

LM-32 - Apprendimento Automatico e Apprendimento Profondo

Progetto di sviluppo di un progetto di machine/deep learning

Infurna Giuseppe
Matricola: 0322500071

Bank Marketing Classification

Introduzione

La classificazione è una tecnica di machine learning in cui un modello viene addestrato utilizzando dati etichettati, con l'obiettivo di assegnare una classe discreta a nuove osservazioni. Ogni esempio del dataset è descritto da un insieme di feature e da una variabile target.

Il problema di Bank Marketing disponibile sul rientra pienamente nella classificazione supervisionata poiché l'obiettivo è prevedere se un cliente sottoscriverà o meno un deposito a termine sulla base delle sue caratteristiche demografiche e del contesto della campagna di marketing.

La variabile target y è binaria e assume valore 'yes' se il cliente sottoscrive il deposito, e 'no' in caso contrario.

Dataset (UCI Bank Marketing)

Il dataset Bank Marketing è distribuito dal repository UCI Machine Learning Repository ed è ampiamente utilizzato come benchmark per problemi di classificazione.

Sono disponibili due file principali: `bank.csv`, che contiene una versione ridotta del dataset, e `bank-full.csv`, che include l'intero insieme di osservazioni raccolte durante le campagne di marketing.

È stato scelto il file `bank-full.csv` per disporre di un numero maggiore di esempi, aumentando la robustezza statistica dei modelli addestrati.

Il dataset completo contiene circa 45.000 osservazioni e 38 feature, tra variabili numeriche e categoriche, oltre alla variabile target.

Obiettivi

Il progetto ha come obiettivo la previsione della sottoscrizione di un deposito a termine da parte dei clienti, utilizzando il dataset di Bank Marketing. Il dataset è caratterizzato da uno sbilanciamento delle classi e da variabili eterogenee, che hanno richiesto una fase di preprocessing accurata. Dopo la pulizia dei dati e la codifica delle variabili categoriche, è stata applicata una riduzione dimensionale tramite PCA, al fine di ridurre la complessità del problema e mitigare la multicollinearità. Sono stati implementati e confrontati diversi algoritmi di classificazione: Logistic Regression, Naive Bayes, Decision Tree e Random Forest. La scelta di questi modelli consente di confrontare approcci lineari, probabilistici e basati su alberi.

Preprocessing dei dati

Il preprocessing è una fase fondamentale per garantire che i dati siano compatibili con gli algoritmi di machine learning e che le prestazioni dei modelli siano affidabili.

Le variabili categoriche vengono trasformate tramite one-hot encoding, una tecnica che converte ogni categoria in una variabile binaria, permettendo ai modelli di interpretarle correttamente.

Le variabili numeriche vengono standardizzate affinché abbiano media zero e varianza unitaria. La standardizzazione è essenziale per la PCA e per modelli sensibili alla scala delle feature.

Da questa prima fase otteniamo i seguenti risultati:

Record totali: 45211

Record utilizzati (senza valori nulli): 7842

Record esclusi (con almeno un valore nullo): 37369

Split 70/30/10

Training set: 5489, 38

Validation set: 1576, 38

Test set: 777, 38

Analisi delle feature – PCA

La Principal Component Analysis (PCA) è una tecnica di riduzione della dimensionalità che consente di proiettare i dati in uno spazio a dimensione ridotta preservando la massima varianza possibile.

La varianza misura la quantità di informazione contenuta nei dati. Le componenti principali sono combinazioni lineari delle feature originali ordinate per varianza spiegata.

Lo scatter plot PC1 vs PC2 consente di osservare la distribuzione dei dati nello spazio ridotto.

