

Analisi predittiva del diabete

Salvatore Megna

7 Aprile 2025

INTRODUZIONE

Negli ultimi anni, l'identificazione precoce del diabete è diventata una priorità per il sistema sanitario, considerando l'elevato impatto clinico, economico e sociale della malattia. Grazie alla disponibilità crescente di dati clinici e biochimici, è oggi possibile sviluppare modelli predittivi in grado di supportare il personale medico nelle decisioni diagnostiche.

In questo progetto, viene analizzato un dataset contenente informazioni su 7897 pazienti, tra cui: valori relativi a glucosio, BMI, età, pressione arteriosa, colesterolo, storia familiare e altri fattori di rischio, tutti riportati nella tabella 1. L'obiettivo principale dell'analisi è confrontare due modelli di classificazione supervisionata : *Random Forest* e *Gradient Boosting*, utilizzati per prevedere l'insorgenza del diabete.

Variabile	Descrizione
Age	Età dell'individuo espressa in anni.
BMI	Indice di massa corporea (kg/m^2), calcolato in base a peso e altezza.
Glucose	Livello di glucosio nel sangue (mg/dL).
BloodPressure	Pressione arteriosa sistolica (mmHg).
HbA1c	Percentuale di emoglobina glicata, media della glicemia negli ultimi 2-3 mesi (%).
LDL	Livello di colesterolo LDL, detto anche "colesterolo cattivo" (mg/dL).
HDL	Livello di colesterolo HDL, detto anche "colesterolo buono" (mg/dL).
Triglycerides	Concentrazione di trigliceridi nel sangue (mg/dL).
WaistCircumference	Circonferenza vita, espressa in centimetri.
HipCircumference	Circonferenza dei fianchi, espressa in centimetri.
WHR	Rapporto vita/fianchi (Waist-to-Hip Ratio).
FamilyHistory	Presenza di familiarità per il diabete (1 = sì, 0 = no).
DietType	Tipo di alimentazione: 0 = squilibrata, 1 = bilanciata, 2 = vegetariana/vegana, 3 = sconosciuta.
Hypertension	Presenza di ipertensione (1 = sì, 0 = no).
MedicationUse	Uso di farmaci specifici per il diabete (1 = sì, 0 = no).
Outcome	Esito diagnostico: 1 = presenza di diabete, 0 = assenza di diabete.

Tabella 1: Desrizione delle variabili del dataset

ANALISI ESPLORATIVA

L'analisi esplorativa ha l'obiettivo di comprendere le caratteristiche generali del dataset, individuare eventuali anomalie nei dati e osservare la relazione tra le variabili esplicative e la variabile target. Come si osserva in Figura 1, la distribuzione della variabile Outcome (Diabete) mostra un forte sbilanciamento tra le classi: oltre l'80% delle osservazioni appartiene alla categoria "Non Diabetico". Questa asimmetria costituisce un elemento critico che dovrà essere gestito nelle fasi successive di modellazione, per evitare che i modelli predittivi risultino sbilanciati e poco sensibili ai soggetti diabetici.

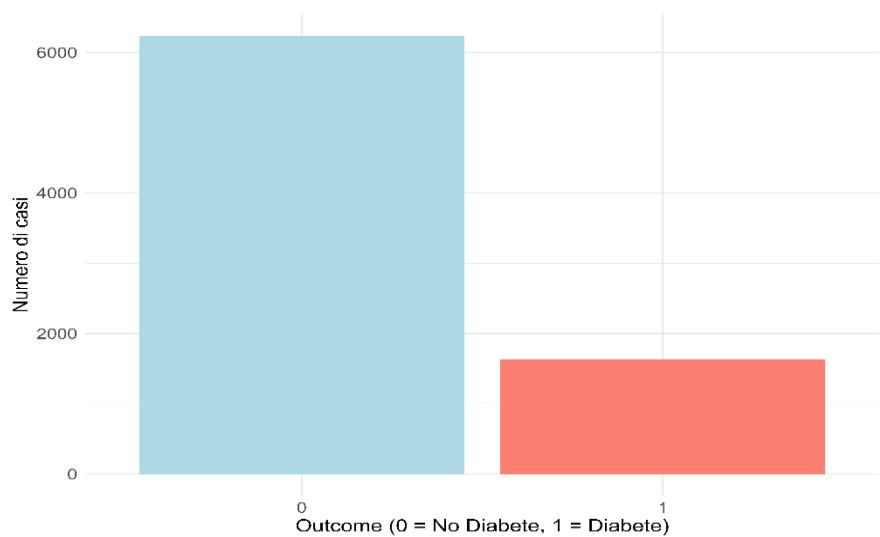


Figura 1: Distribuzione della variabile Outcome (Diabete)

