

HeartRisk AI

Un sistema intelligente per la predizione delle malattie cardiache, basato su Machine Learning e metriche robuste.

Di:

Link al progetto:

- Github: <https://github.com/an70nn/HeartRisk-AI>

Indice

Introduzione	2
Analisi del Dataset	3
Pre-Elaborazione dei dati	5
Analisi e selezione delle Feature	5
Correlazioni con la Variabile Target	6
Correlazioni tra Feature Indipendenti (Multicollinearità)	6
Implicazioni per la Pre-elaborazione	6
Analisi delle distribuzioni delle Feature Numeriche.....	7
Analisi delle correlazioni Feature-Target	9
Apprendimento supervisionato	12

Introduzione

Il progetto **HeartRisk AI** nasce con l'obiettivo di sviluppare un sistema predittivo per la **diagnosi precoce delle malattie cardiache**. Il modello si basa sul dataset *Cleveland Heart Disease* della UCI Machine Learning Repository, scelto per la sua rilevanza clinica, l'ampio utilizzo accademico, la varietà di feature numeriche e categoriche, e la presenza di un target binario già strutturato (presenza o assenza di malattia), che lo rende adatto a classificatori come la Regressione Logistica.

I principali obiettivi del progetto sono:

- **Predire la presenza di malattie cardiache**
Utilizzare un modello di apprendimento supervisionato per classificare i pazienti affetti da patologie cardiache sulla base delle loro caratteristiche cliniche e demografiche.
- **Applicare tecniche di Explainable AI (XAI)**
Integrare visualizzazioni avanzate, come le heatmap del **False Positive Rate** per identificare eventuali **bias** nei confronti di specifici sottogruppi (es. per sesso o tipo di dolore toracico).
- **Testare la robustezza del modello**
Analizzare la sensibilità del modello a variazioni controllate delle feature (approccio Adversarial Machine Learning), per stimarne la stabilità predittiva.
- **Supportare la diagnosi precoce e le decisioni cliniche**
Fornire uno strumento predittivo di supporto alla diagnosi cardiologica, utile per attività di screening o valutazioni preliminari a basso costo.

La metodologia adottata si articola nelle seguenti fasi:

- **Raccolta e Preprocessing dei dati:** Raccolta del dataset e pulizia dei dati, gestione dei valori mancanti, conversione delle variabili categoriche tramite encoding e standardizzazione delle feature numeriche.
- **Addestramento modello:** Utilizzo della Regressione Logistica per addestrare un classificatore binario sulla presenza o assenza di malattia cardiaca.
- **Valutazione delle prestazioni:** Analisi delle performance tramite metriche standard (accuracy, classification report, matrice di confusione) e confronto con la baseline. Uso di Heatmap del False Positive Rate per l'analisi dei BIAS del modello su sottogruppi specifici.
- **Test di robustezza (Adversarial ML):** Valutazione della sensibilità del modello a variazioni controllate di feature selezionate per testarne la stabilità predittiva.
- **Interfaccia utente CLI:** Implementazione di un menu testuale per facilitare l'esplorazione dei dati, il testing del modello e l'analisi interattiva.

Analisi del Dataset

Le informazioni riguardanti il dataset sono reperibili al seguente [link](#) in maniera dettagliata.

Le informazioni principali del dataset sono:

- **Titolo:** Heart Disease Dataset
- **Fonte:** UCI Machine Learning Repository
- **Autori:** Andras Janosi, William Steinbrunn, Matthias Pfisterer, Robert Detrano
- **Data di pubblicazione:** 1988
- **Numero di record:** 303
- **Numero di feature:** 14 (13 predittive + 1 target)
- **Target:** presence of heart disease (valore binarizzato: 0 = assente, 1 = presente)

Le variabili del Dataset possono essere suddivise in tre macrocategorie:

Caratteristiche demografiche:

Feature	Descrizione	Tipo	Valore
Age	Età del paziente	Numerica	Valori numerici (in anni)
Sex	Sesso biologico	Categoriale	0 = Maschio / 1 = Femmina

Indicatori clinici e Sintomi:

Feature	Descrizione	Tipo	Valore
Cp	Tipo di dolore toracico	Categoriale	0 = Angina tipica, 1 = Angina atipica, 2 = Non-anginosa, 3 = Asintomatico
Trestbps	Pressione arteriosa a riposo	Numerica	Valori numerici
Chol	Colesterolo sierico (mg/dl)	Numerica	Valori numerici
FBS	Glicemia a digiuno > 120 mg/dl	Categoriale	0 = No, 1 = Sì
Restecg	Risultato ECG a riposo	Categoriale	0 = Normale, 1 = Anormalità, 2 = Ipertrofia ventricolare SX
Thalach	Frequenza cardiaca massima raggiunta	Numerica	Valori numerici (bpm)
Exang	Angina da sforzo	Categoriale	0 = No, 1 = Sì
Oldpeak	Depressione ST rispetto al riposo	Numerica	Valori numerici
Slope	Pendenza del tratto ST	Categoriale	0 = Decrescente, 1 = Piatta, 2 = Crescente
Ca	Numero di vasi principali colorati (0-3)	Categoriale	Valori numerici 0, 1, 2, 3
Thal	Tipo di Talassemia	Categoriale	3 = Normale, 6 = Difetto fisso, 7 = Difetto reversibile

Variabile Target:

Feature	Descrizione	Tipo	Valore
Target	Presenza di malattia cardiaca	Categoriale (binaria)	0 = Assente, 1 = Presente

Osservazioni: Durante la fase di analisi del progetto, inizialmente non è stato consultato il *paper introduttivo* associato al dataset "Cleveland Heart Disease". Tuttavia, una successiva revisione ha evidenziato elementi rilevanti che meritano di essere integrati. Il dataset originale comprendeva ben 76 attributi, ma nella maggior parte delle applicazioni accademiche, tra cui questo progetto, viene utilizzato un sottoinsieme composto da 14 feature selezionate (noto come "Cleveland subset"), considerate le più significative dal punto di vista clinico.

Un'altra osservazione fondamentale riguarda la variabile target: inizialmente rappresentava cinque livelli di gravità della malattia (da 0 a 4), ma è stata convertita in una variabile binaria (0 = assenza, 1 = presenza di malattia) per rendere più chiara e gestibile la classificazione. Questa decisione, anche se comune in letteratura, comporta una semplificazione del problema clinico, utile però per modelli predittivi di primo livello o sistemi di screening preliminare.

#Crucius su ChatGPT (Mi spiegherà cose sull' Heatmap e Correlazione), salvato tra l'altro anche nella memoria interna di GPT.