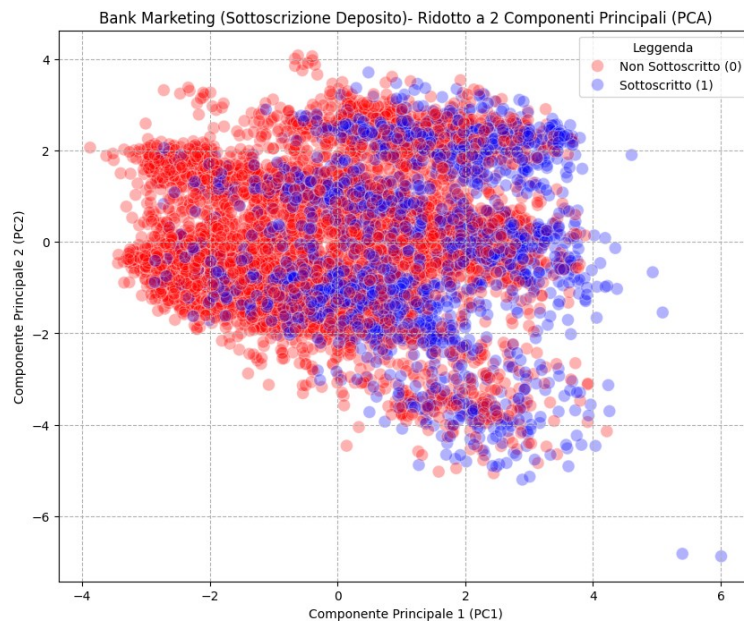
La parziale sovrapposizione delle classi indica che il problema non è linearmente separabile, rendendo necessari modelli di classificazione più complessi.

$$z = \frac{x - \mu}{\sigma}$$

componente principale

$$PC1 = w_1x_1 + w_2x_2 + \dots + w_nx_n$$

La prima componente principale massimizza la varianza spiegata, mentre le successive catturano informazione residua non correlata.



Scatter plot delle prime due component

Lo scatter plot PC1 vs PC2 consente di osservare la distribuzione dei dati nello spazio ridotto.

Algoritmi di classificazione

1. Naive Bayes (Gaussian Naive Bayes)

Il Naive Bayes è un classificatore probabilistico basato sul teorema di Bayes, con l'ipotesi "naive" di indipendenza condizionata tra le feature dato l'esito della classe.

Pro: semplice, veloce, robusto su dataset ad alta dimensionalità

Contro: ipotesi di indipendenza spesso non realistica

2. Decision Tree (Albero di Decisione)

Un Decision Tree costruisce un modello gerarchico che suddivide lo spazio delle feature in regioni omogenee, tramite regole if-then.

Ogni osservazione segue un percorso dall'alto verso il basso fino a una foglia, che assegna la classe più frequente nel nodo.

Pro: altamente interpretabile, non richiede scaling

Contro: tende all'overfitting, alta varianza

3. Random Forest (RF)

La Random Forest è un ensemble di Decision Tree, addestrati su campioni diversi del dataset e con sottoinsiemi casuali di feature.

Pro: alta accuratezza, robusto al rumore

Contro: minore interpretabilità rispetto a un singolo albero

4. Logistic Regression (LogReg)

La Logistic Regression è un modello lineare probabilistico che stima la probabilità di appartenenza a una classe tramite la funzione sigmoide.

Pro: interpretabile, stabile, ottima baseline

Contro: cattura solo relazioni lineari

Algoritmo	Tipo	Interpretabilità	Capacità non lineare
Naive Bayes	Probabilistico	Media	Bassa
Decision Tree	Regole	Alta	Media
Random Forest	Ensemble	Bassa	Alta
Logistic Regression	Lineare	Alta	Bassa

Metriche di valutazione

Per ciascun modello sono state adottate diverse metriche di valutazione, tra cui accuracy, precision, recall, F1-score e ROC AUC, in modo da avere una valutazione completa delle prestazioni.

L'*accuracy* misura la percentuale di predizioni corrette ma può risultare fuorviante in presenza di classi sbilanciate.

Precision e *recall* permettono di valutare rispettivamente l'affidabilità delle predizioni positive e la capacità di individuare correttamente i casi positivi.

L'*F1-score* rappresenta una media armonica tra precision e recall.

