Analisi Predittiva delle Malattie Cardiache

Federico M. Longo, Kyle Pujanes, Matteo Tacchini

Novembre 2025

1 Introduzione

Le malattie cardiovascolari rappresentano una delle principali cause di mortalità a livello globale. Il rilevamento precoce dei fattori di rischio e la predizione dell'insorgenza di una futura patologia cardiaca risulta dunque essere importante in modo da poterne migliorare la prevenzione nonché trattamento.

L'obbiettivo principale di questo studio consiste nello sviluppo di un modello predittivo in grado di stimare, sulla base dei parametri clinici e demografici dei paziente, il rischio relativo di malattia cardiaca.

Il modello è stato addestrato con un dataset che presenta 918 righe e 12 colonne e che comprende sia variabili numeriche che variabili categoriche. Ogni singola colonna rappresenta un attributo, ovvero *feature* che è stato brevemente descritto sotto:

1. **Età (Age)** - il rischio cardiovascolare aumenta con l'età, per questo motivo è un fattore di rischio fondamentale per le malattie cardiache;
2. **Sesso (Sex)** - la variabile è stata codificata come 0 per femmina (F) e 1 per maschio (M);
3. **Dolore Toracico (Chest Pain)** - misura il livello di disagio toracico percepito dal soggetto;
4. **Pressione Sanguigna (Resting Blood Pressure)** - misurata a riposo ed è espressa in $mmHg$;
5. **Colesterolo (Cholesterol)** - indica il livello lipidico nel sangue del paziente;
6. **Glicemia a digiuno (Fasting Blood Sugar)** - variabile codificata come maggiore di $120\frac{mg}{dl}$ (1) o inferiore (0);
7. **Elettrocardiogramma a riposo (Resting ECG)**
8. **Frequenza Cardiaca Massima (Max Heart Rate)**

9. **Dolore da sforzo (Exersice Angina)** - livello di dolore toracico percepito dal paziente durante l'attività fisica;

10. **Old Peak** - depressione del tratto ST, ovvero misura della depressione durante il test da sforzo a riposo. Valori maggiori indicano un aumento del rischio di malattia cardiaca;

11. **Slope** - pendenza del tratto ST, misura della pendenza del tratto ST durante il test da sforzo;

12. **Presenza di Malattia Cardiaca (Heart Disease Presence)** - variabile target che indica la presenza di malattia cardiaca. E' stata codificata come 0 (assenza) o 1 (presenza).

2 Data Pre-Processing

L'analisi preliminare dei dati ha evidenziato assenza di dati duplicati e valori assenti, motivo per il quale non è stato eseguito alcuna rimozione o imputazione.

Per l'uniformità dei dati, le variabili numeriche sono state normalizzate tramite standardizzazione z-score, con l'intento di rendere confrontabili le diverse scale dei valori, mentre le variabili categoriche, come anticipato durante la breve descrizione sopra, sono state codificate in modo da poter essere utilizzabili dal modello.

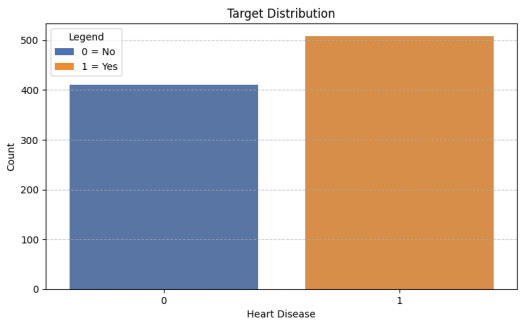

Figure 1: Grafico della distribuzione del target

La distribuzione della variabile target *HeartDisease* nel dataset risulta essere bilanciata, ovvero che la percentuale relativa di ciascuna classe (*pazienti malati* e *pazienti non malati*) rispetto al totale dei pazienti indica che nessuna delle due classi è predominante e pertanto considerabile bilanciata, riducendo così i rischi associati al bias nella classificazione del modello.