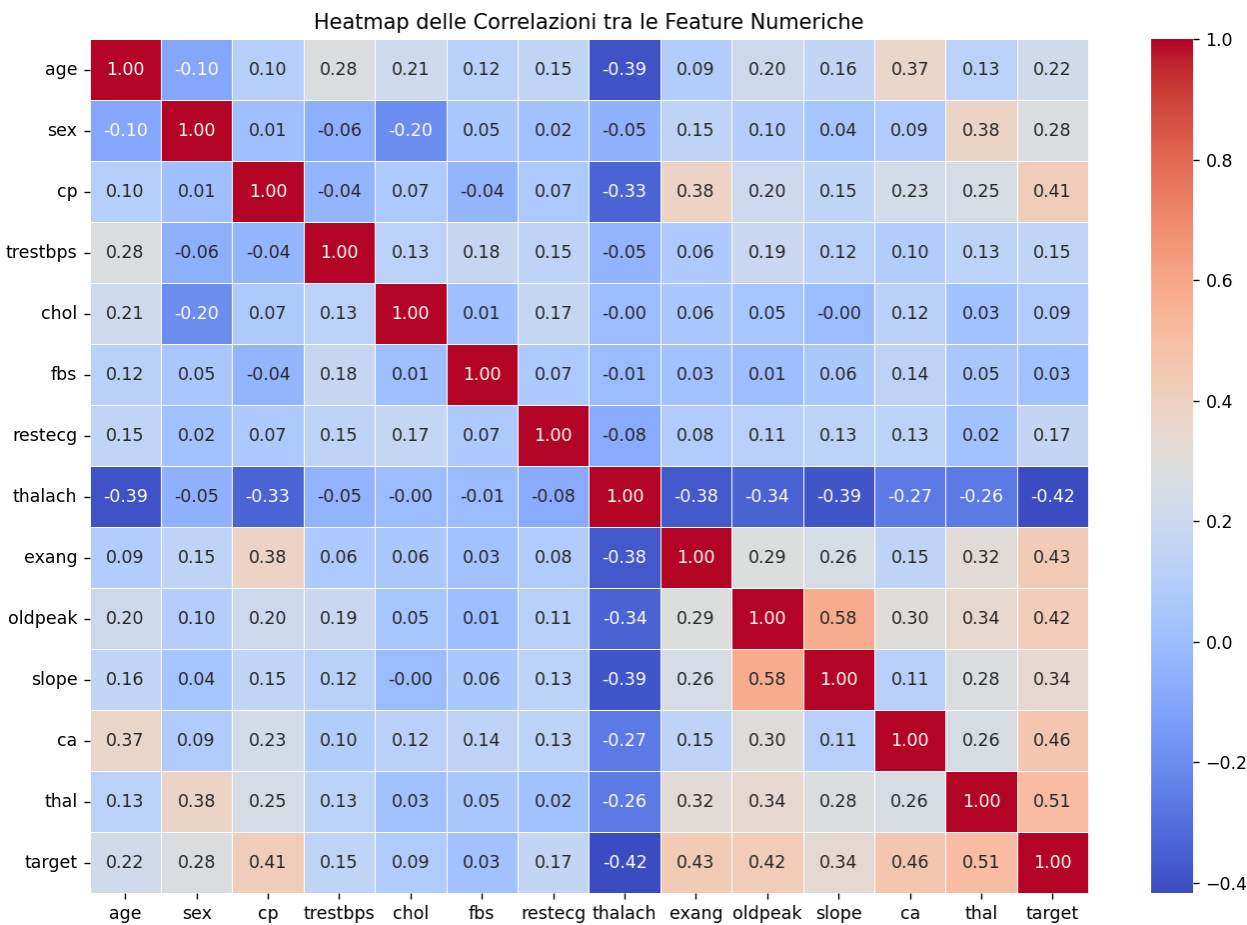
Pre-Elaborazione dei dati

Analisi e selezione delle Feature

La selezione e la comprensione delle feature rappresentano un passaggio cruciale per garantire che i modelli predittivi siano accurati ed efficienti. In questa fase, è stata effettuata un'analisi esplorativa del dataset **Cleveland Heart Disease** con i seguenti obiettivi:

1. **Identificare le feature più correlate** con la variabile target (target), indicativa della presenza o assenza di malattia cardiaca;
2. Evidenziare eventuali **correlazioni forti tra feature indipendenti**, che potrebbero suggerire ridondanze o problematiche di multicollinearità;
3. **Valutare la distribuzione e la trasformazione** dei dati e la necessità di normalizzazione, standardizzazione o encoding delle variabili.

Per raggiungere questi obiettivi, è stata calcolata la **matrice di correlazione** tra le variabili numeriche e categoriche (già binarizzate e trattate come numeriche) presenti nel dataset. I risultati sono stati poi visualizzati tramite una **heatmap delle correlazioni**, che permette di osservare rapidamente i pattern e l'intensità della correlazione positiva (valori vicini a +1, colore rosso) e negative (valori vicini a -1, colore blu).



Dall'analisi della Heatmap delle Correlazioni emergono diverse osservazioni chiave:

Correlazioni con la Variabile Target

Le feature **thal** (tipo di talassemia), **ca** (numero di vasi colorati), **oldpeak** (depressione ST), **exang** (angina da sforzo), **cp** (tipo di dolore toracico) e **thalach** (frequenza cardiaca max) mostrano le correlazioni più forti con il target.

- **thal** e **ca** presentano le correlazioni positive più elevate (vicino a 0.5), indicando che valori più alti di queste feature sono fortemente associati alla presenza di malattia cardiaca.
- **thalach** mostra una correlazione negativa significativa (circa -0.42), suggerendo che una frequenza cardiaca massima più alta è associata all'assenza di malattia (o viceversa).
- **cp** mostra una correlazione positiva (circa 0.41), indicando che certi tipi di dolore toracico (come Angina atipica o Non-anginosa, se il tuo encoding lo permette) sono più associati alla malattia.

