Report Business Intelligence

Bagnato Francesca s346710 | Jovanovic Tamara s344853 | De Toffol Lorenzo s343845 | Foni Gianmarco s347952 | Francabandiera Riccardo s345950

Abstract - Questo progetto si propone di analizzare e validare diversi modelli di classificazione per individuare i pazienti affetti da diabete o a rischio, in base ad una serie di caratteristiche mediche. Dopo un'attenta fase di preprocessing — comprensiva di gestione dei valori mancanti, encoding, rilevamento outlier, normalizzazione e riduzione della dimensionalità tramite PCA — sono stati testati diversi algoritmi: Decision Tree, Random Forest, Gradient Boosting, K-Nearest Neighbors e Support Vector Machines. I modelli hanno mostrato prestazioni eccellenti, con accuracies fino al 100% sui dati di test per Random Forest, SVM (kernel RBF) e Gradient Boosting. I risultati confermano che, con un adeguato preprocessing, anche dataset di modesta dimensione possono fornire ottimi risultati predittivi. L'approccio si dimostra utile in ambito clinico per identificare precocemente soggetti a rischio e ottimizzare le strategie di intervento.

I. Introduzione

Il diabete è una patologia cronica, per la quale ancora non esiste una terapia risolutiva. Le conseguenze a lungo termine del diabete non controllato compromettono la salute degli individui a diversi livelli e sono causa di importanti disabilità e gravi sindromi acute come insufficienza renale, patologie cardiovascolari, tra le più temibili ictus e sindromi coronariche. Riconoscere gli individui affetti da diabete, e quelli che invece sono a rischio di svilupparlo, è sempre stato di fondamentale importanza per la comunità scientifica, e non solo. Una diagnosi precoce ci consente di iniziare le terapie idonee prima che si instaurino danni permanenti sugli altri organi e sistemi, e nei pazienti non ancora affetti da diabete franco iniziare modifiche dello stile di vita e/o introdurre terapie che potrebbero prevenire l'insorgenza della malattia. Le assicurazioni sanitarie hanno particolare interesse a conoscere tempestivamente lo stato glicemico dei loro clienti considerando che il diabete stesso comporta una serie di terapie, analisi ed accertamenti nel corso della vita di un individuo, modificando così l'importo del premio assicurativo. Inoltre, adottare adeguate strategie di prevenzione in questa categoria a rischio aumentato di complicanze non solo diabetiche, potrebbe aiutare a ridurre la spesa sanitaria e migliorare la qualità della vita dei pazienti.

Questo progetto ha come scopo la validazione e l'analisi dei metodi di classificazione dei pazienti. Si descrive l'analisi preliminare dei dati, la scelta degli algoritmi e la loro analisi in termini di validità ed efficacia.

II. Analisi dei dati e preprocessing

Per prima cosa si è fatta un'analisi del dataset al fine di applicare un preprocessing adeguato per gestire dati mancanti, outliers e normalizzazioni. Il dataset è composto da 588 record e 12 features, di cui una rappresenta la label di presenza o meno della patologia del diabete. In particolare, i record sono distribuiti in modo poco omogeneo rispetto a tale etichetta, presentando un paziente diabetico ogni tre pazienti sani (Fig.1).

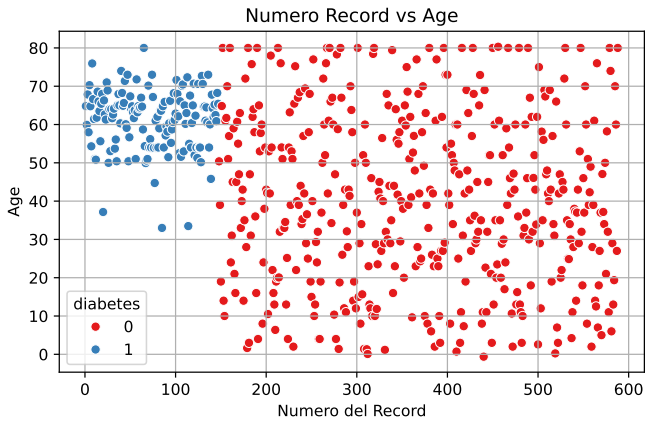


Fig. 1. La prima porzione di dataset è composta da pazienti con diabete. Si noti la differenza di dispersione intorno all'età media nei due gruppi.

