Modello predittivo basato su machine learning
per l'individuazione precoce del diabete di tipo 2

Introduzione

Il diabete è una malattia cronica che compromette la capacità dell'organismo di regolare i livelli di glucosio nel sangue, con possibili conseguenze sulla qualità e sull'aspettativa di vita. Pur non essendo curabile, può essere gestito efficacemente attraverso la perdita di peso, una dieta equilibrata, l'attività fisica regolare e cure mediche appropriate, che aiutano a ridurre i danni associati alla malattia.

Una diagnosi precoce consente di intervenire tempestivamente con modifiche dello stile di vita e trattamenti mirati. In questo contesto, i modelli predittivi del rischio di diabete rappresentano strumenti preziosi per il paziente, il medico e i decisori sanitari.

L'obiettivo di questo studio è sviluppare un algoritmo in grado di identificare il rischio individuale di diabete a partire dalle informazioni fornite dal paziente stesso, fornendo così un supporto concreto all'analisi medica.

Il modello realizzato mostra buone prestazioni predittive, pur con margini di miglioramento, e individua i parametri sui quali il medico può agire per sensibilizzare il paziente.

Grazie a un'interfaccia semplificata per l'inserimento dei dati, il modello può essere utilizzato sia dal medico sia dal paziente, offrendo una visualizzazione immediata della probabilità di essere affetti da diabete.

Descrizione dei dati

Per la costruzione e la validazione del modello è stato utilizzato il dataset[1] derivante dal sondaggio telefonico BRFSS 2015 del CDC (Centers for Disease Control and Prevention) degli Stati Uniti d'America.

La tabella originale è costituita da 253.680 righe, una per ogni intervistato, e da 22 colonne di seguito elencate. La variabile obiettivo è ***Diabetes*** (presenza o assenza di diabete), mentre le restanti colonne costituiscono le variabili predittive (input).

- ***Diabetes***: 0 = "*no diabetes*", 1 = "*prediabetes*", 2 = "*diabetes*".
- ***Hypertension***: 0 = nessuna ipertensione, 1 = ipertensione.
- ***HighChol***: 0 = no colesterolo alto, 1 = sì colesterolo alto.
- ***CholCheck***: 0 = nessun controllo del colesterolo negli ultimi 5 anni, 1 = sì controllo del colesterolo negli ultimi 5 anni
- ***BMI***: indice di massa corporea (kg/m^2).
- ***Smoker***: hai fumato almeno 100 sigarette in tutta la tua vita? (5 pacchetti = 100 sigarette) 0 = no, 1 = sì.

[1] Trattandosi di dati autoriportati, non è stato possibile verificare la veridicità delle risposte.

- ***Stroke***: ictus: 0 = no, 1 = sì.
- ***HeartDiseaseorAttack***: malattia coronarica (CHD) o infarto del miocardio (IM): 0 = no, 1 = sì.
- ***PhysActivity***: attività fisica negli ultimi 30 giorni, lavoro escluso: 0 = no, 1 = sì.
- ***Fruits***: consumo di frutta una o più volte al giorno: 0 = no, 1 = sì.
- ***Veggies***: consumo di verdura una o più volte al giorno: 0 = no, 1 = sì.
- ***HvyAlcoholConsump***: maschio adulto, più di 14 bicchieri a settimana; donna adulta, più di 7 bicchieri a settimana: 0 = no, 1 = sì.
- ***AnyHealthCare***: copertura sanitaria: 0 = no, 1 = sì.
- **NoDocbcCost**: impossibilità di consultare un medico a causa dei costi negli ultimi 12 mesi: 0 = no, 1 = sì.
- ***GenHlth***: stato di salute in generale (scala 1-5): 1 = eccellente, 2 = molto buono, 3 = buono, 4 = discreto, 5 = cattivo.
- ***MentHlth***: stato mentale negativo (stress, depressione, ansia, …) negli ultimi 30 giorni (scala 0-30).
- ***PhysHlth***: stato fisico negativo (malessere generale, malattia, …) negli ultimi 30 giorni (scala 0-30).
- ***DiffWalk***: seria difficoltà a camminare o a salire le scale: 0 = no, 1 = sì.
- ***Sex***: 0 = femmina, 1 = maschio.
- ***Age***: categorie di età: 1 = -24, 2 = 25-29, 3 = 30-34, 4 = 35-39, 5 = 40-44, 6 = 45-49, 7 = 50-54, 8 = 55-59, 9 = 60-64, 10 = 65-69, 11 = 70-74, 12 = 75-79, 13 = 80+.
- **Education**: scolarità: 1 = Never attended, 2 = Elementary, 3 = High school, 4 = High school, 5 = College, 6 = College graduate.
- **Income**: categorie di reddito (in dollari): 1 = <10.000, 2 = 10.000-<15.000, 3 = 15.000-<20.000, 4 = 20.000-<25.000, 5 = 25.000-<35.000, 6 = 35.000-50.000, 7 = 50.000-<75.000, 8 = 75.000+.

Preprocessing

Dalla tabella originale sono state eliminate 4 colonne (*Income*, *Education*, *AnyHealthcare* e *NoDocbcCost*), in quanto non applicabili alla realtà italiana. Sono state altresì eliminate 4.631 righe della classe "*prediabetes*" dalla colonna *Diabetes* considerata non pertinente allo studio in oggetto, così come le righe con valori di BMI estremi (inferiori a 15 o superiori a 50). **La tabella finale** è risultata composta da **246.875 osservazioni**.