Altri feature come **age**, **sex**, **trestbps**, **chol**, **fbs**, **restecg**, **slope** mostrano correlazioni meno marcate con il target, suggerendo un impatto predittivo minore ma comunque rilevante.

Correlazioni tra Feature Indipendenti (Multicollinearità)

Si osservano alcune correlazioni moderate tra le feature indipendenti, ad esempio:

- **thalach** e **exang** (-0.43): C'è una correlazione negativa tra la frequenza cardiaca massima e la presenza di angina da sforzo, il che è clinicamente atteso.
- **oldpeak** e **slope** (0.58): La depressione del segmento ST è correlata alla pendenza del tratto ST, anch'esso atteso.

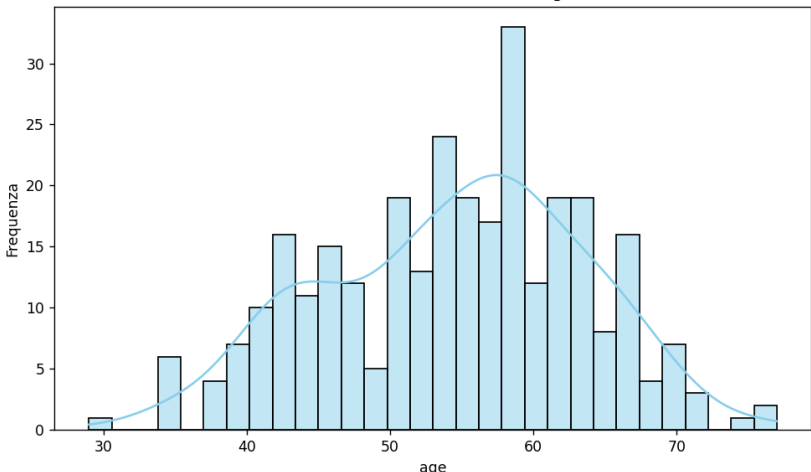
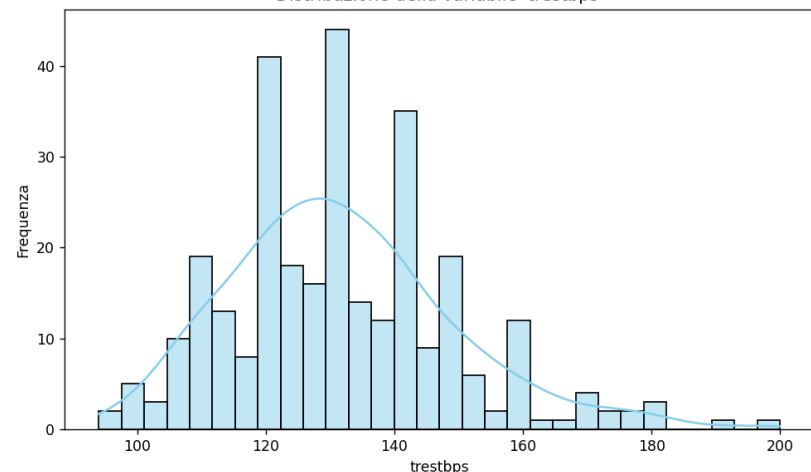
Implicazioni per la Pre-elaborazione