La *ROC AUC* misura la capacità discriminante del modello e ha un'interpretazione geometrica come area sotto la curva ROC.

True Positive RATE

$$TPR = \frac{TP}{TP + FN}$$

False Positive RATE

$$FPR = \frac{FP}{FP + TN}$$

Area Under Curve:

$$AUC = \int_0^1 TPR(FPR) d(FPR)$$

L'AUC rappresenta la probabilità che il modello assegni uno score maggiore a un'istanza positiva rispetto a una negativa.

Risultati sperimentali

I modelli di classificazione sono stati addestrati utilizzando l'intero spazio delle feature preprocessate e valutati sul test set. I risultati quantitativi, riassunti nella tabella delle metriche, mostrano comportamenti differenti tra gli algoritmi considerati:

	Accuracy	Precision	Recall	F1	ROC AUC
Naive Bayes	0.5206	0.3145	0.9366	0.4709	0.8269
Decision Tree	0.7926	0.5417	0.5821	0.5612	0.7184
Random Forest (RF)	0.8419	0.8254	0.3881	0.5279	0.8993
Logistic Regression (LogReg)	0.8406	0.6940	0.5373	0.6057	0.8897

Naive Bayes presenta un'Accuracy pari a 0.5206, ma un Recall molto elevato (0.9366) a fronte di una Precision bassa (0.3145). Questo comportamento è coerente con la matrice di confusione, che evidenzia un numero molto elevato di falsi positivi e una forte tendenza del modello a classificare i record come Sottoscritto. Il modello risulta quindi sensibile ma poco selettivo.

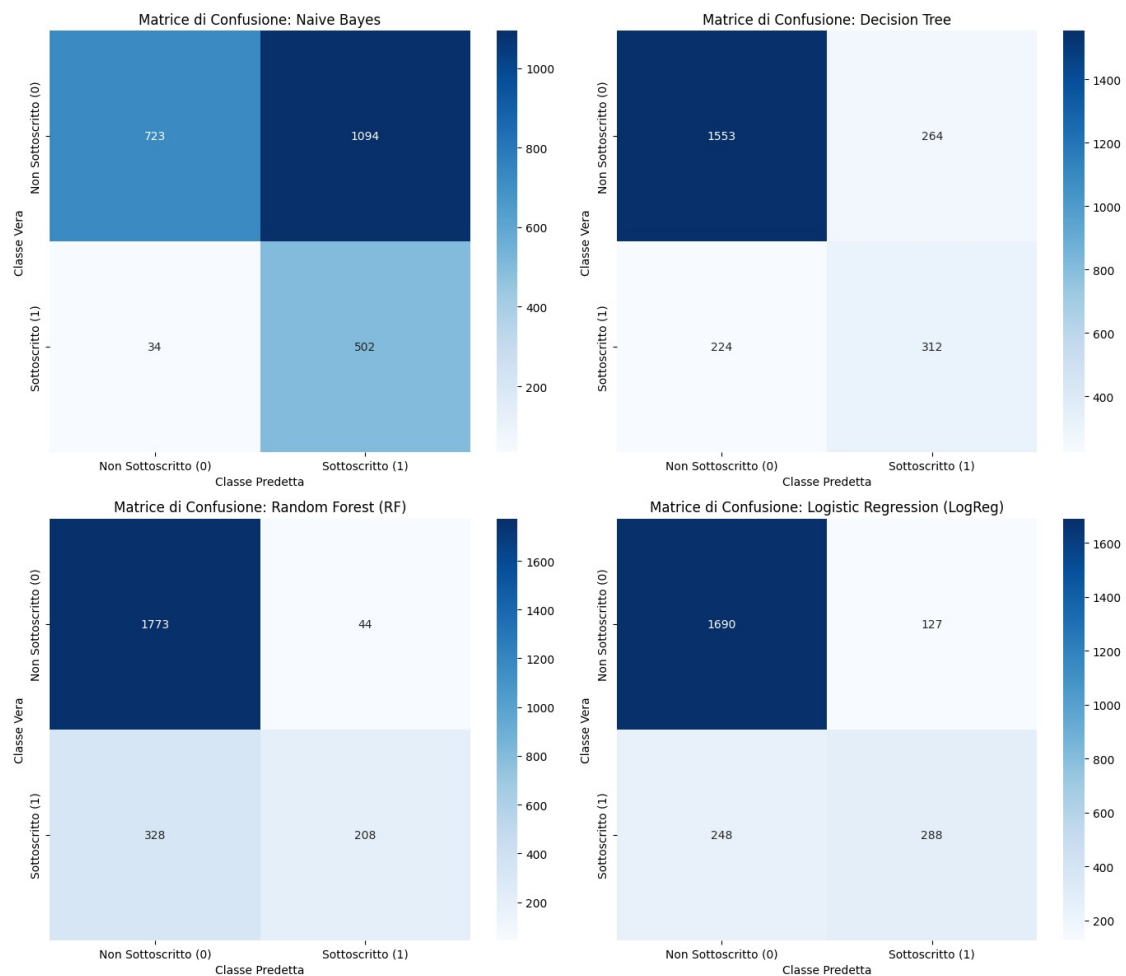
Decision Tree raggiunge un'Accuracy di 0.7926, con valori di Precision (0.5417) e Recall (0.5821) più bilanciati rispetto a Naive Bayes, ma con una ROC-AUC pari a 0.7184, inferiore agli altri modelli. Questo suggerisce una capacità discriminante complessivamente più limitata.