Sono state poi effettuate le seguenti trasformazioni:

- classificazione della colonna *BMI* in cinque gruppi:

  15-*18,4 (sottopeso), 18,5-24,9 (normopeso), 25-29,9 (sovrappeso), 30-39,9 (obesità) e 40-50 (obesità grave);*

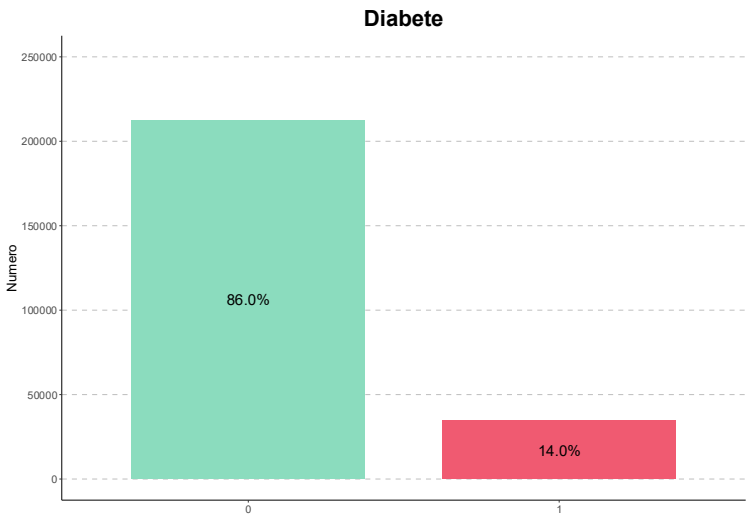
- classificazione della colonna *Age* in cinque gruppi:
  <35, 35-49, 50-64, 65-79, ≥80;

- accorpamento delle colonne *Fruits* e *Veggies* nella colonna *HealthyFood;*

- accorpamento delle colonne *HeartDiseaseorAttack* e *Stroke* nella colonna *VascularDisease;*

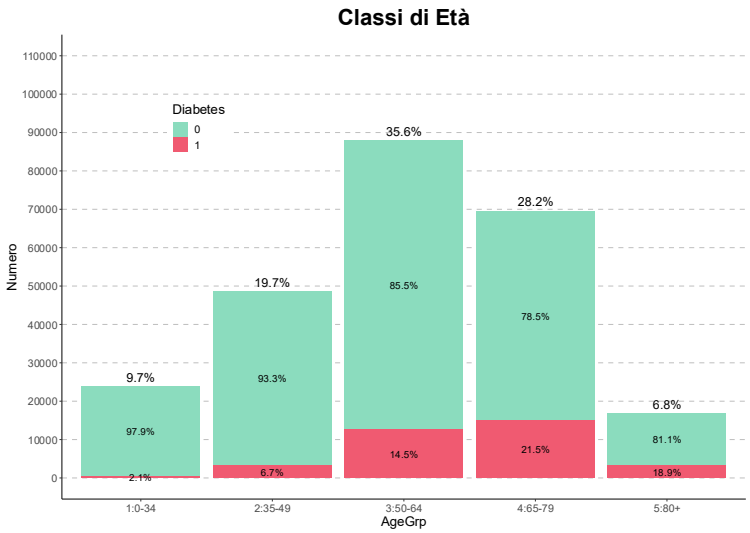
- le colonne *PhysHlth* e *MentHlth* (giorni di stato fisico e mentale negativi) sono state classificate in 2 categorie: 0=buono, 1=non buono.

Esplorazione dei dati



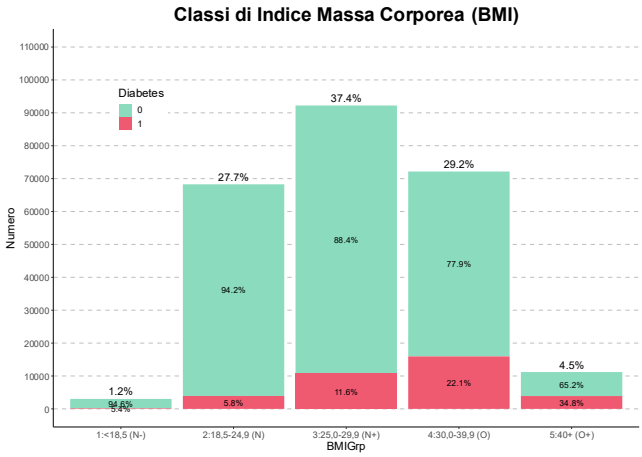
Il diagramma a barre rappresentato a sinistra mostra la **prevalenza di soggetti diabetici (14,0%)** e dei **soggetti non diabetici (86,0%)**.

Nella costruzione del modello si dovrà tener conto della discrepanza delle due classi.

