il modello è stato testato per prevedere i valori della variabile "DIABETE4" e ha raggiunto un'accuratezza di 0,3937, indicando che circa il 39,37% delle previsioni effettuate dal modello erano corrette. Questo valore relativamente basso di accuratezza suggerisce che il modello non riesce a distinguere adeguatamente tra casi positivi e negativi. Nel nostro lavoro, la causa di un valore così basso è dovuto alla dimensionalità del dataset che nel primo approccio è stato eccessivamente ridotto. La precisione, che misura la proporzione di vere predizioni positive tra tutte le predizioni positive, è stata di 0,5. Ciò significa che il 50% delle predizioni positive effettuate dal modello erano corrette, ovvero che quando il modello predice un caso di diabete, ha una probabilità del 50% di essere corretto. La precisione è particolarmente importante in contesti in cui i falsi positivi possono avere conseguenze significative, come nel caso della diagnosi di malattie. Tuttavia, l'attuale livello di precisione suggerisce che ci sono molte predizioni errate, che possono portare a diagnosi non necessarie

La recall, o sensibilità, che valuta la capacità del modello di identificare correttamente tutti i casi rilevanti, è stato di 0,1968. Questo basso valore di richiamo suggerisce che il modello ha mancato un numero significativo di veri casi positivi. Questo indica che solo il 19,68% dei casi reali di diabete sono stati identificati correttamente dal modello, mentre l'80,32% dei casi è stato mancato. Un richiamo basso è particolarmente preoccupante in scenari clinici, dove è cruciale identificare tutti i casi positivi per garantire un trattamento tempestivo.

L'F1 score, che è la media armonica di precisione e richiamo, è stato di 0,2824. L'F1 score fornisce un equilibrio tra precisione e richiamo, e un punteggio di 0,2824 riflette la prestazione complessivamente moderata del modello nel classificare i casi.

Questi risultati evidenziano le difficoltà affrontate dal modello nel prevedere accuratamente la variabile target, suggerendo che sono necessarie ulteriori rifiniture e ottimizzazioni per migliorare le sue prestazioni.

Dal **secondo approccio**, invece, abbiamo le seguenti metriche **accuracy** pari a 0.71, **precision** di 0.27, **recall** di 0.27, e **F1 score** di 0.26. L'accuracy di 0.71 indica che il modello è stato in grado di effettuare correttamente il 71% delle previsioni. Tuttavia, le metriche di precision, recall e F1 score mostrano valori inferiori suggerendo alcune limitazioni del modello nel distinguere correttamente le classi. In particolare, la bassa precision indica una quantità significativa di falsi positivi, mentre il recall basso segnala un numero considerevole di falsi negativi. L'F1 score, che rappresenta una media armonica tra precision e recall, conferma ulteriormente la difficoltà del modello nel bilanciare tra queste due metriche. Le basse performance complessive di precision, recall e F1 score suggeriscono che il modello potrebbe trarre beneficio da un'ulteriore ottimizzazione dei parametri o dall'inclusione di nuove feature informative nel dataset. Nonostante ciò, l'accuracy relativamente elevata offre una base promettente per miglioramenti futuri.

Infine, nel **terzo approccio** sono state modificate le variabili target ponendo l'attenzione sul tipo di diabete e non più solo sulla presenza o meno di esso in un paziente. Mediante le modifiche apportate, riportate precedentemente, abbiamo ottenuto le seguenti metriche per i pazienti affetti da **diabete di tipo 1**: **accuracy** pari a 0.96, **precision** di 0.97, **recall** di 0.96, e **F1 score** di 0.96; invece, per i pazienti affetti da **diabete di tipo 2**: **accuracy** pari a 0.74, **precision** di 0.82, **recall** di 0.76, e **F1 score** di 0.74. Ciò indica un buon funzionamento del nostro modello di rete Bayesiana quando predice sulle variabili "DIABTYPE1" e "DIABTYPE2"