L'analisi delle variabili numeriche, rappresentata in Figura 2, evidenzia pattern interessanti: in media, i soggetti diabetici presentano valori più elevati di glicemia (Glucose) ed emoglobina glicata (HbA1c), due indicatori ben noti nella diagnosi e nel monitoraggio della malattia. Anche le misure antropometriche come l'indice di massa corporea (BMI), la circonferenza vita e il rapporto vita/fianchi (WHR) risultano più alte nei soggetti diabetici, suggerendo un legame con la distribuzione del grasso corporeo. L'età media dei diabetici è maggiore rispetto ai non diabetici, confermando che il rischio aumenta con l'età. Variabili come HDL e LDL, invece, mostrano differenze meno evidenti.

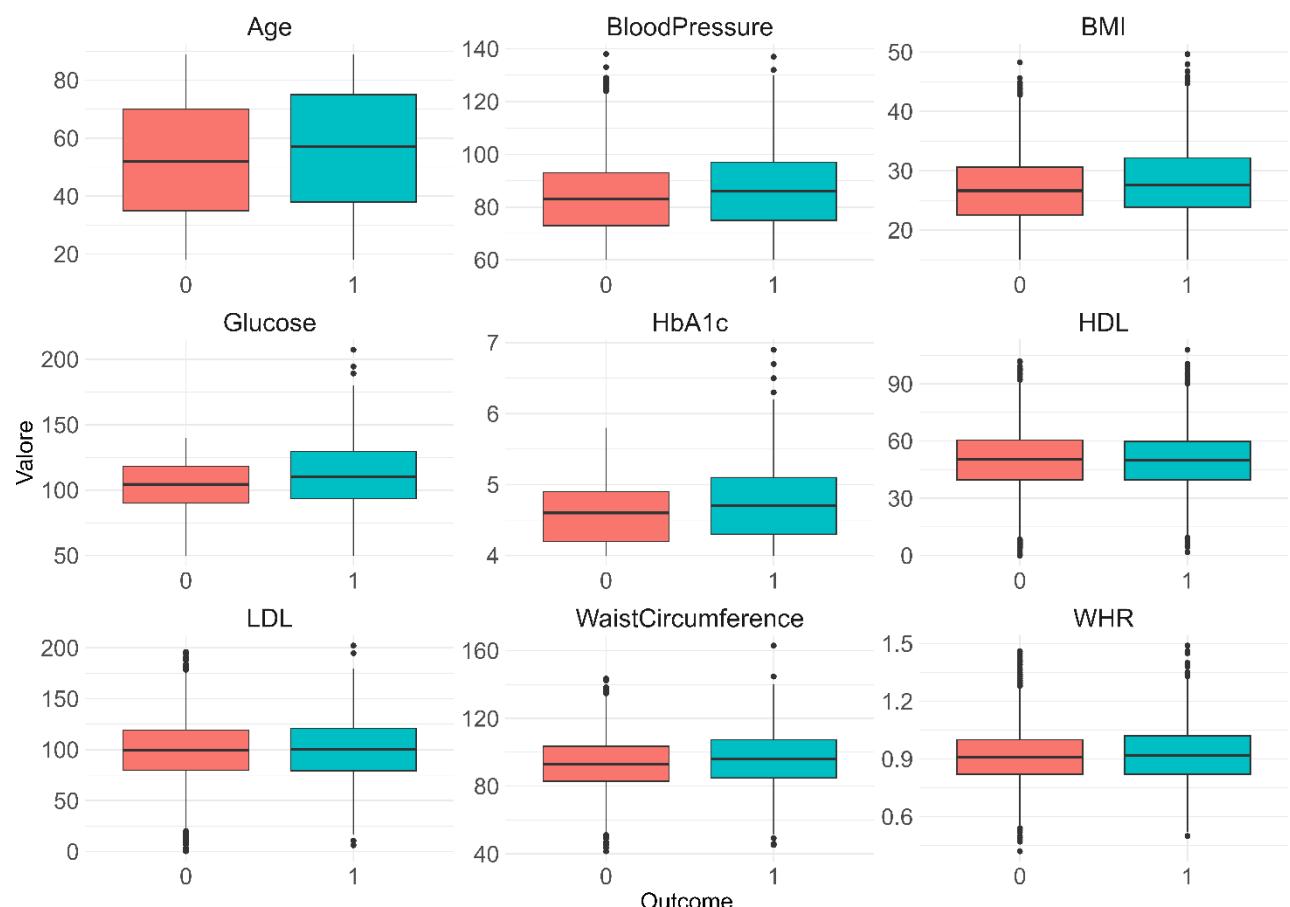


Figura 2: Boxplot delle variabili quantitative condizionate all'outcome

Per quanto riguarda le variabili categoriche, i grafici riportati in Figura 3 evidenziano diverse associazioni con la condizione diabetica. In particolare, la variabile FamilyHistory mostra una chiara relazione: i soggetti con una storia familiare di diabete presentano una prevalenza decisamente maggiore di casi positivi, di questo dovremo tener conto in fase di modellizzazione per evitare che i modelli apprendano troppo semplicemente a quale gruppo appartengono i pazienti. La variabile DietType, sebbene presenti una maggioranza di soggetti non diabetici in tutte le categorie, mostra una proporzione più alta di diabetici nella categoria "Sconosciuta", un dato che potrebbe riflettere una maggiore incidenza di dati mancanti tra soggetti a rischio. La variabile Hypertension evidenzia una netta associazione con la condizione diabetica: i soggetti ipertesi mostrano una prevalenza molto maggiore di diabete, confermando il legame tra rischio cardiovascolare e insorgenza della malattia. Infine, anche la variabile MedicationUse mostra una prevalenza leggermente più alta di diabete nei soggetti che assumono farmaci, anche se la differenza è meno marcata rispetto ad altre variabili. Questo dato potrebbe riflettere l'uso di farmaci antidiabetici o di trattamenti per patologie correlate.