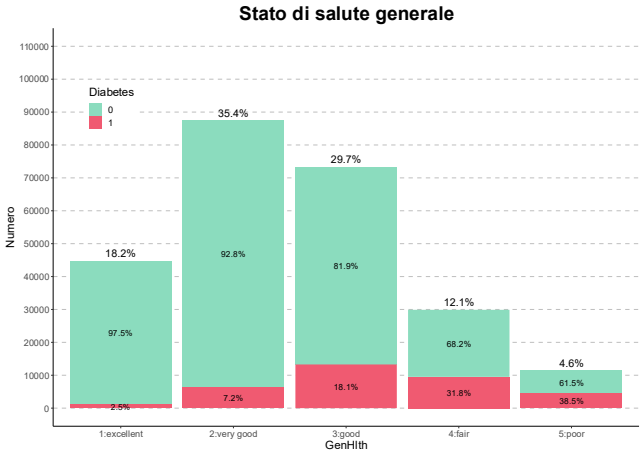


I diagrammi a barre sovrapposte rappresentano ogni categoria con una barra la cui altezza è proporzionale alla sua frequenza, mentre i segmenti all'interno di ogni barra corrispondono alla prevalenza di soggetti diabetici e no.

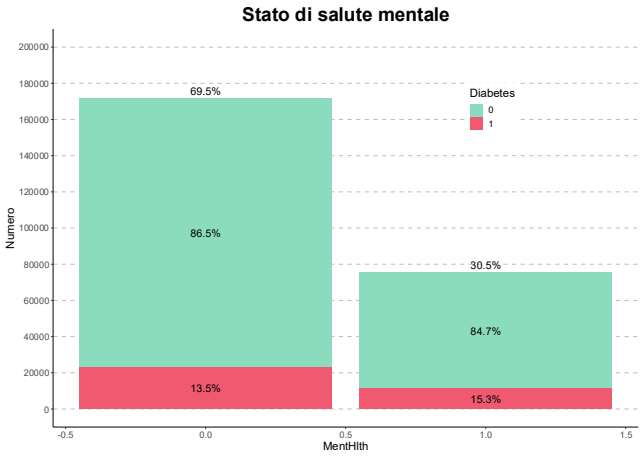
Nel diagramma delle classi di età la prevalenza di soggetti diabetici si verifica nella fascia 65-79 anni (21,5%), a seguire in quella degli ultraottantenni (18,9%).



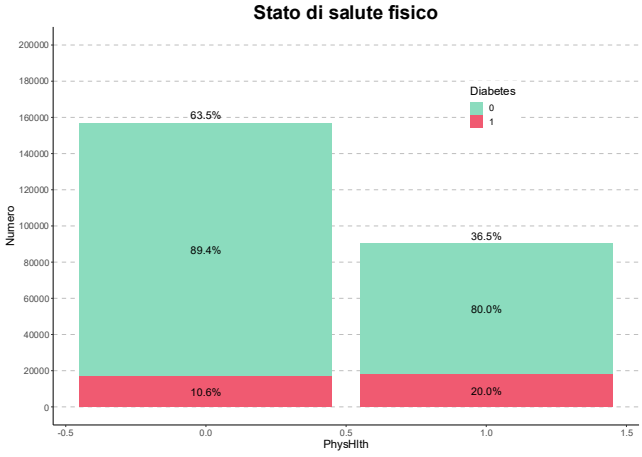
Le classi di BMI che hanno prevalenza di soggetti diabetici sono quella di "*obesità grave*" (34,8%), seguita da quella di "*obesità*" (22,1%) e da quella di "*sovrappeso*" (11,6%).

