Implementazione di un modello di machine learning per la classificazione precoce del diabete mellito di tipo 2

Introduzione

Il diabete è una grave malattia cronica in cui gli individui perdono la capacità di regolare efficacemente i livelli di glucosio nel sangue, che può portare a una riduzione della qualità e dell'aspettativa di vita. Sebbene non esista una cura per il diabete, strategie come perdere peso, mangiare sano, essere attivi e ricevere cure mediche possono contribuire a mitigare i danni di questa malattia in molti pazienti.

La diagnosi precoce può portare a cambiamenti nello stile di vita e a trattamenti più efficaci, rendendo i modelli predittivi per il rischio del diabete importanti strumenti per il pubblico, i medici e i dirigenti sanitari.

Lo scopo di questo studio è l'implementazione di un efficace algoritmo d'identificazione del rischio di diabete del singolo paziente, da utilizzarsi per supportare l'analisi medica di diabete in base alle informazioni fornite dal paziente stesso.

Lo studio ha portato all'elaborazione di un modello in grado di prevedere con un'elevata accuratezza il diabete nei pazienti sulla base d'informazioni ricevute direttamente da questi, in modo tale da consentire al medico di riconoscere su quali parametri e attributi agire e sensibilizzare il malato in merito.

L'utilizzo di questo modello a cura del medico e/o del paziente, attraverso un'interfaccia semplificata per l'inserimento dei dati, fornisce una visualizzazione immediata della probabilità di una persona di essere diabetica.

Descrizione dei dati

Per la costruzione e la validazione del modello si è utilizzato il dataset[1] risultante 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 prima, *Diabetes* (presenza o meno di diabete), è la variabile obiettivo che si cerca di prevedere in funzione delle restanti colonne che vengono utilizzate come variabili di input (o predittori).

- ***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

[1] Trattandosi di risposte fornite dai partecipanti al sondaggio, non c'è stato modo di controllare la veridicità della maggior parte delle informazioni.

- ***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.
- ***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*** salute in generale (scala 1-5): 1 = eccellente, 2 = molto buona, 3 = buona, 4 = discreta, 5 = non buona.
- ***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 si è ritenuto di eliminare 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, come pure le righe in cui il valore di *BMI* è risultato inferiore a 15 oppure superiore a 50[2], portando così la dimensione finale della tabella a 246.875 righe.

[2] Corrispondenti rispettivamente a "*severe malnutrition*" e "superobese".

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);*

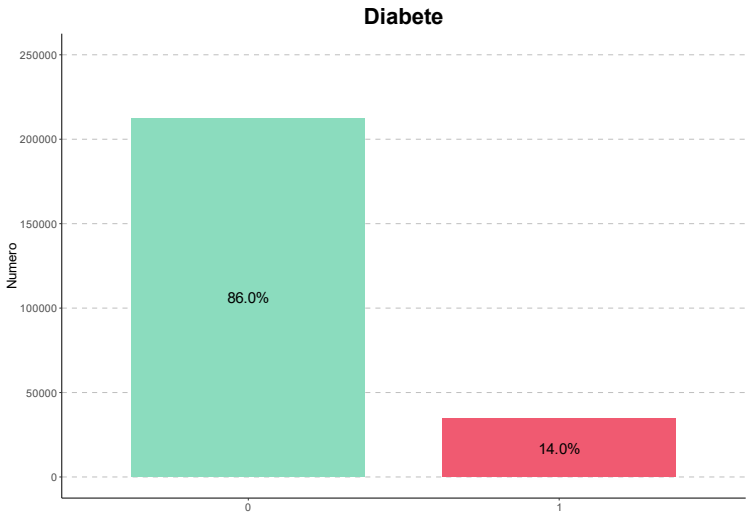
- classificazione della colonna *Age* in cinque gruppi:
  <35, 35-49, 50-64, 65-79, ≥80;