3 Scelta del Modello

Considerata la struttura del dataset, con un target binario che indica la presenza o assenza di malattia cardiaca, sono stati considerati i diversi algoritmi di classificazione supervisionata tra cui: Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, AdaBoost, Gradient Boosting e Naive Bayes.
La valutazione di ciascuna è stata effettuata adoperando la ***k-fold cross-validation***, con $k = 5$, analizzando le curve ***ROC*** con le relative ***AUC*** e tramite le *performance metrics*, ovvero parametri che forniscono informazioni sulle prestazioni di un modello tra cui:

- **Accuracy (Accuratezza)** - proporzione di predizioni corrette sul totale delle osservazioni;
- **Precision (Precisione)** - proporzione di veri positivi rispetto al totale dei casi predetti come positivi;
- **Recall (Sensibilità)** - proporzione di veri positivi rispetto al totale dei casi effettivamente positivi;
- **F1-score** - media armonica tra precision e recall.

3.1 Risultati della valutazione dei modelli

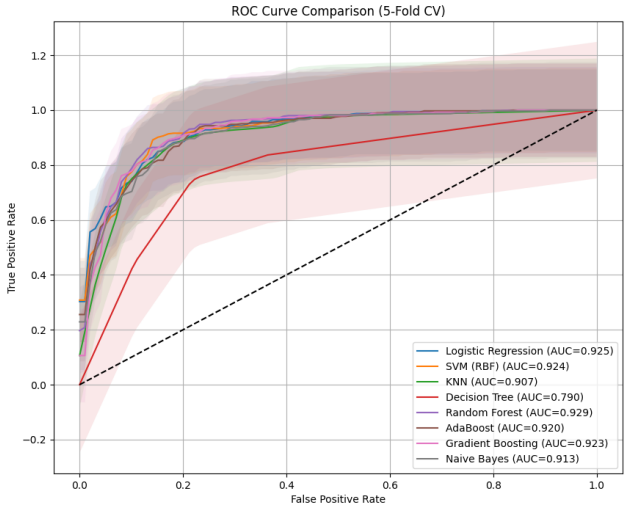

Figure 2: Confronto delle curve ROC con bande di variabilità (5-fold cross-validation). Le aree sfumate indicano la deviazione standard delle curve ROC al variare dei fold

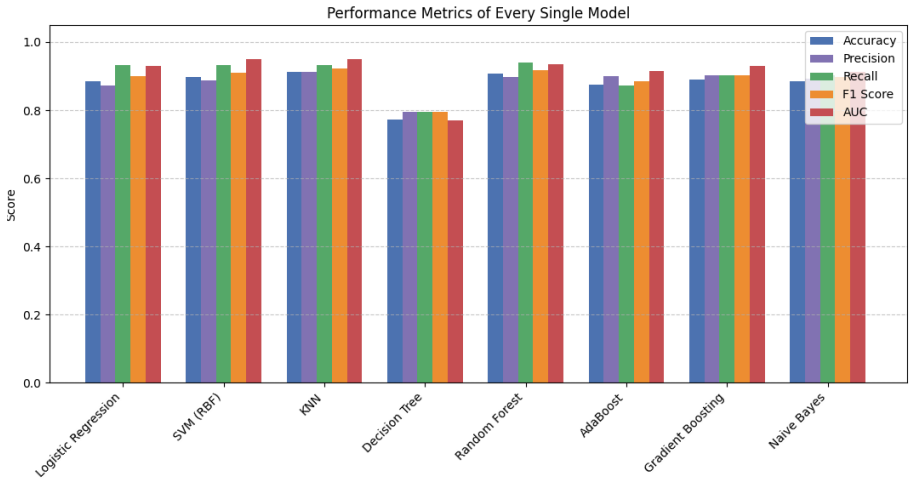

Figure 3: Grafico comparativo delle metriche di performance

I valori finali delle metriche sono riportati nella tabella seguente:

	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.886	0.872	0.931	0.900	0.930
SVM (RBF)	0.897	0.888	0.931	0.909	0.949
KNN	0.913	0.913	0.931	0.922	0.950
Decision Tree	0.772	0.794	0.794	0.794	0.769
Random Forest	0.908	0.897	0.941	0.919	0.934
AdaBoost	0.875	0.899	0.873	0.886	0.915
Gradient Boosting	0.891	0.902	0.902	0.902	0.930
Naive Bayes	0.886	0.893	0.902	0.898	0.912

Table 1: Tabella riassuntiva delle metriche di performance dei modelli

Dai risultati ottenuti si evidenzia che:

- KNN, k-Nearest Neighbors è il modello con miglior accuratezza, precisione, F1-score e AUC;
- SVM (RBF), Support Vector Machine (Radial Basis Function), presenta invece la seconda AUC più slta;
- Random Forest, modello migliore in base alla Recall.