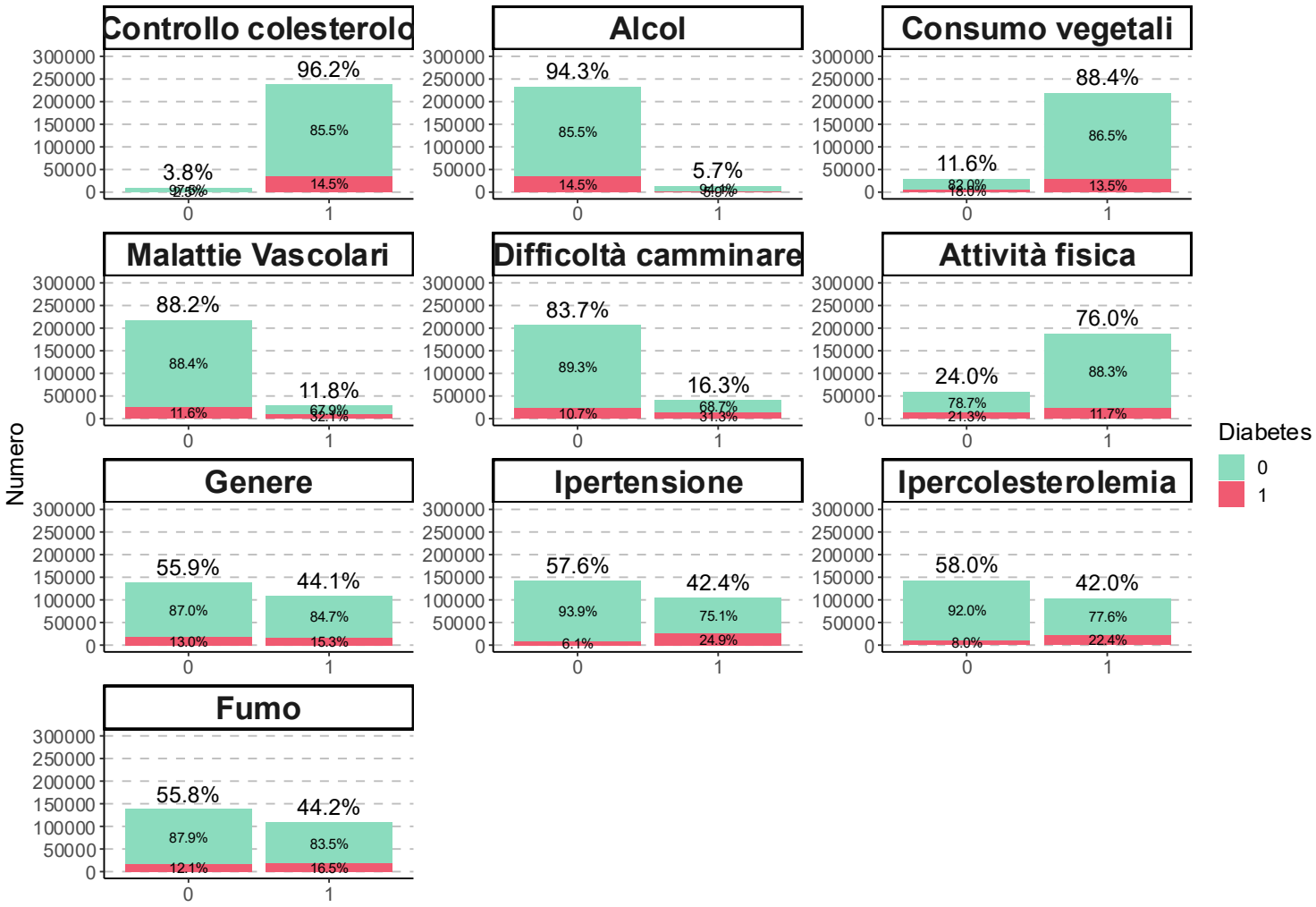


Anche lo stato di salute generale rivela una prevalenza di soggetti diabetici tra coloro che lo dichiarano come "*cattivo*" (38,5%) e da quelli che lo dichiarano "*discreto*" (31,8%).



La prevalenza di soggetti diabetici si ha tra coloro che lamentano uno stato di malessere mentale o fisico negli ultimi 30 giorni (rispettivamente il 15,3% e il 20,0%).





I diagrammi sopra riportati rappresentano l'analogo per le variabili binarie (es. colesterolemia alta o normale).

Da questa serie di diagrammi si evincono delle classi sovra rappresentate, ovvero:

- il 96,2% dei soggetti si è sottoposto a un controllo del colesterolo almeno una volta negli ultimi 5 anni;
- il 94,3% dichiara di non essere un forte consumatore di alcol;
- l'88,4% dei soggetti consuma di solito frutta o verdura;
- l'88,2% dei soggetti non ha avuto malattie di tipo vascolare.

Dalle restanti variabili si evince che i diabetici, per lo più, sono soggetti che:

- ***Sono di sesso maschile (15,3% contro il 13,0%)***
- ***Sono ipertesi (24,9% contro il 6,1%);***
- ***Hanno valori alti di colesterolo (22,4% contro l'8,0%)***
- ***Non praticano alcuna attività fisica (21,3% contro l'11,7%)***
- ***Fumano (16,5% contro il 12,1%)***
- ***Hanno difficoltà nel camminare (31,3% contro il 10,7%)***

Analisi dei dati

L'**analisi delle correlazioni tra le variabili numeriche** conferma quanto già emerso dall'esplorazione descrittiva.

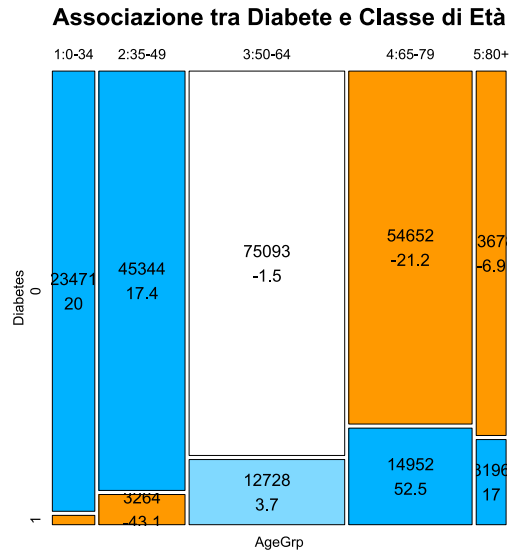
Il diabete risulta debolmente correlato in modo positivo con:

- ***L'ipertensione (0,27)***
- ***Il colesterolo alto (0,21)***
- ***La difficoltà nel camminare (0,22)***
- ***I problemi di tipo vascolare (0,19)***

Si osserva inoltre una debole correlazione negativa con la pratica di attività fisica (-0,12).

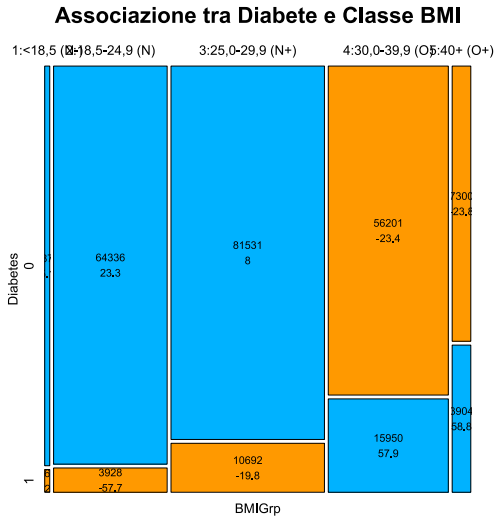
I **diagrammi a mosaico** hanno permesso di analizzare le associazioni tra variabili categoriche:

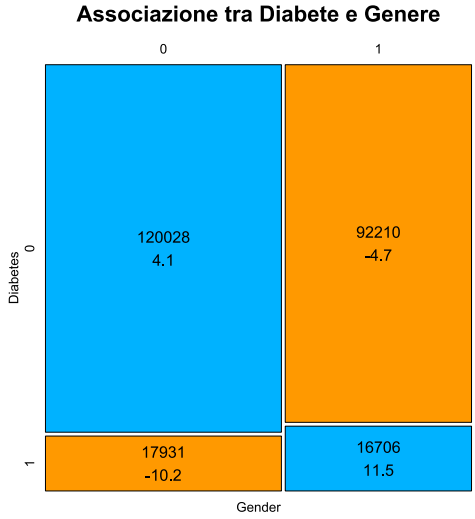
- la dimensione delle celle è proporzionale alle frequenze osservate in ogni combinazione delle categorie;
- l'intensità del colore è proporzionale al valore assoluto dei residui[2] e quindi rappresenta la "forza" dell'associazione (il colore azzurro indica associazione positiva, il colore arancione negativa).
- I valori all'interno delle celle indicano rispettivamente le frequenze delle combinazioni e il valore dei residui.



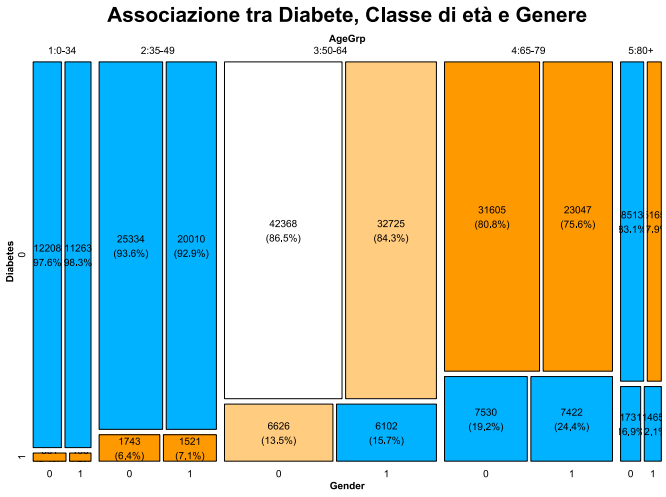
Ad esempio, considerando la **classe di età** si nota una forte **associazione positiva tra il diabete e la classe di età 65-79, seguita da quella 80+** e poi, ma più debole, quella 50-64. Viceversa, c'è una forte associazione negativa tra il diabete e l'età inferiore ai 50 anni.

Il diagramma a sinistra mostra che **il diabete** è meno frequente del previsto nelle classi "*sottopeso*", "*normopeso*" e "*sovrappeso*", mentre diventa molto **più frequente nelle classi di "*obesità*", soprattutto in quella grave**. Quello a destra evidenzia invece che le donne sono più rappresentate tra i soggetti non diabetici e meno tra quelli diabetici, mentre per gli uomini accade l'opposto, con un'**associazione positiva tra diabete e genere maschile**.

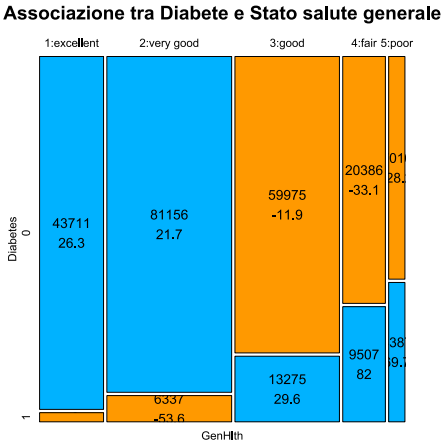




[2] I residui sono le differenze (positive o negative) tra le frequenze osservate e le frequenze attese sotto l'ipotesi di indipendenza statistica.

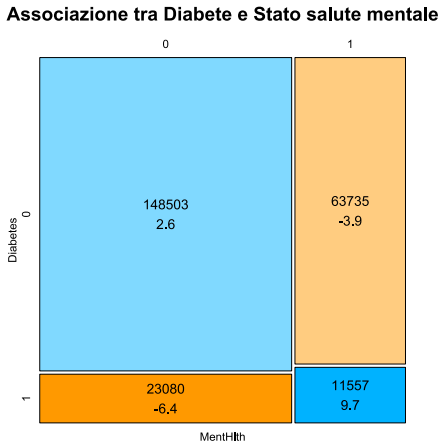


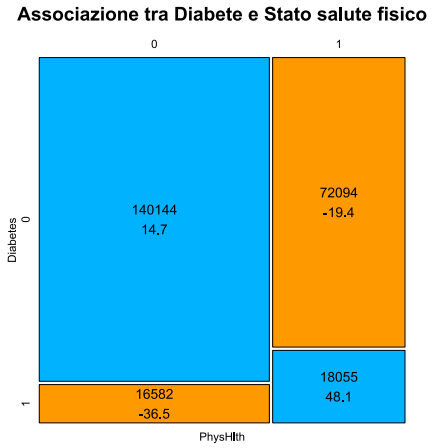
Da questo diagramma **a tre dimensioni** si evince che nelle **fasce più giovani l'associazione con il diabete è negativa**, indicando una frequenza di casi inferiore a quella attesa, mentre **dai 65 anni in poi** la tendenza si inverte e **l'associazione diventa positiva**, con un eccesso di casi che si accentua negli ultraottantenni, **senza però differenze particolari legate al genere**. Qui i valori tra parentesi non indicano i residui ma le percentuali.