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

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

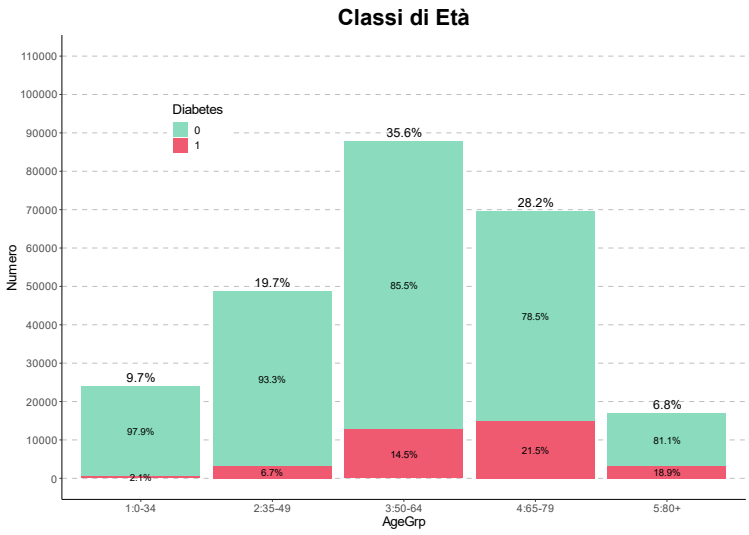
- le colonne *PhysHlth* e *MentHlth* (giorni di stato fisico e mentale negativi) sono state classificate in 2 gruppi: 0=buono, 1=non buono.

Esplorazione dei dati



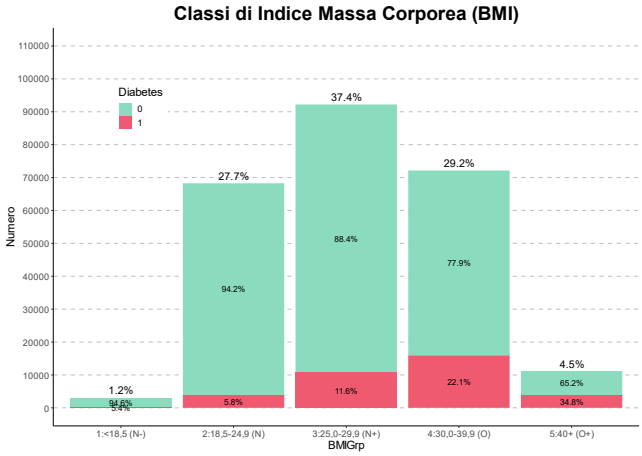
Il diagramma a barre rappresentato a sinistra mostra la prevalenza di diabetici (14,0%) e dei non diabetici (86,0%).

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

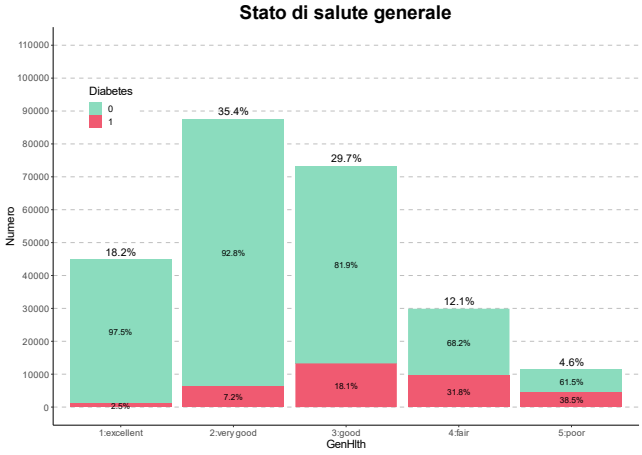


I diagrammi a barre sovrapposte rappresentano ogni categoria con una barra la cui altezza è proporzionale alla sua frequenza, mentre i segmenti all'interno di ogni barra corrispondono alla prevalenza di soggetti diabetici e no.