4 Conclusioni

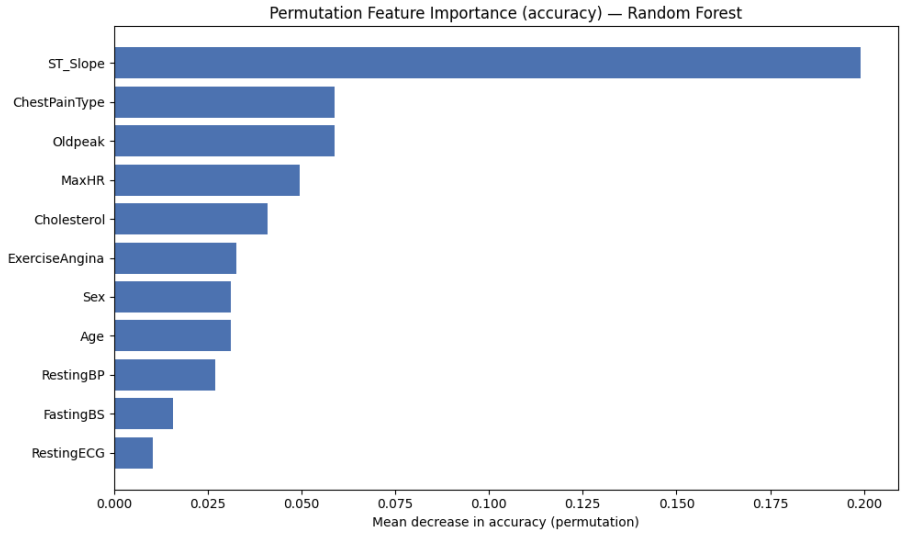

Figure 4: Grafico che riporta la classifica delle feature: è stato misurato l'impatto che ha una feature nelle predizioni del modello

Nonostante il modello KNN presenti valori numericamente superiori sulle metriche, si consiglia di utilizzare la ***Random Forest Classifier***, motivata da:

- **Miglior Recall (0.941)**, in ambito clinico-sanitario è fondamentale che il modello sia in grado di riconoscere effettivamente un paziente malato, ossia che i **falsi negativi** siano minimizzati;
- **Robustezza al rumore e agli *outlier***, variando dati il modello deve mantenere le prestazioni o al massimo fare piccole oscillazioni;
- Capacità di gestire sia dati numerici che categorici senza bisogno di trasformazioni complesse;
- **Interpretabilità** grazie alle feature importance.

La stabilità di questo modello lo rende dunque preferibile rispetto ad altri che forniscono valori di prestazioni elevati, come appunto la KNN.

In ultima analisi, è possibile dunque dire che modelli di Machine Learning possono **supportare** in maniera efficace il medico curante per diagnosticare preventivamente il rischio di malattie cardiovascolari sulla base di dati clinico-sanitari.

5 Feedback

Grazie simone per avermi learnato un po di shit sulle AI. Avrei preferito fare più ore e trattare materie diverse con lei. Sai spiegare molto bene, con ottimi esempi, facili da capire e semplici. Sei stato molto coinvolgente, anche se sei arrivato alla fase finale; capisco che dipenda dagli impegni. Ti do un voto 8 su diesciiii. Se diventerai docente, fammi sapere dove, così potrò mandare lì mio figlio.Avrei voluto approfondire ancora di più con Python , in particolare aspetti più avanzati.

- Federico M. Longo

I professori che mi hanno effettivamente insegnato qualcosa sono pochi, e posso dire con certezza che Lei ne fa parte. Più che la materia, che ho sempre trovato affascinante ed interessante da approfondire, provo tanta ammirazione per il tipo di docente che si presenta: un professore senza bias (o quasi ;)), sincero e appassionato, non solamente della materia bensì anche del voler traasmettere il proprio knowledge a chiunque. (sono le stesse cose che cerco di trasmettere anch'io ai ragazzi a cui faccio ripetizioni). Tanta stima e tante grazie Prof. Rancati, futuro relatore? (HAHAHAHHA JK..... or maybe not... who knows? I just hope to cross roads with you one day)

- Kyle V. Pujanes

Bella prof. Ci ho provato. Prometto che non salto più durante le sue ore.

- Matteo Tacchini