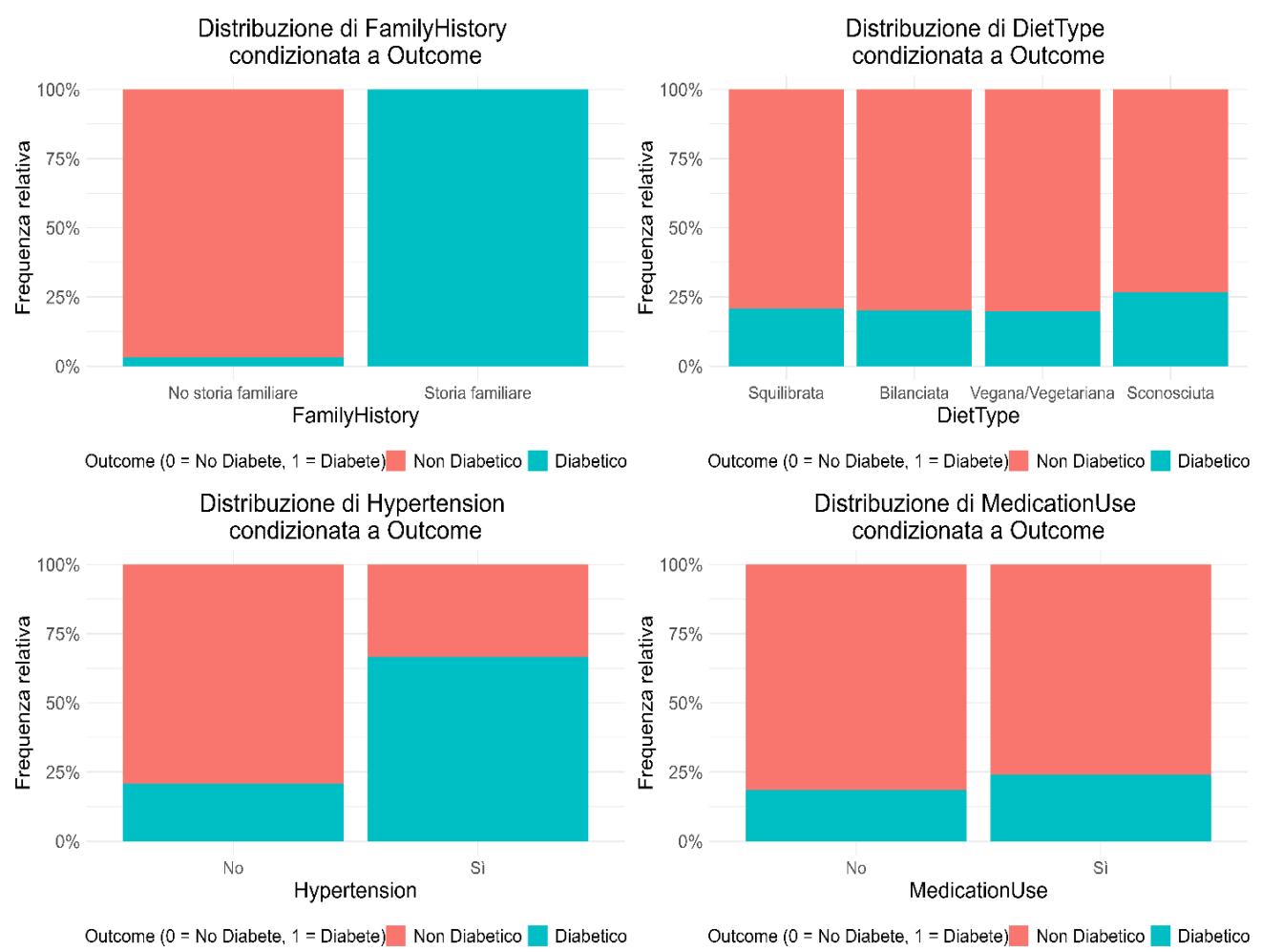


Figura 3: distribuzione delle variabili categoriali condizionate all'Outcome.

Dal punto di vista della qualità dei dati, sono state riscontrate alcune anomalie: valori negativi in HDL e LDL, e valori estremamente elevati in BMI superiori a 120. Considerando che tali osservazioni rappresentano meno dello 0.3% del dataset (20 osservazioni su circa 7900), si è scelto di rimuoverle per garantire la robustezza dell'analisi, senza intaccare la significatività statistica complessiva.

La matrice delle correlazioni riportata in Figura 4 mostra le relazioni lineari tra le variabili quantitative del dataset. Tra i valori più significativi si nota una correlazione molto alta tra Glucose e HbA1c ($r = 0.80$), due indicatori chiave del controllo glicemico che risultano naturalmente collegati, essendo il primo una misura puntuale della glicemia e il secondo un indicatore della sua media nel tempo.. Anche le variabili antropometriche mostrano forti associazioni: WaistCircumference e BMI sono chiaramente correlate con $r = 0.76$, mentre HipCircumference e BMI raggiungono $r = 0.66$. Questi dati confermano la coerenza interna tra le misure relative alla composizione corporea. La pressione arteriosa (BloodPressure) mostra correlazioni moderate con più variabili, tra cui BMI ($r = 0.60$), Glucose ($r = 0.54$) e WaistCircumference ($r = 0.53$). Questo suggerisce una possibile relazione tra obesità, ipertensione e alterazioni metaboliche, coerente con quanto riscontrato nella letteratura clinica. Al contrario, alcune variabili lipidiche come LDL, HDL e Triglycerides risultano debolmente correlate con le altre. In particolare, i valori di correlazione sono prossimi a zero, indicando che queste variabili potrebbero fornire un contributo informativo indipendente nella fase di modellazione. Nel complesso, l'analisi esplorativa ha permesso di individuare variabili potenzialmente rilevanti nella previsione della condizione diabetica, gettando le basi per una modellizzazione predittiva consapevole e informata.

