Rapporto tecnico

Classificazione del diabete tramite apprendimento automatico

1.0 Introduzione e obiettivi

La diagnosi precoce del diabete è un pilastro fondamentale per la gestione efficace della patologia. In questo contesto, i modelli di machine learning si affermano come potenti strumenti di supporto decisionale, capaci di identificare pattern complessi nei dati clinici. Questo rapporto documenta l'analisi e l'ottimizzazione di un modello Random Forest per la classificazione dei pazienti, ma il suo scopo va oltre la semplice valutazione metrica.

L'analisi dimostra come un modello standard, pur mostrando un'accuratezza apparentemente buona, possa nascondere un rischio clinico inaccettabile e come, attraverso un processo sistematico di ottimizzazione guidata dal contesto, possa essere trasformato in uno strumento di screening affidabile e sicuro.

L'obiettivo di questo documento è quindi triplice: valutare l'efficacia del classificatore Random Forest nella sua configurazione di base per identificarne i limiti operativi, analizzare i suoi driver predittivi per validarne la coerenza clinica e, infine, definire una configurazione ottimizzata che ne massimizzi l'utilità e la sicurezza in un contesto sanitario. La struttura del report seguirà un percorso analitico che parte dalla metodologia sperimentale, procede con la valutazione delle performance di baseline, approfondisce la capacità discriminante del modello, descrive l'ottimizzazione strategica della soglia di classificazione e culmina con una valutazione finale e un confronto con algoritmi alternativi.

Questa analisi rigorosa è essenziale per dimostrare come un modello venga trasformato da un algoritmo performante a uno strumento di supporto decisionale clinicamente valido.

2.0 Metodologia e configurazione sperimentale

Dataset

L'analisi è stata condotta sul "Pima Indians Diabetes Database", un dataset di riferimento per questo problema clinico. Il dataset contiene 8 feature per ciascun paziente: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction e Age.

La variabile target, Outcome, è binaria e indica la presenza (1) o l'assenza (0) di diabete.

Preparazione dei dati

Per garantire una valutazione oggettiva delle performance, il dataset è stato suddiviso in un set di addestramento (80% dei dati) e un set di test (20%). È stata applicata una suddivisione stratificata rispetto alla variabile target (Outcome) per assicurare che la proporzione delle classi fosse preservata in entrambi i sottoinsiemi, prevenendo così bias di valutazione.

Configurazione del modello

Il modello oggetto di questa analisi è un classificatore Random Forest (RandomForestClassifier). Per l'esperimento è stata scelta una configurazione robusta con n_estimators=1000, al fine di garantire stabilità e ridurre la varianza delle previsioni, fornendo una solida base di partenza per l'analisi.

Definita la configurazione sperimentale, procediamo ora alla valutazione delle performance di baseline del modello, utilizzando la soglia di classificazione standard come punto di riferimento critico.

3.0 Analisi delle prestazioni con soglia standard

La prima fase di valutazione analizza le performance del modello con la soglia di classificazione standard (0.5). Questa analisi funge da baseline per comprendere i punti di forza e, soprattutto, per identificare le debolezze operative intrinseche del modello prima di qualsiasi ottimizzazione.

3.2 Valutazione di baseline: accuratezza e limiti nascosti

Applicato al set di test, il modello raggiunge un'accuratezza complessiva di 0.81, indicando che l'81% delle previsioni totali è corretto. Sebbene questo valore suggerisca una buona capacità predittiva generale, un'analisi più approfondita è necessaria per svelare potenziali limiti non evidenti.

3.3 Analisi per classe: identificazione della debolezza critica (Recall)

Un'analisi granulare delle metriche per singola classe rivela una debolezza operativa inaccettabile. La tabella seguente riassume i risultati.

Classe	Precision	Recall	F1-score
0 (Non Diabetico)	0.82	0.91	0.86
1 (Diabetico)	0.79	0.63	0.70

Dall'analisi emerge che, mentre il modello è eccellente nell'identificare i pazienti non diabetici (Recall di 0.91), il Recall per la classe 1 (Diabetico) è solo di 0.63. Questo implica che il modello, nella sua configurazione standard, non è in grado di identificare il 37% dei pazienti effettivamente diabetici, un limite inaccettabile per qualsiasi applicazione clinica volta allo screening.