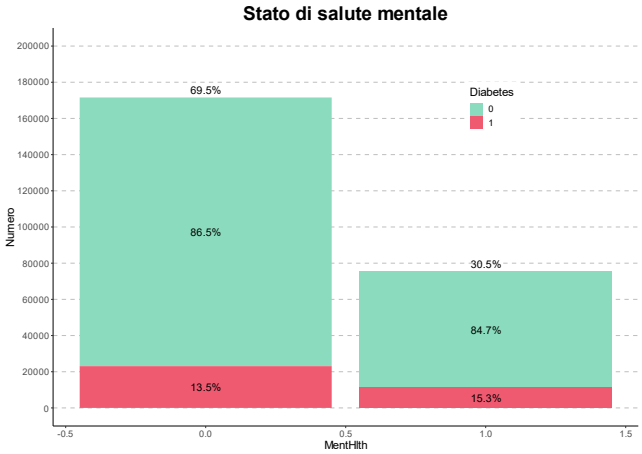
Nel diagramma delle classi di età la prevalenza di diabetici si verifica nella fascia 65-79 anni (21,5%), a seguire in quella degli ultraottantenni (18,9%).



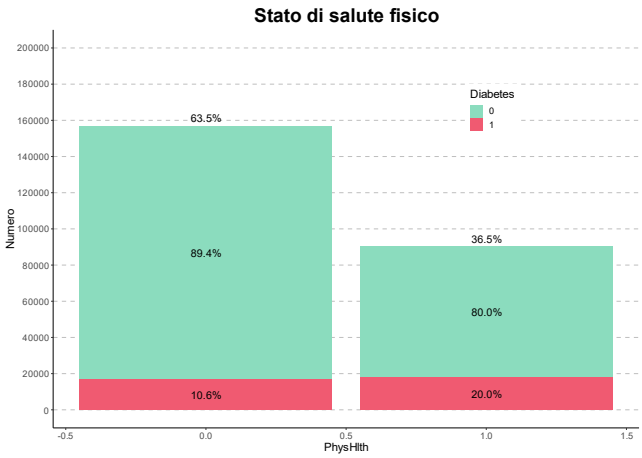
Le classi di BMI che hanno prevalenza di diabetici sono quella di “obesità grave” (34,8%), seguita da quella di “obesità” (22,1%) e da quella di “sovrappeso” (11,6%).