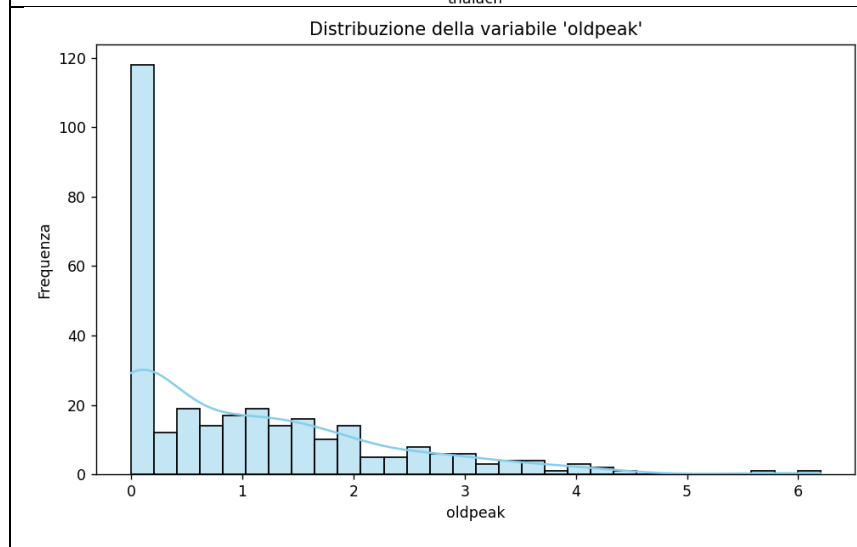
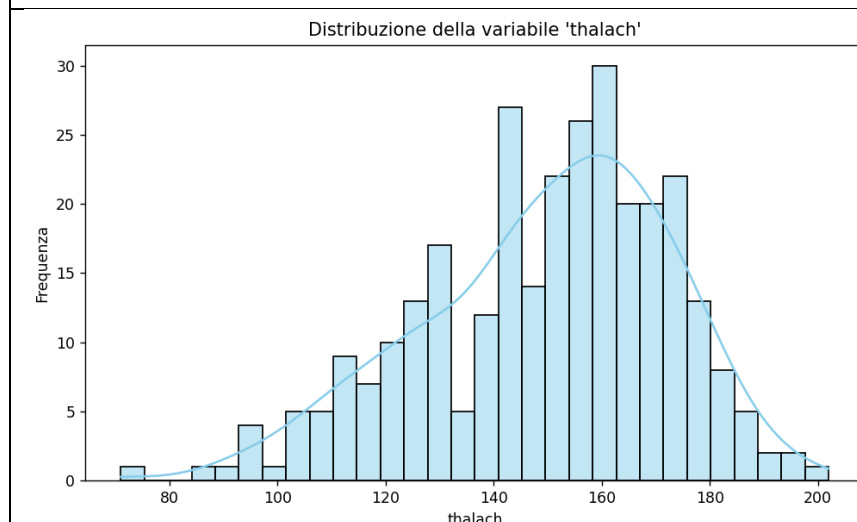
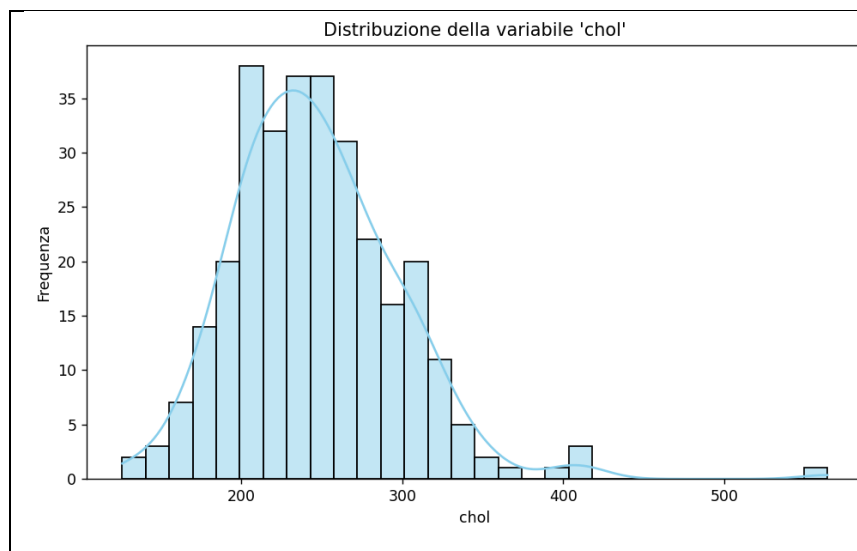
L'analisi delle correlazioni ha rafforzato la necessità di trattare adeguatamente le feature numeriche (tramite **Standard Scaler**) e quelle categoriche (tramite **One-Hot Encoding**), come implementato nella successiva fase di pre-elaborazione. Questo garantisce che tutte le feature contribuiscano in modo equo al modello e che quelle categoriche siano interpretate correttamente.