Questo diagramma mostra che tra i soggetti non diabetici prevalgono quelli nello **stato di salute generale** "*eccellente*", mentre l'**associazione con il diabete diventa positiva** già con **stato "*buono*"** e si rafforza nelle condizioni **"*discreto*" e "*cattivo*"**.

Nel diagramma sottostante a destra emerge che il diabete è meno frequente nei soggetti senza problemi fisici e molto più frequente in quelli con problemi fisici.





Questo risultato implica che esiste un'associazione positiva molto forte tra il diabete e:

- ***Età: soprattutto nelle classi "65-79" e "80+" anni***
- ***BMI: nelle classi "obesità" e "obesità grave"***
- ***Stato di salute generale: nelle condizioni "discreto" e "cattivo"***
- ***Genere: nel sesso maschile, seppur con una tendenza meno marcata***

Modeling

Il dataset è stato suddiviso in modo casuale, con stratificazione sulla variabile obiettivo (per mantenere le stesse proporzioni di soggetti diabetici e non diabetici), in due parti:

- una parte (80% dei dati), chiamato "*training set*", è stata utilizzata per la costruzione del modello;
- il rimanente (20% dei dati), denominato "*test set*", è servito per misurare la capacità del modello di produrre buone predizioni su nuovi dati.

La fase di costruzione del modello si è svolta attraverso i seguenti passaggi:

1. Per la fase di modeling si è utilizzata una funzione di "*feature selection*"[3] per scegliere un sottoinsieme rilevante e informativo delle variabili usate come predittori presenti nella tabella. Le variabili selezionate sono state: *Hypertension*, *HighChol*, *BMIGrp*, *AgeGrp*, *GenHlth*, *DiffWalk*, *VascularDisease* e *PhysActivity*.

2. Essendo la tabella sbilanciata rispetto alla variabile obiettivo (i soggetti diabetici rappresentano solo il 14%, contro l'86% di quelli non diabetici), si è adottata una tecnica di sovracampionamento per compensare questa disparità.

3. Per evitare problemi di "*overfitting*" (comportamento che si verifica quando il modello impara troppo dai dati usati per l'addestramento, compromettendo così le capacità di produrre predizioni affidabili su nuovi dati), si è usata la tecnica della cross validation[4].

4. Il modello considerato è un *Gradient Boosting*[5], una tecnica molto utilizzata per la sua precisione e velocità su dati particolarmente complessi e di grandi dimensioni.

Per realizzare quanto esposto si è utilizzata la piattaforma open-source KNIME[6] per la lettura dei dati, la loro trasformazione e l'interfaccia utente; il software R[7] con la funzione *ggplot* della libreria *ggplot2* per i grafici, la funzione *xgboost* della libreria *XGBoost* per la costruzione del modello e la funzione *predict_parts* della libreria *DALEX* basata sul metodo *SHAP* (*Shapley Additive exPlanations*) che consente di interpretare il contributo di ogni variabile alla previsione del modello.

I parametri del modello sono stati impostati in base ai risultati ottenuti in precedenza da un ottimizzatore sviluppato dall'autore di questo documento.

---

[3] L'obiettivo della *feature selection* è quello di identificare le variabili predittive più significative per un determinato problema, riducendo al contempo la complessità del modello e migliorando le prestazioni. https://proceedings.mlr.press/v10/salehi10a/salehi10a.pdf

[4] https://en.wikipedia.org/wiki/Cross-validation_(statistics)

[5] https://en.wikipedia.org/wiki/Gradient_boosting

[6] https://www.knime.com

[7] https://www.r-project.org

Valutazione

Per verificare la **capacità predittiva del modello sui nuovi dati**, è stato utilizzato il *test set* e costruita la matrice di confusione[8].

Si è scelto come valore di soglia ottimale, oltre il quale la probabilità viene classificata come diabete, il valore ottenuto dalla "*precision-recall curve (o PR curve)*" che minimizza la distanza tra il valore predittivo positivo ("*Pos Pred Value*" o "*Precision*") e la sensibilità ("*Sensitivity*" o "*Recall*"). Questo metodo risulta appropriato per tabelle sbilanciate come quella del presente studio.

In questo caso, il valore di soglia ottenuto dalla suddetta curva risulta essere 0,32.

	Prediction	
Reference	0	1
0	38.476	3.805
1	3.948	3.146

Accuracy	0.843
Sensitivity	0.45260
Specificity	0.90694
Pos Pred Value	0.44347
Neg Pred Value	0.91001
Prevalence	0.14078
Kappa	0.3565
F1	0.44799