Analisi della matrice di confusione

La matrice di confusione quantifica la distribuzione degli errori del modello:

- Veri Negativi (TN): **91**
- Falsi Positivi (FP): **9**
- Falsi Negativi (FN): **20**
- Veri Positivi (TP): **34**

Il dato più critico è il numero di Falsi Negativi (FN), pari a 20. Questi rappresentano 20 casi di diabete non diagnosticati, confermando che il basso recall per la classe positiva costituisce il rischio operativo primario del modello. La necessità di ridurre questo numero è la principale spinta per le fasi successive di analisi.

4.0 Analisi dei driver predittivi e capacità discriminante

Oltre le metriche di classificazione, è essenziale comprendere quali variabili biologiche influenzano le decisioni del modello e qual è la sua reale capacità di distinguere tra le classi. Questa analisi ne valida la coerenza clinica e rivela il potenziale di ottimizzazione.

Analisi della Feature Importance

L'analisi dell'importanza delle feature rivela una chiara e clinicamente plausibile gerarchia nei fattori che guidano le previsioni del Random Forest:

- Driver Principale: Glucose (importanza ≈ 0.25), confermandosi come l'indicatore più potente.
- Contributo Significativo: BMI, Age e DiabetesPedigreeFunction seguono con un impatto rilevante.

- Impatto Moderato: BloodPressure e Pregnancies.
- Importanza Marginale: SkinThickness e Insulin.

Questa gerarchia, che privilegia indicatori metabolici universalmente riconosciuti, rafforza la validità del modello e la sua aderenza ai noti fattori di rischio per il diabete.

Curva ROC e Punteggio AUC

Il punteggio AUC (Area Under the Curve) valuta la capacità discriminante del modello indipendentemente dalla soglia. Il modello ottiene un eccellente punteggio AUC di 0.871.

L'elevato valore di AUC (0.871) crea un'apparente contraddizione con il basso recall (0.63) osservato nella configurazione standard. Questo non indica un fallimento del modello nel "comprendere" i pattern, ma piuttosto un disallineamento della sua soglia di decisione (0.5) con l'obiettivo clinico di massimizzare la sensibilità. La capacità discriminante esiste; il nostro compito è sfruttarla regolando il punto di intervento.

L'elevato valore di AUC suggerisce quindi che è possibile ottimizzare il trade-off tra sensibilità e specificità, introducendo la necessità di ricalibrare la soglia di decisione.

5.0 Ottimizzazione della soglia di classificazione

In un contesto sanitario, l'impatto di un errore non è simmetrico: una mancata diagnosi (falso negativo) è più grave di un falso allarme (falso positivo). La regolazione della soglia è quindi un passo strategico per trasformare un modello tecnicamente valido in uno strumento clinicamente utile.

Analisi del Trade-off Precisione-Recall

Esiste una relazione inversa tra Precisione e Recall, modulabile agendo sulla soglia di decisione. Abbassando la soglia, il modello diventa più "sensibile", aumentando il Recall a scapito della Precisione. Aumentandola, diventa più "selettivo", con l'effetto opposto. L'obiettivo è identificare il punto operativo sulla curva Precision-Recall che massimizza l'F1-score, bilanciando in modo ottimale la sensibilità diagnostica con il contenimento dei falsi positivi.

Identificazione della soglia ottimale

Per trovare il miglior equilibrio tra Precisione e Recall per la classe positiva, è stata individuata la soglia che massimizza l'F1-score. L'analisi ha rivelato che la soglia ottimale per questo modello è 0.38, che permette di raggiungere un F1-score massimo di 0.78. Questo valore, inferiore allo standard 0.5, indica che una maggiore sensibilità è necessaria per ottimizzare le performance complessive.

Procediamo ora a valutare l'impatto di questa nuova soglia sulle metriche di performance finali del modello.

6.0 Valutazione finale del modello con soglia ottimizzata (0.38)

Questa sezione presenta la valutazione definitiva del modello configurato con la soglia ottimizzata a 0.38, per verificare la risoluzione della debolezza critica iniziale e la creazione di un classificatore clinicamente affidabile.

Analisi del classification report ottimizzato

L'applicazione della nuova soglia ha prodotto un miglioramento mirato e significativo delle performance, come riassunto di seguito.

Classe	Precision	Recall	F1-score
0 (Non Diabetico)	0.90	0.83	0.86
1 (Diabetico)	0.73	0.83	0.78

I miglioramenti chiave sono:

- Il Recall per la classe 1 (Diabetico) è aumentato drasticamente da 0.63 a 0.83. Il modello ora identifica correttamente l'83% dei pazienti diabetici.
- L'F1-score per la classe 1 è migliorato da 0.70 a 0.78, indicando un equilibrio molto più robusto.
- L'accuratezza complessiva è salita da 0.81 a 0.83, e il weighted avg F1-score di 0.83 conferma un miglioramento bilanciato su entrambe le classi.