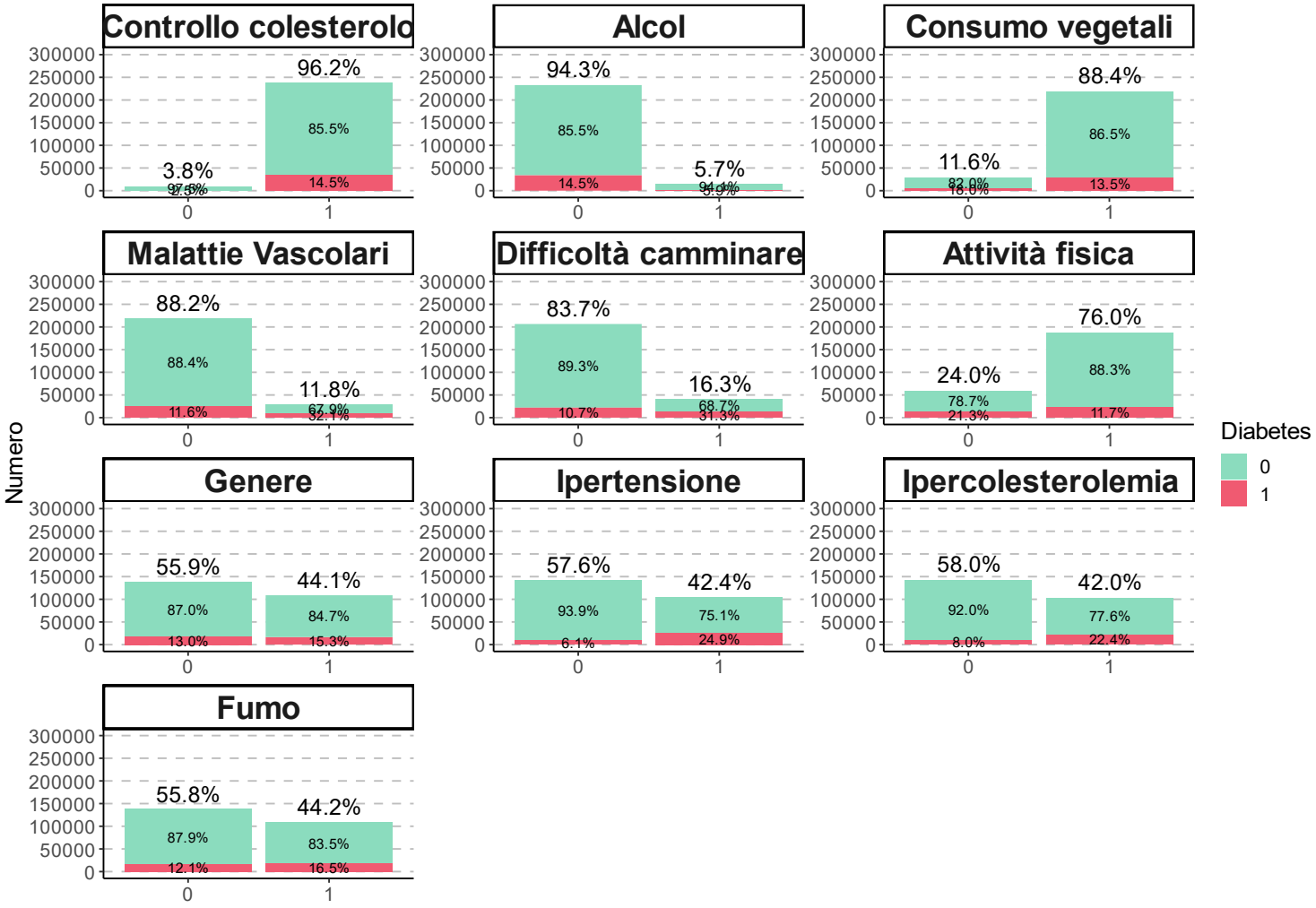


Anche lo stato di salute generale rivela una prevalenza di diabetici tra coloro che lo dichiarano come “non buono” (38,5%) e da quelli che lo dichiarano “discreto” (31,8%).



La prevalenza di 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, per le variabili numeriche, conferma quanto si poteva dedurre dai diagrammi di distribuzione.

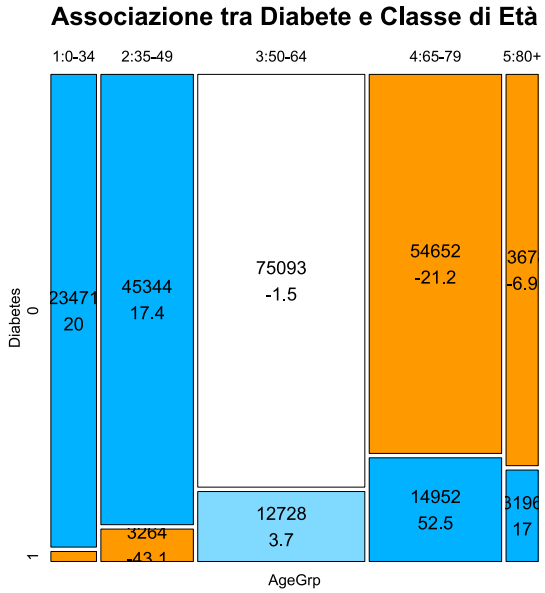
Il diabete sembra, infatti, (debolmente) correlato positivamente con:

- ***l'ipertensione (0,27);***
- ***il colesterolo alto (0,21);***
- ***la difficoltà nel camminare (0,22);***
- ***i problemi di tipo vascolare (0,19).***

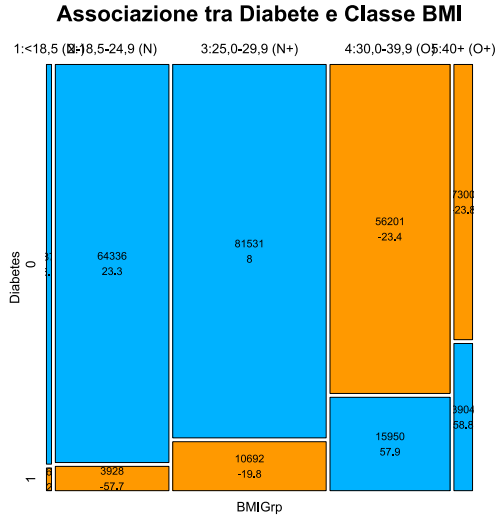
Si nota anche una correlazione (debolmente) negativa con la pratica di attività fisica (-0,12).