Random Forest ottiene l'Accuracy più elevata (0.8419) e una Precision molto alta (0.8254), ma a fronte di un Recall contenuto (0.3881). La matrice di confusione mostra che il modello è fortemente conservativo, privilegiando la corretta classificazione dei Non Sottoscritti e riducendo i falsi positivi, ma penalizzando l'individuazione dei casi positivi. La ROC-AUC pari a 0.8993 indica comunque una buona capacità di separazione complessiva.

Logistic Regression presenta un'Accuracy di 0.8406, comparabile a quella della Random Forest, ma con un miglior equilibrio tra Precision (0.6940) e Recall (0.5373). Questo equilibrio si riflette nel F1-score più elevato (0.6057) tra i modelli analizzati, rendendo la Regressione Logistica il modello con il miglior compromesso complessivo tra sensibilità e selettività. La ROC-AUC pari a 0.8897 conferma una buona capacità discriminante.

Nel complesso, l'analisi evidenzia come il dataset presenti una struttura informativa sufficiente a supportare modelli di classificazione efficaci, ma anche un marcato trade-off tra Precision e Recall, influenzato dallo sbilanciamento delle classi. I risultati mostrano che modelli diversi ottimizzano aspetti differenti del problema: Naive Bayes massimizza il Recall, Random Forest la Precision, mentre la Regressione Logistica fornisce il miglior equilibrio complessivo.

L'integrazione delle matrici di confusione, delle metriche quantitative e dell'analisi PCA consente una valutazione completa del comportamento dei modelli, evidenziando punti di forza e limiti di ciascun approccio nel contesto specifico del problema affrontato.



Curve ROC

Gli stessi risultati sono stati estratti dalla curva di ROC dove emerge chiaramente che:

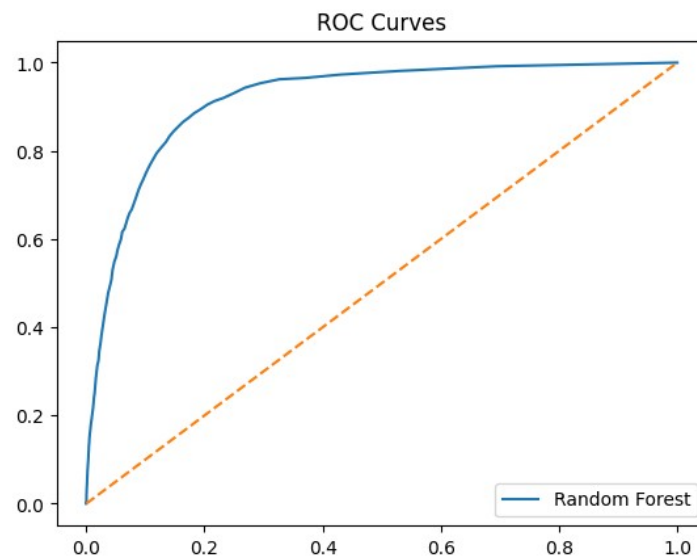
Random Forest è il modello con le migliori prestazioni complessive, raggiungendo un $AUC \approx 0.90$, indice di un'elevata capacità discriminante tra le due classi. La curva ROC si mantiene stabilmente sopra quelle degli altri modelli, soprattutto nelle regioni a basso tasso di falsi positivi (FPR), aspetto cruciale in contesti applicativi sensibili agli errori.

Logistic Regression mostra prestazioni molto competitive ($AUC \approx 0.89$), risultando solo leggermente inferiore alla Random Forest. Questo risultato è particolarmente rilevante considerando la maggiore semplicità e interpretabilità del modello.