Gestione valori nulli

In un'analisi preliminare del dataset è stata trovata la presenza di dati nulli per le seguenti features: *smoking history e Insulin Sensitivity Est*. Nel primo caso, la mancanza del dato è rappresentata dal valore categorico "No Info" ed è stata gestita nella seguente maniera: (Fig.2)
Se l'età del paziente è

- inferiore a 15 anni, "No Info" è stato sostituito con "Never". Si suppone che i pazienti giovani non abbiano fatto esperienza di fumo.
- superiore a 15 anni, "No Info" è stato sostituito con "Current", in modo da considerare il caso peggiore possibile. L'approccio adottato nella nostra analisi è di preferire un falso positivo a un falso negativo.

Nel secondo caso, i valori mancanti sono rappresentati da "NaN" e sono stati sostituiti con la media del dataset. Si è provato a vedere se fosse possibile ricavare il valore mancante da altre features presenti, ma vista la scarsa correlazione, si è deciso di optare per il valor medio.

Encoding

È stato necessario effettuare un processo di encoding delle variabili 'gender' e 'smoking history'.

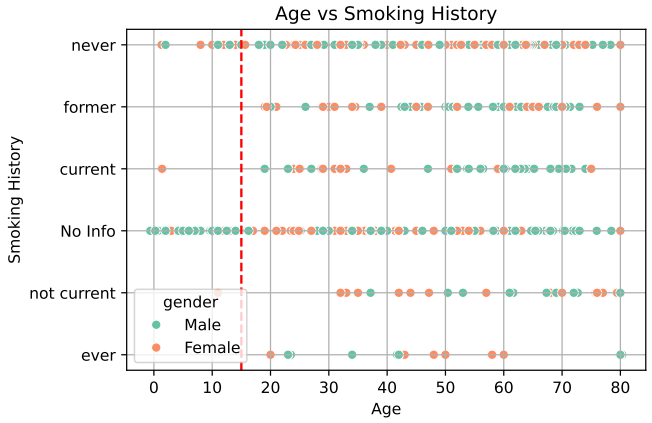


Fig. 2. In rosso la linea corrispondente al valore di età 15 anni

Per ‘gender‘ è stato utilizzato Label Encoding, poiché si tratta di una variabile binaria, e l’assegnazione di valori numerici distinti (0 e 1) è sufficiente.
Per ‘smoking history‘, che rappresenta più categorie, si è optato per One-Hot Encoding, per evitare di introdurre un ordine arbitrario tra le categorie.

Outlier Detection

Sono stati rimossi i record in cui il valore dell’indice di massa corporea (BMI) era superiore a 60, in quanto valori così elevati sono estremamente rari nella popolazione e difficilmente realistici, quindi tali record possono essere classificati come outliers.

Il tipo dell’attributo ’age‘ è stato trasformato da float a intero e i record con valori di ’age‘ < 1 sono stati rimossi. Un’osservazione che risulta interessante sull’età è che, per pazienti più giovani di 50 anni l’analisi può essere semplificata, in quanto gli unici indicatori significativi risultano essere HbA1c_level e blood_glucose_level, mentre hypertension, smoking_history e heart_disease difficilmente sono presenti.

Normalizzazione

Alcune variabili numeriche rappresentavano scale differenti, pertanto è stata applicata la Min Max Normalization per riportarle tutte all’interno dell’intervallo [-1,1]. Le variabili normalizzate sono:

- age,
- bmi,
- HbA1c_level,
- blood_glucose_level,
- Random_Lab_Marker.