Analisi della matrice di confusione aggiornata

Il cambiamento è ancora più evidente analizzando la nuova distribuzione degli errori:

- Veri Negativi (TN): 83
- Falsi Positivi (FP): 17
- Falsi Negativi (FN): 9
- Veri Positivi (TP): 45

Il risultato più importante è che il numero di falsi negativi è stato ridotto del 55% (da 20 a 9), mitigando drasticamente il rischio operativo più critico del modello.

Implicazioni Strategiche e Cliniche

La regolazione della soglia a 0.38 ha trasformato il modello. Da strumento con una grave lacuna nella sensibilità, è diventato uno strumento di screening clinicamente sicuro, capace di catturare la stragrande maggioranza dei casi positivi (recall 83%) mantenendo una precisione accettabile (73%).

7.0 Analisi comparativa e raccomandazione

Per validare la scelta del Random Forest, è fondamentale un benchmarking rispetto ad altri algoritmi di boosting ad alte prestazioni come XGBoost e LightGBM. Questo confronto contestualizza le performance e assicura che la soluzione scelta sia la più efficace.

Confronto delle prestazioni

La tabella seguente sintetizza le metriche chiave per i tre modelli, tutti valutati dopo aver ottimizzato la loro soglia per massimizzare l'F1-score.

Modello	Accuracy	Precision (Classe 1)	Recall (Classe 1)	F1-score (Classe 1)	AUC
Random Forest	**0.83**	**0.73**	0.83	**0.78**	**0.8708**
XGBoost	0.73	0.60	0.78	0.68	0.7761
LightGBM	0.70	0.55	**0.84**	0.67	0.7679

Nota: i modelli sono stati valutati su partizioni di test generate in modo indipendente; sebbene le proporzioni siano identiche (80/20), i dati specifici nel test set possono variare leggermente.

Analisi Comparativa

Dall'analisi emerge la netta superiorità del modello Random Forest:

- Supera significativamente XGBoost e LightGBM in Accuracy, F1-score e AUC.

- Sebbene LightGBM raggiunga il recall più elevato in assoluto (0.84), questo risultato è ottenuto a fronte di un crollo della precisione a 0.55, rendendolo operativamente insostenibile a causa dell'eccessivo numero di falsi allarmi. Il Random Forest offre un equilibrio nettamente superiore, mantenendo un recall quasi identico (0.83) ma con una precisione (0.73) che lo rende affidabile in un contesto reale.

Il confronto conferma che, per questo specifico problema, il Random Forest ottimizzato rappresenta la scelta più robusta ed equilibrata.

8.0 Conclusioni e raccomandazioni

Sintesi dei Risultati

L'analisi è partita da un modello con una debolezza operativa inaccettabile (recall del 63% sui casi positivi), nonostante una buona accuratezza di baseline. L'identificazione di un'elevata capacità discriminante (AUC = 0.871) ha permesso di attribuire il problema a una soglia di classificazione disallineata con gli obiettivi clinici. L'ottimizzazione della soglia a 0.38 ha risolto questa criticità, aumentando il recall all'83% e riducendo le mancate diagnosi del 55%. Infine, il benchmarking ha confermato la superiorità di questa configurazione rispetto a modelli alternativi come XGBoost e LightGBM.

Raccomandazione Finale

Azione Raccomandata: Adottare il modello RandomForestClassifier (n_estimators=1000) per l'identificazione dei pazienti a rischio.

Parametro Operativo Critico: Impostare la soglia di classificazione a 0.38.

Giustificazione: Questa configurazione offre il miglior equilibrio tra performance predittive (Accuracy: 0.83, AUC: 0.87), massimizzando la capacità di identificare i casi positivi (Recall: 0.83) e riducendo il rischio di mancate diagnosi del 55% rispetto alla configurazione standard.

Considerazioni sull'Impatto

Questo studio dimostra come un'attenta ottimizzazione, guidata dal contesto applicativo, sia fondamentale per trasformare un modello di machine learning in uno strumento di valore, affidabile e sicuro. La priorità di minimizzare le mancate diagnosi ha guidato il processo, producendo una soluzione di intelligenza artificiale che non è solo potente, ma anche responsabile.