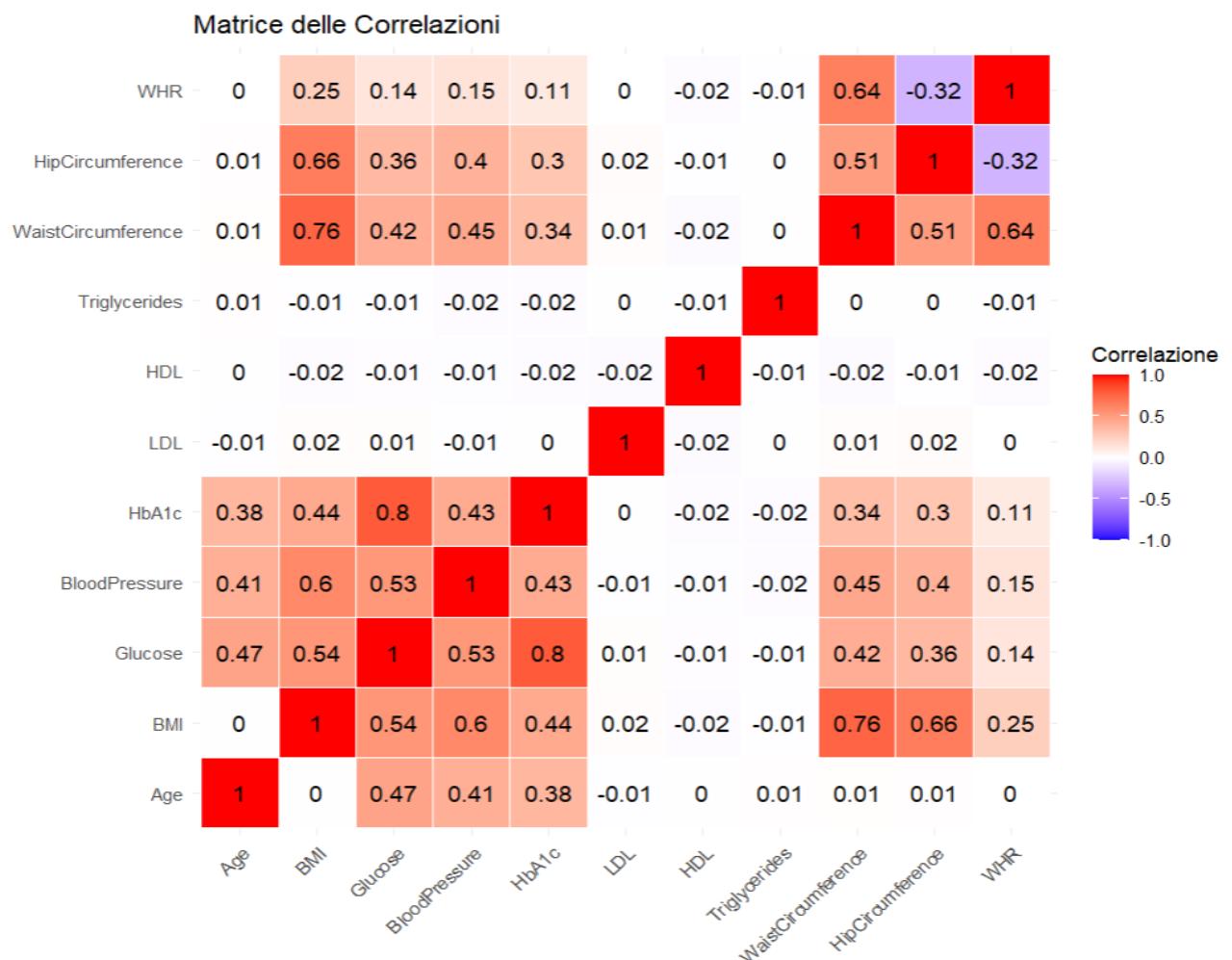


Figura 4: matrice delle correlazioni per variabili quantitative.

ANALISI DEI MODELLI

Per valutare correttamente le prestazioni dei modelli predittivi, il dataset è stato suddiviso in due sottoinsiemi: il 75% dei dati è stato utilizzato per l'addestramento (training set), mentre il 25% restante è stato riservato alla fase di valutazione finale (test set). Inizialmente, la modellazione è stata avviata

utilizzando tutte le variabili a disposizione, con l'obiettivo di prevedere l'outcome (presenza o assenza di diabete). Tuttavia, fin dalle prime fasi è emerso un aspetto critico: la presenza della variabile FamilyHistory, indicante la familiarità con il diabete, condizionava fortemente le performance dei modelli. In particolare, i modelli di Random Forest raggiungevano valori di F1-score prossimi a 1 anche con un numero estremamente ridotto di alberi (10, 20 o 30), rendendo di fatto impossibile qualsiasi confronto sensato tra diverse configurazioni di iperparametri o tra algoritmi distinti.

Un'analisi esplorativa più approfondita ha confermato che la variabile FamilyHistory era praticamente in grado di predire da sola l'outcome: la totalità dei soggetti che presentavano familiarità risultava diabetica, mentre tra i soggetti che non presentavano familiarità, solo il 3,17% era affetta dalla malattia. Per evitare che i modelli apprendessero in modo banale questa relazione, si è deciso di escludere FamilyHistory dal training, in modo da rendere il compito predittivo più realistico e interessante da analizzare.