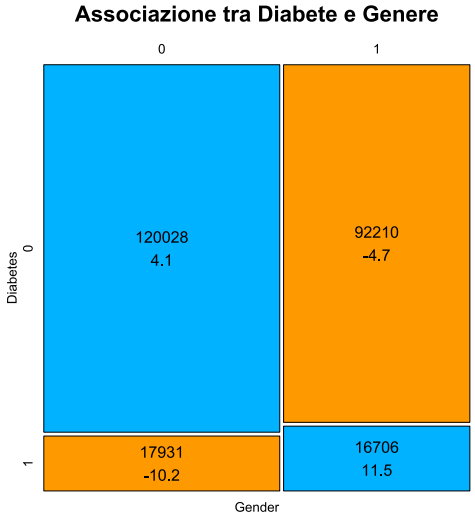
Un diagramma a mosaico permette di rappresentare graficamente l'associazione tra due o più variabili categoriche dove:

- 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[3] 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.



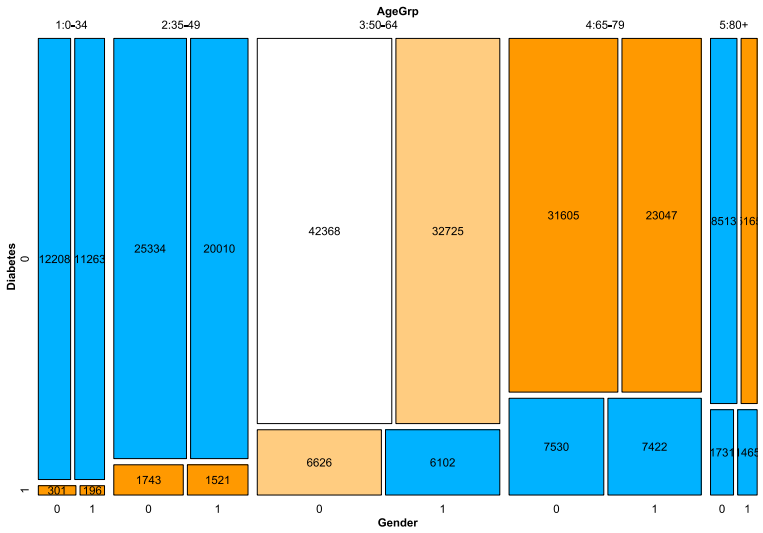
Ad esempio, considerando la variabile *AgeGrp*: si nota una forte associazione positiva tra il diabete e la classe di età 65-79, seguita da quella 80+ e poi, ma più debole, quella 50-64. Viceversa, c'è una forte associazione negativa tra il diabete e l'età inferiore ai 50 anni.





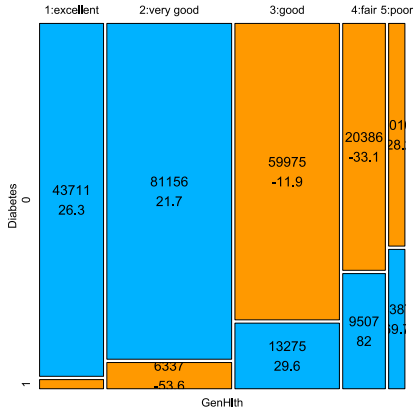
[3] I residui sono le differenze (positive o negative) tra le frequenze osservate e le frequenze attese sotto l'ipotesi di indipendenza statistica.