È stata rimossa la feature BMI_Glucose_Interaction in quanto è possibile ricavarla da altri attributi (”bmi” e ”blood_glucose_level”). Facendo un’analisi di correlazione, possiamo vedere quanto BMI_Glucose_Interaction sia fortemente correlata alle altre due features.

	bmi	blood_glucose_level	BMI_Glucose_Interaction
bmi	1.00	0.243822	0.668118
blood_glucose_level	0.243822	1.00	0.864132
BMI_Glucose_Interaction	0.668118	0.864132	1.00

Dopo aver completato il preprocessing del dataset di training, i dati sono stati suddivisi in feature (X_train) e target label (y_train).

Preprocessing del test set

Il test set è stato preprocessato seguendo le stesse procedure applicate ai dati di training (encoding e normalizzazioni), con l’eccezione del rilevamento e rimozione degli outlier.

Riduzione della dimensionalità tramite PCA

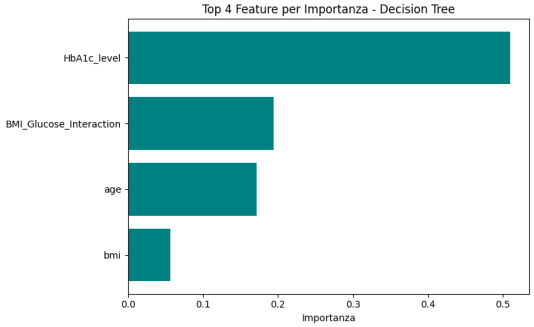
È stata applicata l’analisi delle componenti principali (PCA) per ridurre la dimensionalità del dataset e darne una rappresentazione grafica. Non si è poi rivelato necessario applicarla per migliorare le prestazioni, in quanto i risultati ottenuti risultavano già più che soddisfacenti. La trasformazione è stata eseguita impostando come soglia il 90% della explained variance. Il risultato ha mostrato che:

- il numero di componenti principali selezionate è pari a 7;
- l’explained variance complessiva è pari al 90.53

III. Metodologia

A. Decision Trees

I decision tree permettono di individuare le features più importanti nello studio del fenomeno d’interesse. Dalla tabella possiamo vedere i 4 parametri più significativi in ordine decrescente:



Si allega al fondo l’immagine del decision tree ottenuto.

B. Random Forest

Il metodo Random Forest combina gli output di diversi decision trees in un unico risultato. Ha una maggiore flessibilità, in quanto gestisce con alta precisione sia problemi di regressione che di classificazione. È efficace nello stimare valori mancanti, mantenendo una buona accuratezza. Inoltre permette di gestire bene i dati con valori mancanti grazie al suo approccio robusto.

C. Gradient Boost

Nonostante l'elevata accuratezza di Decision Trees ottenuta, si è deciso di esplorare l'utilizzo di Gradient Boosting Machines in questo contesto, come naturale evoluzione della classificazione. In effetti, GBMs si basano sull'ottimizzazione dei modelli più semplici, come Decision Trees, perché a ogni passo cercano di migliorare l'errore. I GBM consentono all'utente di ottimizzare una funzione di perdita specificata in base all'obiettivo desiderato.

D. K-Nearest Neighbors

Il K-Nearest Neighbors (KNN) è un metodo semplice ed efficace, basato sulla distanza tra punti nel feature space: un nuovo punto viene classificato in base alle classi dei suoi k vicini più prossimi. Il metodo non assume alcuna ipotesi sulla distribuzione dei dati, rendendolo flessibile ma anche sensibile alla scala delle feature e alla presenza di rumore.

In questo studio si è scelto di includere KNN proprio perché il dataset utilizzato è relativamente piccolo, condizione in cui KNN può risultare competitivo rispetto a modelli più complessi. Inoltre, è stato effettuato un preprocessing adeguato per garantire l'efficacia del metodo: in particolare, le features numeriche sono state normalizzate per evitare che quelle con range più ampio dominassero la distanza euclidea. Questo ha permesso al KNN di operare correttamente anche in presenza di feature eterogenee. Sono stati anche eliminati gli outliers, ai quali il modello è molto sensibile.