Naive Bayes ottiene un'AUC intermedia (≈ 0.83), indicando una buona capacità predittiva ma una minore flessibilità nel catturare relazioni complesse tra le variabili.

Decision Tree risulta il modello meno performante ($AUC \approx 0.72$), con una curva ROC più vicina alla diagonale casuale, segnale di una capacità discriminante limitata e di una maggiore sensibilità al rumore dei dati.

Nel complesso, la curva ROC conferma che Random Forest e Logistic Regression rappresentano il miglior compromesso tra sensibilità (TPR) e specificità ($1 - \text{FPR}$).



Spiegabilità con approccio SHAP

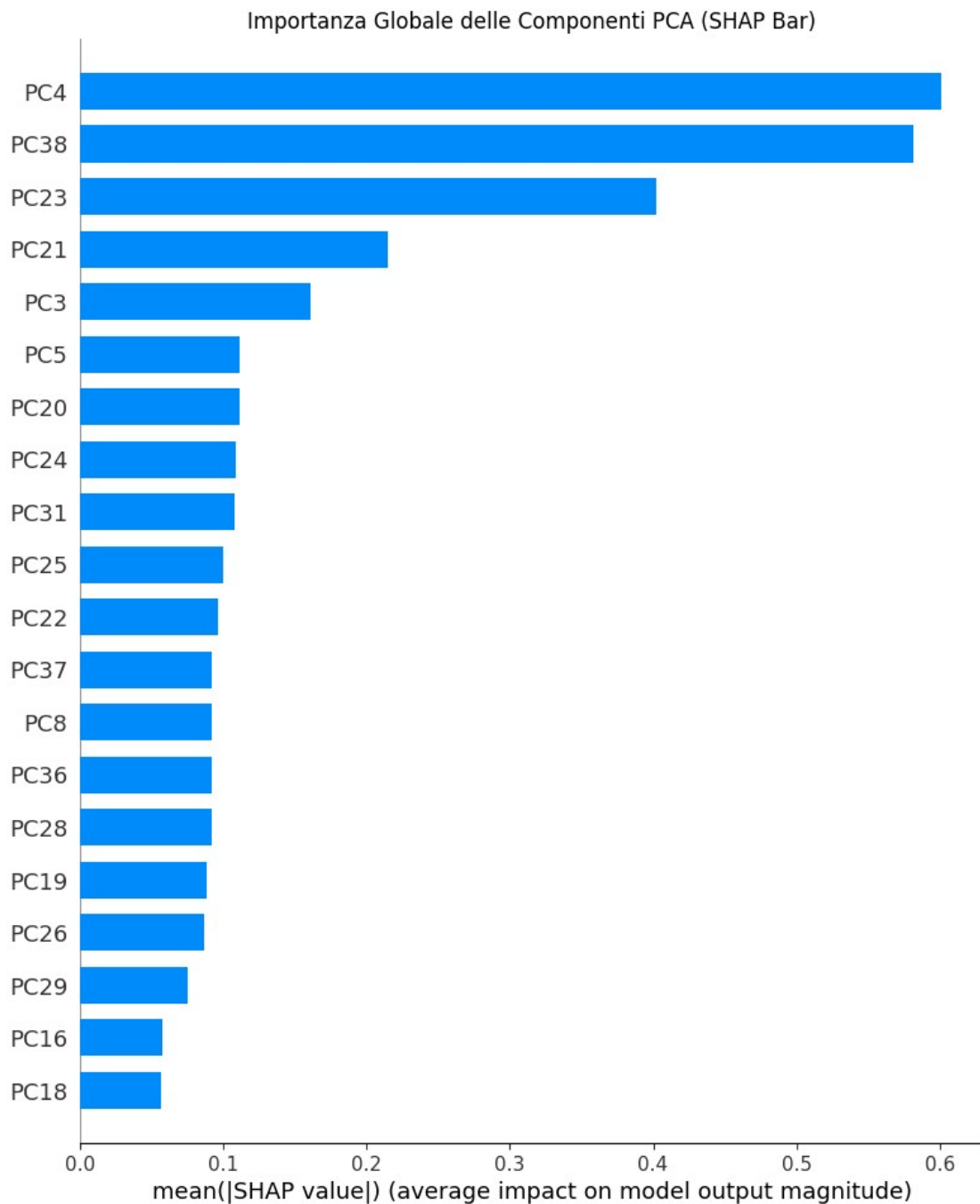
L'uso dei valori SHAP applicati alla Logistic Regression consentono di interpretare in modo dettagliato il contributo delle singole componenti PCA alla decisione del modello.