Dopo l'esclusione della variabile FamilyHistory, le performance dei modelli si sono ridimensionate, in particolare per quanto riguarda la sensibilità, ovvero la capacità di identificare correttamente i soggetti diabetici. In ambito medico, la sensibilità è una metrica particolarmente importante, poiché è preferibile incorrere in un falso positivo piuttosto che trascurare un paziente potenzialmente a rischio. Per questo motivo, si è scelto di variare la soglia di classificazione (threshold) per ciascun modello, al fine di individuare quella che massimizza l'F1-score, una metrica che bilancia in modo ottimale precisione e sensibilità.

Per quanto riguarda gli algoritmi utilizzati, sono stati implementati e confrontati Random Forest e Gradient Boosting (GBM).

Nel caso della Random Forest, si è agito su due iperparametri fondamentali:

- ntree: numero di alberi nella foresta (valori testati: 500, 1000, 1500)
- mtry: numero di variabili considerate a ogni split (2, 3, 4)

L'analisi delle metriche (Tabella: Metriche Random Forest) mostra che, al variare dei parametri e della soglia, i modelli raggiungono performance F1 tra 0.360 e 0.366, con sensibilità che arriva fino a 0.989 ma a costo di un calo importante in precisione.

Modello	Threshold	Accuracy	Sensitivity	Precision	F1
rf_500_2	0.15	0.313	0.897	0.230	0.366
rf_1500_2	0.15	0.297	0.913	0.228	0.365
rf_500_3	0.10	0.305	0.899	0.228	0.364
rf_1000_4	0.10	0.311	0.885	0.228	0.363
rf_500_4	0.10	0.319	0.874	0.229	0.362
rf_1000_2	0.10	0.227	0.989	0.221	0.362
rf_1500_4	0.10	0.309	0.883	0.227	0.361
rf_1500_3	0.10	0.295	0.899	0.226	0.361
rf_1000_3	0.10	0.293	0.897	0.225	0.360

Tabella 2: Iperparametri Random Forest

Nel caso del Gradient Boosting, sono stati testati diversi modelli combinando:

- interaction.depth: profondità massima degli alberi (2, 3, 4)

- shrinkage: tasso di apprendimento (0.001, 0.005, 0.01)

Questi due parametri regolano la complessità e l'apprendimento progressivo del modello: profondità maggiore consente di catturare interazioni più complesse, mentre un tasso di apprendimento più basso rende il modello più conservativo, ovvero meno propenso a correggere gli errori del modello precedente, evitando in questo modo un adattamento eccessivo ai dati di training e quindi riducendo il rischio di overfitting, ma richiede più alberi.

Il numero di alberi (n.trees) è stato inizialmente impostato a 1000. Per determinare il numero ottimale di alberi, si è utilizzata la funzione gbm.perf, basata sulla devianza di Bernoulli in cross-validation. Come visibile nel grafico (Figura 5), la curva della devianza si appiattisce intorno ai 500 alberi, indicando che l'aggiunta di ulteriori alberi, comporta una diminuzione marginale in termini di errore, tuttavia si è deciso di utilizzare 1000 alberi per aumentare le prestazioni del modello, evitando però di spingerci oltre aggiungendo costi computazionali superflui.

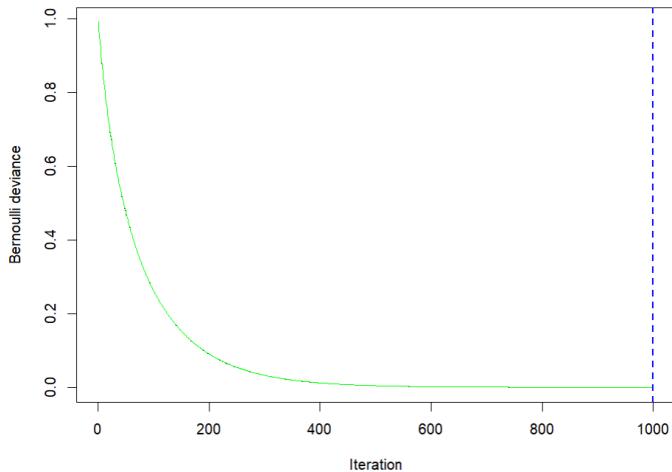


Figura 5: Grafico della devianza di Bernoulli al variare del numero di alberi

Anche nel caso del boosting, le performance si sono mantenute su valori di F1 simili a quelli ottenuti con la Random Forest (tra 0.362 e 0.373), con modelli che raggiungono sensibilità anche superiori a 0.96, ma ancora una volta a scapito della precisione.