E. SVM

Il metodo SVM (Support Vector Machines) consiste nel cercare un modo di separare le classi nello spazio. In particolare, se si usa una SVM lineare, si cerca il miglior iperpiano di separazione tra le classi. Nell'analisi seguente si è utilizzato infatti l'SVM lineare, ma anche SVM con kernel non lineari al fine di confrontare i risultati ottenuti. Questo metodo presenta spesso l'accuratezza migliore tra gli altri, per questo si è deciso di utilizzarlo, nonostante non sia incrementabile, come paragone per il resto dell'analisi.

IV. Risultati sperimentali

Per ogni modello si riportano le metriche di accuratezza sui dati di training e test.

Decision tree

L'informazione fornita dai Decision Trees è che ci sono due feature principali particolarmente significative per la diagnosi del diabete: la prima è l'età e la seconda è HbA1c_level. Abbiamo quindi che, all'aumentare dell'età, aumenta il rischio di diabete, e che uno degli indicatori più significativi è il livello di glucosio nel sangue. Ci sono anche altri fattori incisivi, ad esempio una storia di fumo, ma questi sono i prevalenti. Come anticipato, il ramo dell'albero per cui age $<$ 50 raggiunge prima un nodo foglia, in quanto non esplora i parametri indicati sopra (smoking history, heart disease, ipertensione).

Random Forest

E' stato utilizzato un modello di Random Forest per migliorare le prestazioni rispetto a un singolo albero decisionale, il quale non raggiungeva un'accuratezza perfetta sul training set. La Random Forest ha ottenuto un'accuratezza, precision e recall pari a 1 sia sul training set che sul test set. Questo risultato indica un'elevata capacità di classificare correttamente le istanze, dovuta alla struttura del modello, che combina più alberi decisionali e consente una riduzione della varianza e una maggiore robustezza rispetto all'impiego di un solo albero.

Gradient boosting

Nonostante gli ottimi risultati forniti da Decision trees e Random forest, si è deciso di utilizzare il metodo Gradient boosting per confermare la performance ottenuta con le metodologie prima citate.

K-Nearest Neighbors

Nel caso del modello **KNN** la scelta del parametro k è stata guidata da due approcci complementari:

- Confronto tra training e test accuracy (Elbow Method): sono state calcolate le accuratezze su training set e test set per valori di k compresi tra 1 e 20. Questo ha permesso di osservare il punto in cui l'accuracy sul test set si stabilizza o inizia a decrescere, evidenziando l'eventuale overfitting per k troppo piccoli o underfitting per k troppo grandi. Fig.4
- Validazione incrociata (5-fold CV): per ogni valore di k, è stata calcolata l'accuracy media tramite validazione incrociata su 5 fold. È stato selezionato il valore di k che ha massimizzato questa media, ottenendo un equilibrio ottimale tra bias e varianza. (Fig.3)

Combinando le due analisi, si è identificato il valore ottimale $k = 5$, che ha fornito la migliore generalizzazione sul validation set.

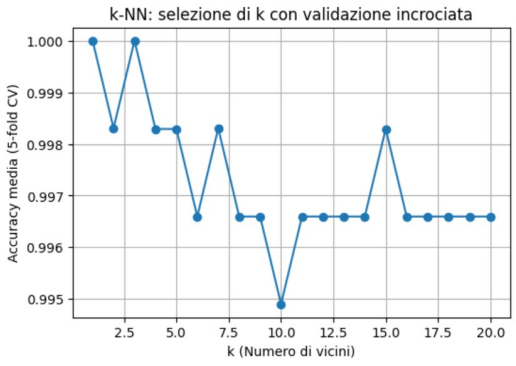


Fig. 3. Accuracy media ottenuta tramite validazione incrociata a 5 fold in funzione del parametro k del classificatore k-NN.

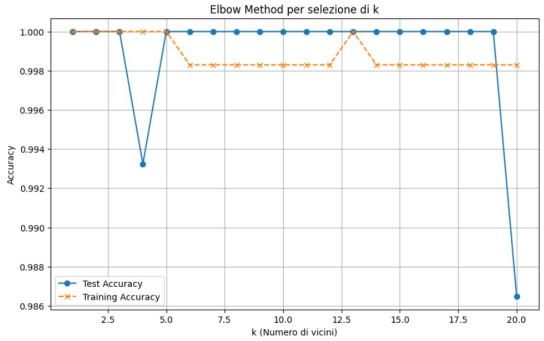


Fig. 4. Accuracy su training e test set al variare di k (Elbow Method).