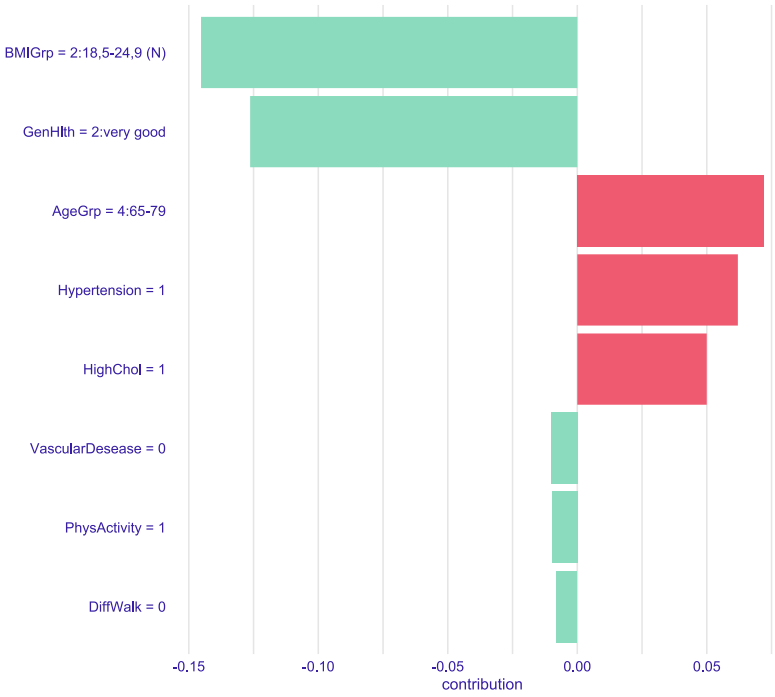
In base a questo valore si è ottenuta una matrice di confusione con un'accuratezza del **84,3%,** una sensitività del **45,2%**, una specificità del **90,7%** (ovvero 38.476 soggetti non diabetici predetti correttamente come non diabetici) e un valore predittivo positivo del **44,3%**.

Questa soglia consente di contenere il numero di **falsi negativi** (**3.948 soggetti diabetici non riconosciuti**) senza incrementare eccessivamente i falsi positivi (3.805 soggetti non diabetici classificati come tali)

[8] https://en.wikipedia.org/wiki/Confusion_matrix

Interpretazione

Per identificare le caratteristiche più determinanti nelle previsioni è stata utilizzata la tecnica della *feature importance*[9].



Questa metodologia assegna a ciascun attributo un coefficiente che misura l'importanza della variabile predittiva nella stima della variabile obiettivo.

Quanto più la barra è lunga, tanto maggiore è l'impatto della variabile sul risultato (se il coefficiente è positivo aumenta la probabilità di diabete; se negativo la riduce).

Come prevedibile dai risultati dell'analisi descrittiva, le principali caratteristiche che incrementano il rischio di diabete sono:

- ***Età compresa tra i 65 e 79 anni***
- ***Presenza di ipertensione***
- ***Colesterolo alto***

Al contrario, i fattori associati a una minore probabilità di sviluppare diabete sono:

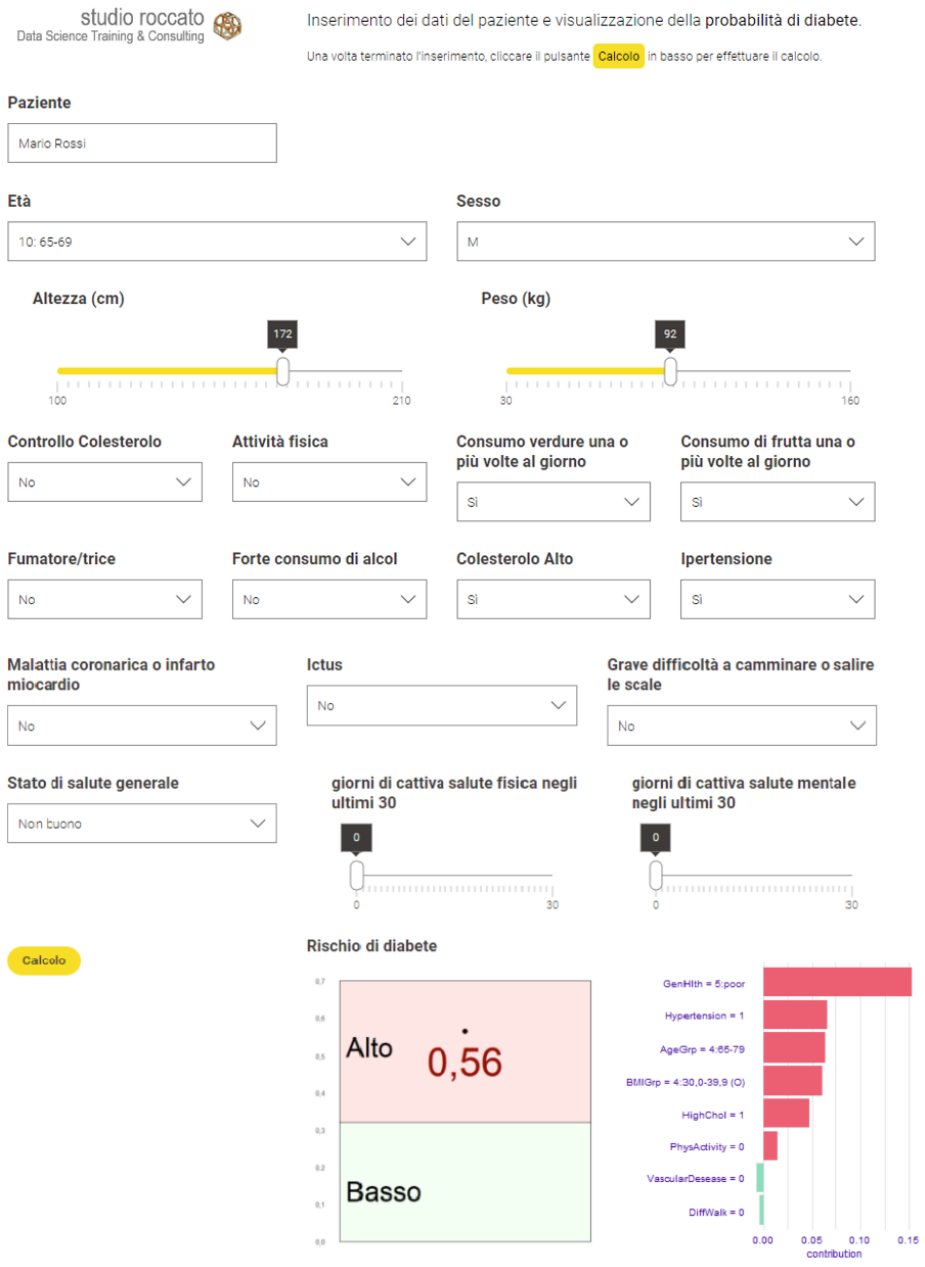
- ***BMI nella classe "normale" (18,5–24,9)***
- ***Stato di salute "molto buono"***
- ***Assenza di malattie vascolari***
- ***Pratica di attività fisica***
- ***Assenza di difficoltà nel camminare***