Modello	Threshold	Accuracy	Sensitivity	Precision	F1
gbm_d3_sh001	0.15	0.358	0.865	0.238	0.373
gbm_d2_sh001	0.15	0.331	0.892	0.234	0.371
gbm_d4_sh001	0.15	0.379	0.828	0.239	0.371
gbm_d4_sh0005	0.15	0.301	0.924	0.231	0.369
gbm_d2_sh0005	0.15	0.258	0.968	0.226	0.366
gbm_d3_sh0005	0.15	0.280	0.931	0.226	0.364
gbm_d2_sh0001	0.10	0.221	1.000	0.221	0.362
gbm_d3_sh0001	0.10	0.221	1.000	0.221	0.362
gbm_d4_sh0001	0.10	0.221	1.000	0.221	0.362

Tabella 3: Iperparametri Gradient Boosting

Le tabelle finali evidenziano come entrambi gli algoritmi siano in grado di offrire buone performance, ma ciascuno con un proprio equilibrio tra sensibilità e precisione, aspetto fondamentale da considerare in fase di scelta del modello da adottare.

IMPORTANZA DELLE VARIABILI

Un aspetto fondamentale dell'analisi riguarda l'importanza attribuita alle variabili predittive dai due modelli. Nel caso del Random Forest, l'importanza risulta distribuita in modo relativamente uniforme tra diverse variabili: Glucose è al primo posto, ma anche LDL, Triglycerides, HDL, HipCircumference, WaistCircumference, BMI e persino Age contribuiscono in misura rilevante alla classificazione. Questo comportamento è coerente con la struttura dell'algoritmo: il Random Forest costruisce molti alberi in parallelo, ciascuno addestrato su un campione casuale di variabili. Questo approccio garantisce una certa equità nella selezione delle variabili e una maggiore robustezza alle relazioni spurie, distribuendo così il peso delle decisioni su più predittori.

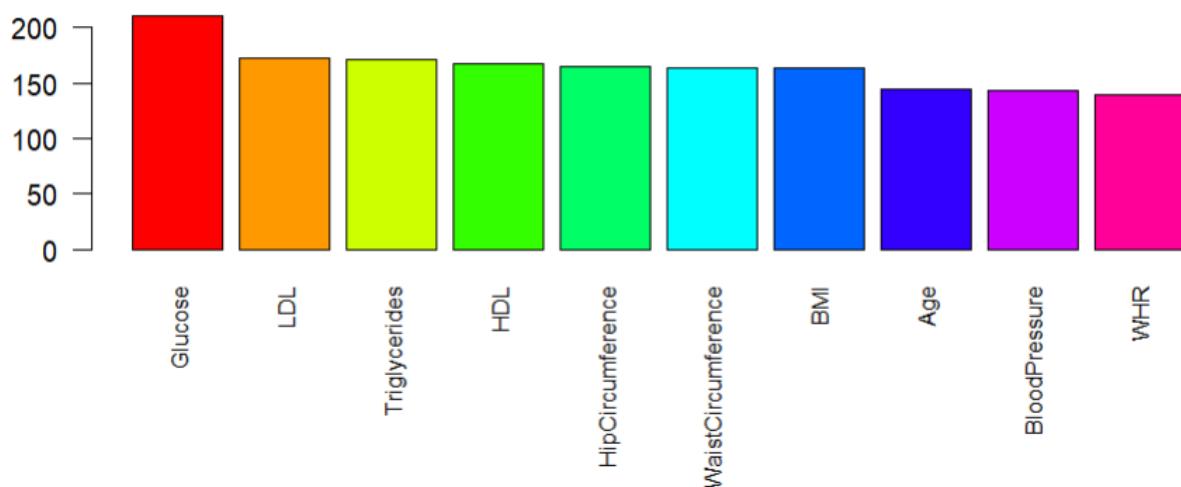


Figura 6: Importanza variabili del modello Random Forest migliore (rf_500_2)

Al contrario, il modello di Gradient Boosting (in particolare il modello gbm_d3_sh001) attribuisce un'importanza nettamente sbilanciata alla variabile Glucose, seguita da HipCircumference e Triglycerides con valori molto più bassi. Questo riflette la logica sequenziale dell'algoritmo boosting, che costruisce alberi in serie, dove ciascun albero si concentra sugli errori commessi dai precedenti. In questa dinamica, una variabile particolarmente efficace nel discriminare tra classi – come appunto Glucose – può essere selezionata ripetutamente, dominando così il processo di apprendimento. Inoltre, nel caso di variabili altamente correlate, come Glucose e HbA1c (correlazione pari a 0.80 secondo la matrice delle correlazioni), il boosting tende a preferirne una sola, riducendo l'importanza attribuita all'altra. Questo comportamento spiega la pressoché nulla rilevanza di HbA1c nel modello boosting, pur essendo ben nota dal punto di vista clinico la sua rilevanza.