Associazione tra Diabete, Classe di età e Genere

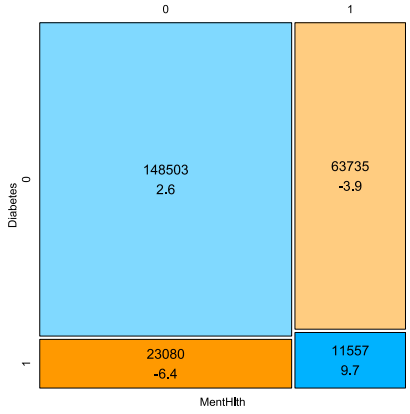


Da questo diagramma a tre dimensioni si evince che, a differenza di quello precedente riferito al solo genere, la prevalenza dei diabetici si ha, nella fascia di età dai 65 anni in poi, tra le donne invece che tra gli uomini.

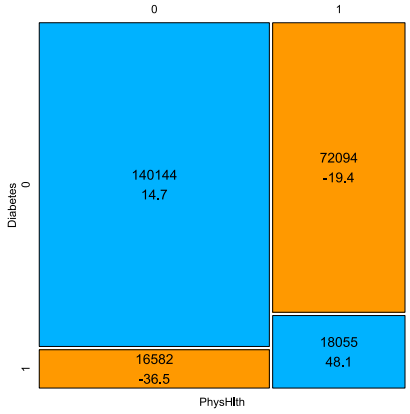
Associazione tra Diabete e Stato salute generale



Associazione tra Diabete e Stato salute mentale



Associazione tra Diabete e Stato salute fisico



In sintesi, si può notare un'associazione positiva molto forte tra il diabete e:

- ***la classe di età "65-79" e "80+" anni;***
- ***la classe "obesità" e "obesità grave";***
- ***lo stato di salute generale "discreto", "non buono" e "buono";***
- ***in leggera prevalenza, il sesso maschile.***

Modeling

La tabella dei dati è stata suddivisa, su base randomica e stratificata sulla variabile obiettivo (ovvero mantenendo le stesse proporzioni di diabetici), in 2 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*"[4] 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*, *VascularDesease* e *PhysActivity*.

2. Essendo la tabella sbilanciata rispetto alla variabile obiettivo (il numero di diabetici è rappresentato da un numero molto inferiore rispetto al numero di non diabetici), si è adottata una tecnica di sovra-campionamento[5] 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[6].

4. Il modello considerato è un Gradient Boosting[7], 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[8] per la lettura dei dati, la loro trasformazione e l'interfaccia utente; il software R[9] 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*" per la sua interpretazione.

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

---

[4] 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

[5] Serve a evitare che il modello venga influenzato in modo sproporzionato dalla classe maggioritaria, migliorando al contempo la capacità del modello di riconoscere correttamente la classe minoritaria. https://www.jair.org/index.php/jair/article/view/10302/24590

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

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

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

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

Valutazione

Per verificare la capacità di produrre affidabili predizioni su nuovi dati, il modello è stato utilizzato sui dati del "test set" e, dai risultati ottenuti, si è costruita la matrice di confusione[10].

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

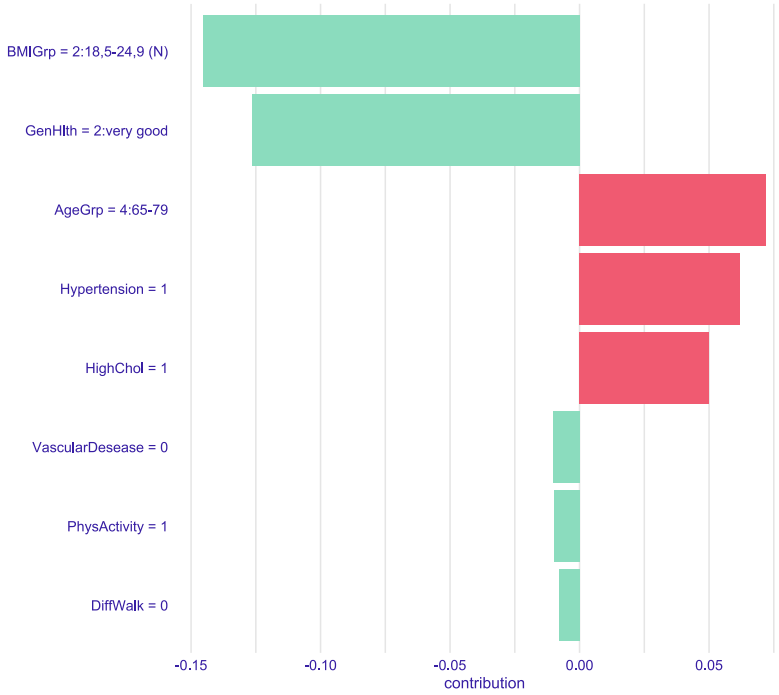
In base a questo valore si è ottenuta una matrice di confusione con un'accuratezza del **84,3%,** una sensitività del **45,2%**, una specificità del **90,7%** (ovvero 38.476 non diabetici predetti correttamente come "non diabetici") e un valore predittivo positivo del **44,3%**.