6. Conclusions

Questo studio ha presentato un'analisi completa sull'utilizzo delle reti bayesiane (BN) per prevedere l'insorgenza del diabete e identificare le sue cause. Sfruttando il dataset del Behavioral Risk Factor Surveillance System (BRFSS) 2022, che contiene un'ampia gamma di variabili correlate alla salute, siamo stati in grado di esplorare le complesse relazioni tra diversi fattori di rischio e il diabete. La ricerca dimostra con successo l'utilità delle reti bayesiane come modello predittivo per le malattie croniche, in particolare il diabete. L'obiettivo è riuscire a predire la presenza o meno di diabete avendo a disposizione le risposte dei soggetti a delle interviste e non dati o analisi cliniche.

Lo studio ha impiegato algoritmi Random Forest e misure di correlazione per effettuare la selezione delle caratteristiche. Questo approccio ci ha permesso di *ridurre la dimensionalità* del dataset mantenendo i predittori più rilevanti per il diabete, migliorando così l'efficienza del modello.

Il nostro *modello di rete bayesiana* ha mostrato forti capacità predittive, mostrando come migliorano le metriche modificando i dati che gli vengono dati in input. Sono stati ottenuti alti punteggi su diverse metriche, tra cui l'accuratezza (circa il 71%). Ciò indica che le BN sono utili ed efficaci per la previsione dell'insorgenza del diabete ma anche per comprendere le relazioni causali tra vari fattori di salute e stile di vita.

L'analisi ha rivelato che età, obesità, ipertensione e mancanza di attività fisica sono i *fattori di rischio più significativi* per il diabete.

L'*analisi demografica* ha indicato una maggiore prevalenza del diabete tra le fasce di età più avanzate, sottolineando la necessità di misure preventive specifiche per età.

Dal terzo approccio, invece, possiamo affermare che il diabete di tipo 1 è influenzato da fattori genetici, mentre il diabete di tipo 2 è influenzato da una gamma più ampia di fattori, in particolare dall'indice di massa corporea e da fattori di rischio legati allo stile di vita.

La capacità di prevedere l'insorgenza del diabete consente una migliore allocazione delle risorse nei sistemi sanitari, garantendo che le risorse mediche siano destinate agli individui con i profili di rischio più elevati.

I risultati ottenuti dalla rete bayesiana possono essere utilizzati per *personalizzare* gli interventi medici, migliorando l'efficacia dei piani di trattamento e delle modifiche dello stile di vita per i pazienti a rischio di sviluppare il diabete.

In conclusione, questa ricerca evidenzia il potenziale delle reti bayesiane come strumento per prevedere il diabete e analizzare la complessa rete di fattori che contribuiscono alla sua insorgenza. Mediante l'analisi di tali relazioni, lo studio contribuisce al campo più ampio della previsione e prevenzione delle malattie croniche, con l'obiettivo finale di ridurre il carico globale del diabete attraverso interventi di salute pubblica informati.

Riferimenti bibliografici

[1] , URL: https://www.cdc.gov/brfss/annual_data/annual_2022.html.

[2] Danli Kong, Rong Chen, Y.C., 2024. Bayesian network analysis of factors infuencing type 2 diabetes, coronary heart disease, and their comorbidities .

[3] Kai Bian, Mengran Zhou, F.H., Lai, W., 2020. Rf-pca: A new solution for rapid identification of breast cancer categorical data based on attribute selection and feature extraction .

[4] Li, K., Peng, H., Zhou, X., Li, S., 2016. Feature selection based on multiple correlation measures for medical examination dataset, in: 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), IEEE. pp. 845–849.

[5] MARWA HUSSEIN MOHAMED, MOHAMED HELMY KHAFAGY NESMA MOHAMED MAHMOUD KAMEL, W.S., 2024. Diabetic mellitus prediction with brfss data sets .