Il grafico a barre dell'importanza globale (mean |SHAP value|) conferma quantitativamente quanto evidenziato successivamente nel summary plot:

PC4 e PC38 dominano nettamente il ranking delle componenti più importanti.

Le prime 4-5 componenti spiegano una quota significativa dell'impatto totale sul modello, mentre le restanti contribuiscono in modo via via decrescente.

Questo risultato suggerisce che, nonostante l'uso della PCA, il modello mantiene una struttura decisionale relativamente concentrata, potenzialmente utile anche per future riduzioni dimensionali.

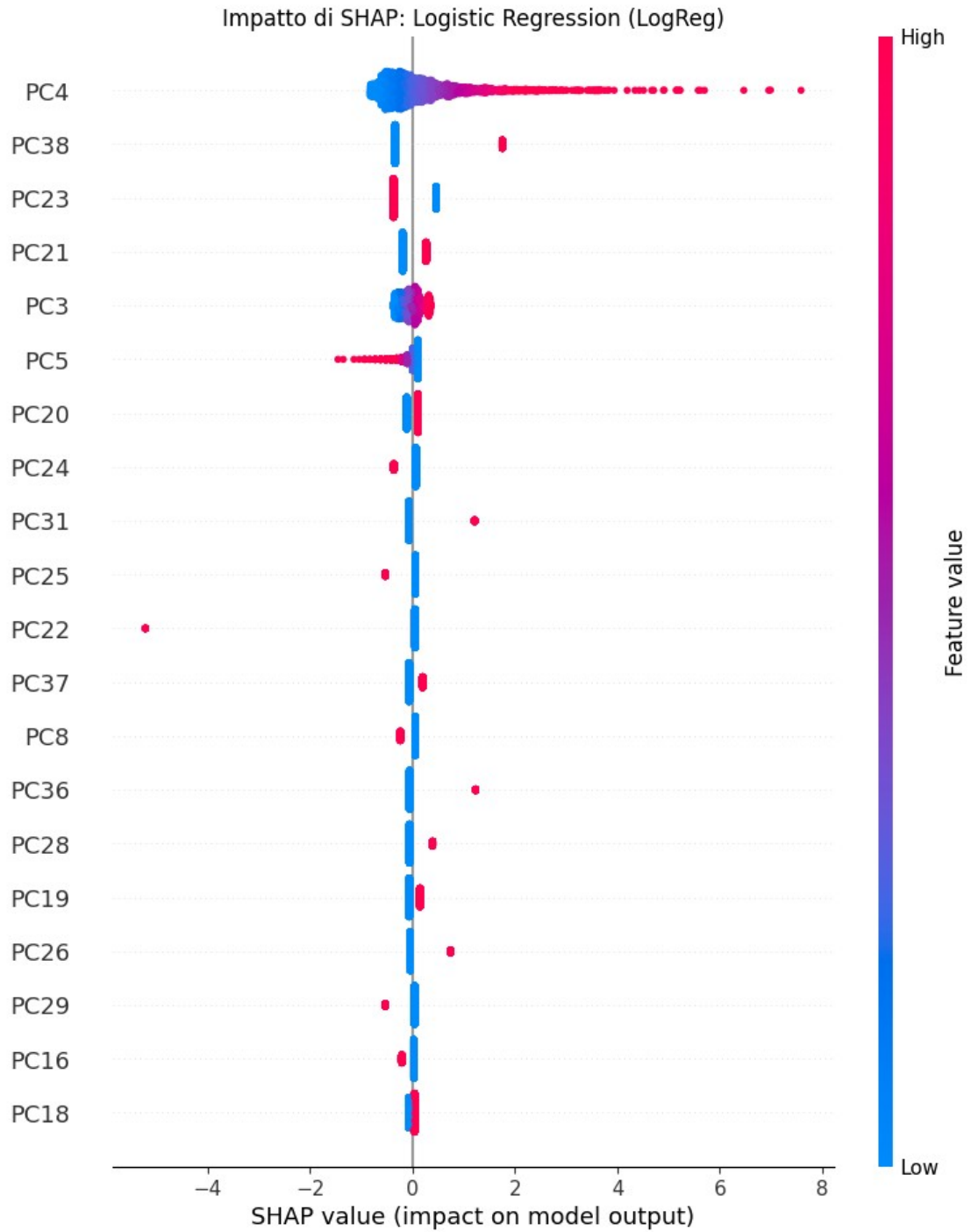


Nello SHAP Summary Plot (Beeswarm) la PC4 emerge come la variabile più influente: valori elevati di questa componente contribuiscono fortemente ad aumentare la probabilità della classe positiva, come evidenziato dall'ampia dispersione dei valori SHAP positivi.

PC38, PC23 e PC21 hanno anch'esse un impatto rilevante, seppur inferiore rispetto a PC4, confermando che l'informazione discriminante non è concentrata su una sola dimensione ma distribuita su più componenti.

Le componenti con valori SHAP prossimi allo zero (es. PC16, PC18) hanno un'influenza marginale sul modello, suggerendo un contributo limitato alla previsione finale.

Il gradiente cromatico (da blu a rosso) indica inoltre una relazione coerente tra valore della feature e direzione dell'impatto, rafforzando l'affidabilità delle interpretazioni.



Conclusioni

Il progetto ha seguito un workflow completo di data science, dalla preparazione dei dati alla spiegazione delle decisioni dei modelli.

Il Random Forest è il modello con le migliori prestazioni predittive, ma presenta una minore interpretabilità.

Logistic Regression, pur leggermente meno performante, offre un eccellente compromesso tra accuratezza e spiegabilità, come dimostrato dall'analisi SHAP.

L'integrazione tra curva ROC e SHAP consente non solo di valutare quanto il modello funziona bene, ma anche perché prende determinate decisioni.

Alla luce di questi risultati, la Logistic Regression con supporto SHAP rappresenta una scelta particolarmente solida in contesti applicativi in cui la trasparenza del modello è un requisito fondamentale, mentre la Random Forest risulta preferibile quando l'obiettivo primario è massimizzare la performance predittiva.