Questo valore di soglia permette altresì di avere un **numero basso di falsi negativi** (3.948 casi di soggetti non classificati come diabetici quando lo sono), senza per questo aumentare il numero dei falsi positivi (3.805 casi di soggetti classificati come diabetici quando non lo sono).

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

Interpretazione

Per comprendere quali caratteristiche siano più importanti per la previsione si è utilizzata la tecnica *feature importance*[11].



Questa tecnica assegna, attraverso l'algoritmo utilizzato, un coefficiente agli attributi delle variabili predittive in base alla loro importanza nel predire la variabile obiettivo.

Quanto più le barre sono grandi, tanto più quegli attributi hanno un peso maggiore nel determinare il risultato (se il coefficiente è positivo, contribuiscono a spostare il risultato verso valori positivi, e viceversa).

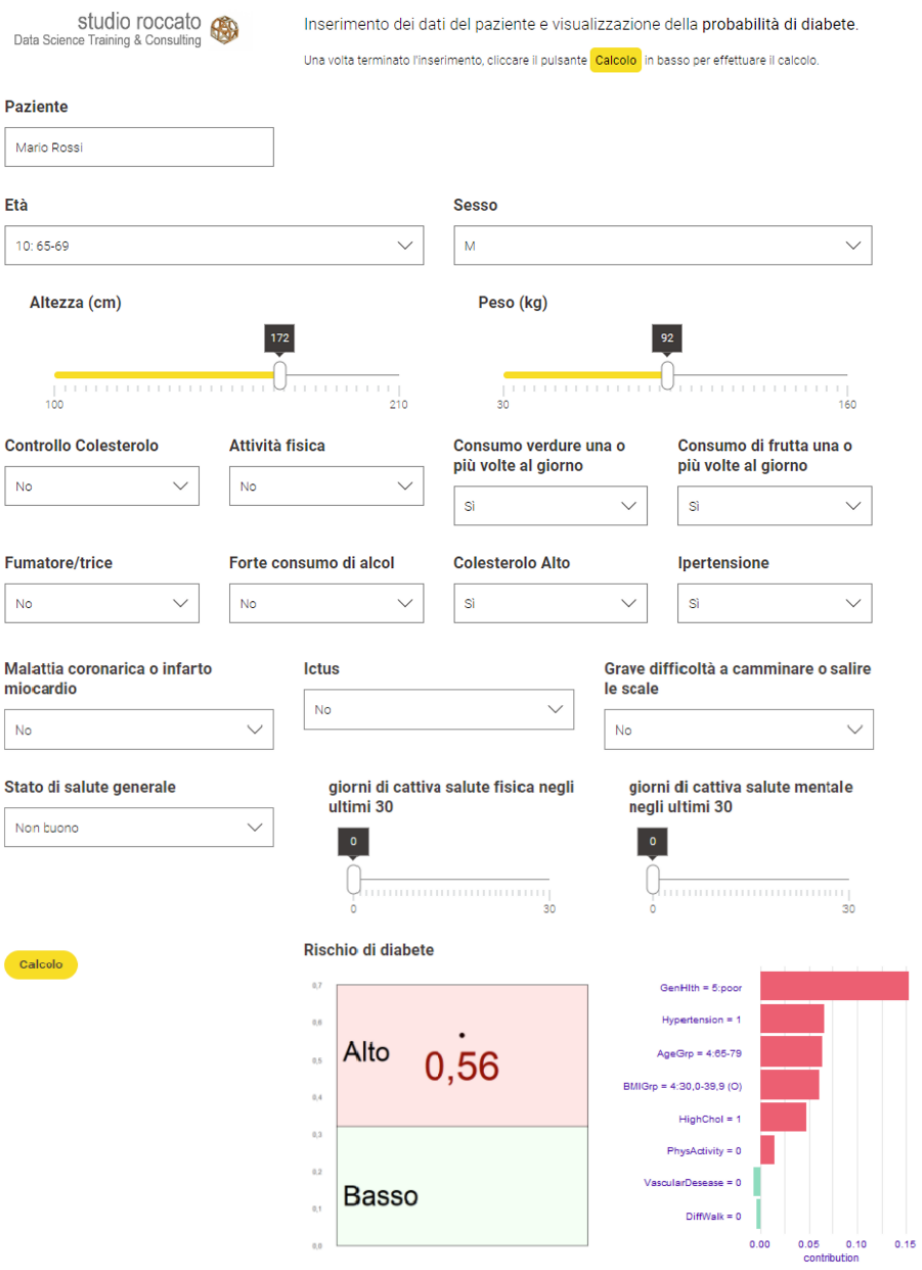
Come prevedibile, dalle analisi svolte precedentemente, le principali caratteristiche che secondo il modello rendono un individuo più a rischio di diabete sono:

- ***la classe di età 65-79;***
- ***l'ipertensione;***
- ***l'ipercolesterolemia.***

mentre la classe BMI "*normale*" (18,5-24,9), lo stato di salute "*molto buono*", il non aver avuto problemi di tipo vascolare, il fare attività fisica e non aver difficoltà nel camminare sono fattori che abbassano il rischio di diabete.

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

Interfaccia grafica



Attraverso il prospetto mostrato a fianco il singolo utente può inserire i propri dati (in conformità a quelli utilizzati dal modello) per consentire la determinazione della propria probabilità di sviluppare il diabete o meno.

Questa funzione è stata progettata per essere semplice da usare, intuitiva e rapida.

Nella figura a fianco sono riportati i dati di un ipotetico paziente, i cui risultati sono immediatamente visualizzabili.

Nella parte inferiore i grafici mostrano, come output, il rischio di diabete del soggetto e il peso delle singole variabili.

Appare, pertanto, immediatamente chiaro su quali parametri/elementi il medico possa agire in sinergia col paziente.

Conclusioni

In questo studio si è voluto affrontare il problema della classificazione della diagnosi precoce del diabete mellito di tipo 2. L'obiettivo principale è l'identificazione di un modello in grado di prevedere con un'elevata accuratezza il diabete nei pazienti sulla base di informazioni ricevute direttamente da questi in modo tale da consentire al medico di riconoscere su quali parametri e attributi agire e sensibilizzare il paziente in merito.

Il Gradient Boosting si è dimostrato un ottimo modello dalle prestazioni ottenute sul "test set": la capacità di fornire predizioni accurate e affidabili è la condizione essenziale per sviluppare uno strumento efficace di prevenzione della malattia.

Sviluppi futuri

Riconoscendo un possibile spazio per miglioramenti, il lavoro futuro potrebbe concentrarsi sul perfezionamento del modello per aumentarne le prestazioni, esplorando tecniche di modellazione alternative, incorporando ulteriori fonti di dati, conducendo ulteriori test e convalide su altri gruppi di pazienti per garantire la coerenza delle prestazioni.

Altra nota degna di merito è la necessità di reperire dati non necessariamente provenienti da interviste telefoniche, ma archivi medici possibilmente della realtà italiana.

Risulta infine di fondamentale importanza integrare le competenze fornite da medici specialisti per ulteriori affinamenti dello strumento e per la validazione delle ipotesi effettuate.

studio roccato
Data Science Training & Consulting

alfredo.roccato(at)fastwebnet.it

www.alfredoroccato.it