SVM

Dall'analisi del dataset con SVM, si evince che i dati sono facilmente separabili linearmente. Infatti, nonostante si utilizzi un margine C hard (10^{10}) l'accuratezza ottenuta è 1. Con kernel diversi il risultato è simile a patto di inserire un C soft, quindi molto minore. Un fattore C più piccolo permette al modello di commettere alcuni errori senza che ciò penalizzi eccessivamente il modello. Per il kernel sigmoidale è stato effettuato un parameter tuning su C per ottenere un'accuratezza vicina a 1. Nonostante questa procedura non è stato possibile arrivare alla classificazione perfetta del dataset. Questo significa che i dati non sono facilmente divisibili attraverso sigmoidi. Nell'ultimo caso si è utilizzato il kernel RBF con C soft partendo dal valore considerato nel caso sigmoidale. Il risultato conferma la perfetta classificazione del dataset con questo metodo, senza ulteriore utilizzo di tuning di C. In conclusione, il dataset risulta essere ben divisibile da iperpiani o da curve dei kernel usati, dimostrando quindi che i dati sono di facile analisi una volta preprocessati.

Fig. 5. Decision Tree (per l'immagine completa consultare la png nel folder).

TABLE I
ACCURATEZZA DEI MODELLI SUI SET DI TRAINING E TEST

Modello	Training (%)	Test (%)
Decision Tree	94.46	100.00
Random Forest	100.00	100.00
k-NN ($k = 5$)	100.00	100.00
Gradient Boosting	100.00	100.00
SVM hard margin	100.00	100.00
SVM kernel sigmoidale	99.31	98.65
SVM kernel RBF	100.00	100.00

TABLE II
MATRICI DI CONFUSIONE PER I MODELLI CLASSIFICATORI

Modello	**Actual \Predicted**	**No**	**Yes**
Decision Tree	No	419	12
	Yes	16	131
Random Forest (CV 5-fold)	No	431	0
	Yes	0	147
Gradient Boosting	No	431	0
	Yes	0	147

V. CONCLUSIONI

Il fenomeno descritto da questi dati è perfettamente modellizzabile attraverso gli strumenti da noi utilizzati. Il preprocessing svolto si è rivelato adeguato a gestire le criticità del dataset, permettendo di raggiungere un tale livello di precisione. Non si è reso necessario l'uso della cross-validation (tranne che per k-NN), in quanto il dataset risulta sufficientemente semplice e bilanciato, senza evidenti rischi di overfitting o di varianza elevata. Inoltre, non è stata applicata la tecnica di riduzione della dimensionalità tramite PCA, poiché le variabili non presentavano un numero elevato di caratteristiche tali da richiedere una riduzione della dimensionalità.
Il dataset riflette una particolarità intrinseca alla patologia, ovvero diverse forme di diabete che colpiscono popolazione più giovane (diabete di tipo I) e popolazione di età > 50 anni (diabete di tipo II). In quest'ultimo caso, l'analisi potrebbe essere ulteriormente raffinata per comprendere meglio il contributo dei singoli fattori di rischio.
In conclusione, la semplicità intrinseca del dataset e l'efficacia del preprocessing hanno reso superflue ulteriori tecniche di validazione o trasformazione, consentendo di raggiungere risultati ottimali.

VI. CONTRIBUTI

Ciascun membro del gruppo ha contribuito attivamente allo sviluppo del progetto. In particolare:

- Bagnato Francesca: modello *Random Forest* e stesura del report;
- Jovanovic Tamara: preprocessing, implementazione dei dati e modello *Gradient Boosting*;
- De Toffol Lorenzo: modello *Decision Tree* e stesura del report;
- Foni Gianmarco: modello *k-NN* e stesura del report;
- Francabandiera Riccardo: preprocessing, implementazione dei dati e modello *SVM*.