Questi risultati sottolineano il ruolo fondamentale di uno **stile di vita corretto nella prevenzione del diabete**.

[9] https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

Interfaccia grafica

È stata progettata un'interfaccia che, grazie alla sua semplicità e immediatezza, consente all'utente di inserire i propri dati (coerenti con le variabili del modello) per ottenere in tempo reale la probabilità di sviluppare diabete.



L'immagine a fianco illustra un caso d'uso: l'inserimento dei dati di un paziente e la successiva visualizzazione immediata dei risultati.

Nella parte inferiore i grafici mostrano, come output, il rischio di diabete del soggetto e il peso delle variabili che hanno influito nella previsione.

In questo modo, l'interfaccia fornisce al medico un supporto decisionale per intervenire su parametri specifici, promuovendo un approccio condiviso e consapevole alla prevenzione con il paziente.

Conclusioni

In questo studio è stato affrontato il tema della **diagnosi precoce del diabete di tipo 2** attraverso l'applicazione di tecniche di machine learning. L'obiettivo era sviluppare un modello predittivo in grado di **supportare il medico nell'individuazione dei fattori di rischio e nel rafforzamento delle strategie preventive**.

Il modello, basato su *Gradient Boosting*, ha mostrato prestazioni incoraggianti: è stato in grado di individuare le variabili più rilevanti per la previsione e di restituire risultati interpretabili grazie all'analisi della loro importanza.

La progettazione di un'**interfaccia intuitiva** ha reso lo strumento immediatamente **fruibile sia dai clinici sia dai pazienti**, sottolineandone la potenziale utilità pratica.

Pur **senza sostituire la diagnosi medica**, il modello si configura come un valido ausilio per **sensibilizzare il paziente e supportare il processo decisionale** clinico.

Limiti dello studio

I risultati ottenuti devono essere interpretati con cautela, in quanto lo studio presenta alcune limitazioni. In primo luogo, l'utilizzo di **dati autoriportati provenienti da questionari telefonici** espone il modello al **rischio di *bias* di risposta e di incompletezza delle informazioni**.

Inoltre, la **mancanza di variabili cliniche specifiche**, come i valori dei trigliceridi e la distribuzione del grasso corporeo, **riduce la capacità predittiva complessiva**. Allo stesso modo, l'assenza di informazioni contestuali (ad esempio l'area di residenza, utile per valutare l'influenza dell'ambiente) e di parametri fisiologici misurati direttamente rappresenta un'ulteriore criticità.

Infine, il fatto che il modello sia stato sviluppato su dati statunitensi ne limita la trasferibilità immediata al sistema sanitario italiano, caratterizzato da differenze demografiche, socio-economiche e culturali significative.

Sviluppi futuri

Alla luce di queste considerazioni, i futuri sviluppi dovranno concentrarsi sul **perfezionamento del modello e sul rafforzamento della sua validità esterna**.

Un passo fondamentale sarà l'**integrazione di dati clinici strutturati riferiti al contesto italiano**, così da aumentarne affidabilità e rilevanza per la pratica sanitaria nazionale. Sarà inoltre necessario testare il modello su campioni più ampi e diversificati, per verificarne la robustezza e l'applicabilità in scenari differenti.

Infine, un **ulteriore valore** potrà derivare dalla **collaborazione diretta con medici specialisti,** che consentirà di affinare **la selezione delle variabili cliniche, validare le ipotesi formulate e garantire un utilizzo appropriato dello strumento** come supporto alle decisioni mediche.

Riferimenti

Liu et al. (2024). *Use of Machine Learning to Predict the Incidence of Type 2 Diabetes Among Relatively Healthy Adults: A 10-Year Longitudinal Study in Taiwan*. https://pubmed.ncbi.nlm.nih.gov/39795600/

Zhang et al. (2025), *Development of a 5-Year Risk Prediction Model for Transition From Prediabetes to Diabetes Using Machine Learning: Retrospective Cohort Study*. https://www.jmir.org/2025/1/e73190

studio roccato
Data Science Training & Consulting

alfredo.roccato(at)fastwebnet.it

www.alfredoroccato.it