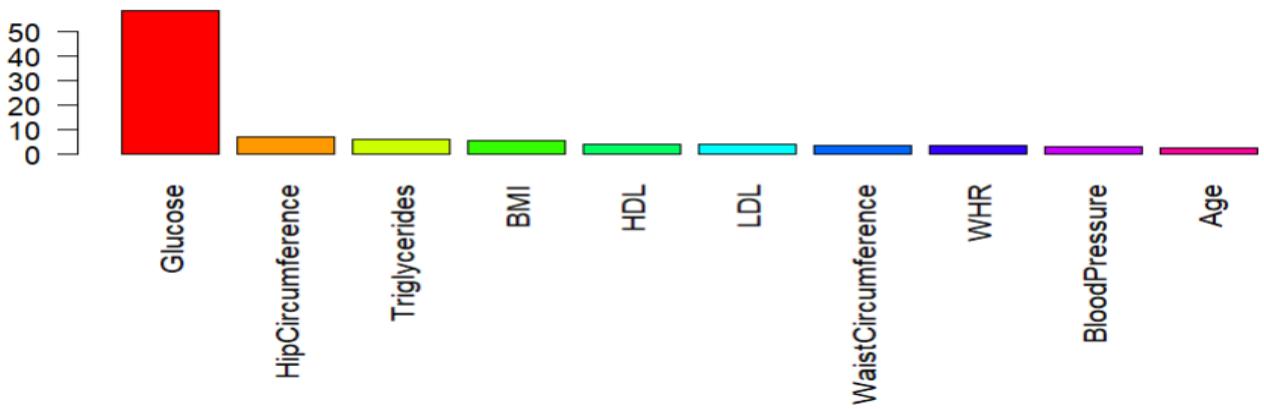


Figura 7: Importanza variabili del modello Gradient Boosting migliore (gbm_d3_sh001)

Nel complesso, entrambi i modelli confermano quanto emerso dall'analisi esplorativa: variabili come Glucose, BMI, circonferenze corporee e Triglycerides sono sistematicamente tra le più informative per la previsione dell'Outcome. Il Random Forest restituisce una visione più bilanciata, mentre il Gradient Boosting tende a focalizzarsi maggiormente su poche variabili molto forti, come Glucose.

CONCLUSIONI

L'analisi svolta ha evidenziato come, a partire da un dataset clinico strutturato, sia possibile sviluppare modelli predittivi efficaci per la classificazione del diabete. Dopo un'approfondita fase di pulizia ed esplorazione dei dati, si è proceduto alla costruzione di modelli supervisionati utilizzando due approcci noti e ampiamente consolidati: Random Forest e Gradient Boosting. Durante la modellazione, è emersa l'importanza di valutare con attenzione l'impatto delle singole variabili predittive. In particolare, la variabile FamilyHistory, se non opportunamente gestita, determinava una separabilità quasi perfetta tra le classi, riducendo artificialmente la complessità del problema. La sua esclusione ha permesso una valutazione più realistica delle capacità predittive dei modelli. L'ottimizzazione degli iperparametri e della soglia di classificazione ha permesso di individuare le configurazioni migliori per ciascun modello, tenendo conto del delicato equilibrio tra sensibilità e precisione. Entrambi gli algoritmi hanno mostrato buone performance, con F1-score simili, ma con differenze significative nella distribuzione degli errori: mentre alcuni modelli privilegiavano la sensibilità, altri mantenevano una maggiore precisione. In ambito clinico, dove è preferibile non trascurare i casi positivi, l'attenzione alla sensibilità si è rivelata una scelta opportuna. I risultati ottenuti confermano la validità dell'approccio adottato e forniscono un esempio concreto di come strumenti di analisi dei dati possano supportare le decisioni mediche, contribuendo all'individuazione precoce di condizioni potenzialmente critiche come il diabete.