Analisi delle distribuzioni delle Feature Numeriche

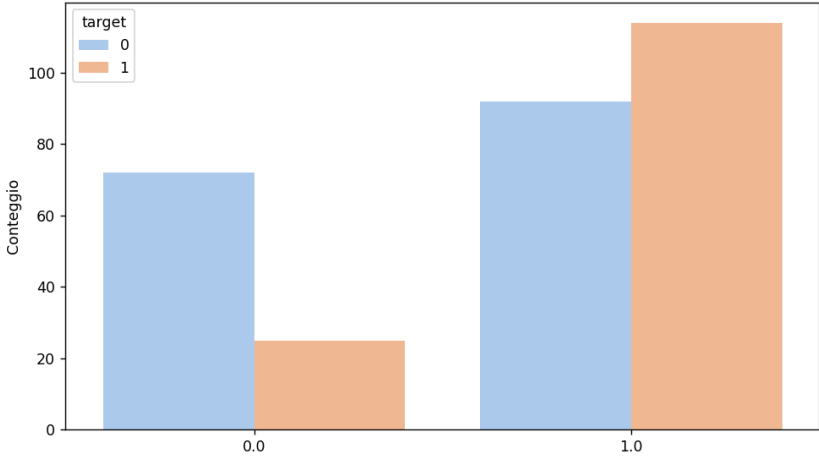
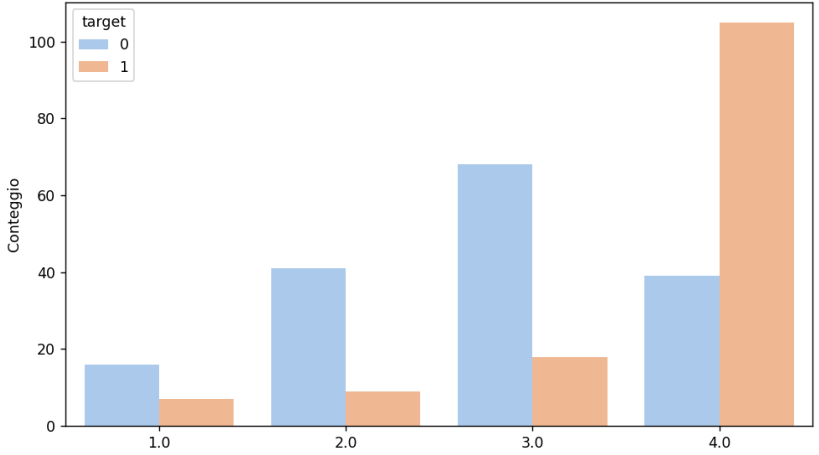
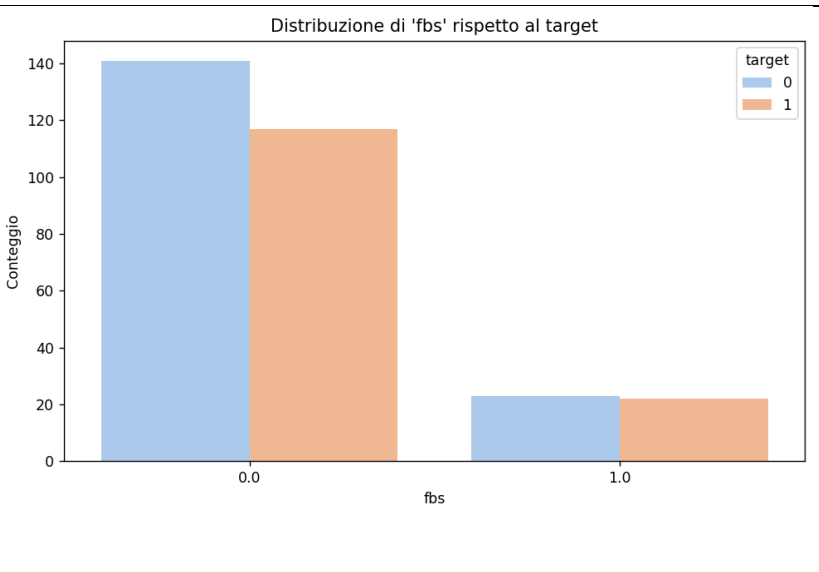
Per comprendere meglio la struttura dei dati e individuare eventuali anomalie o squilibri nelle variabili numeriche, è stata condotta un'analisi esplorativa dei principali features quantitative presenti nel dataset. A tal fine, sono stati utilizzati **istogrammi affiancati a curve di densità (KDE)** per ciascuna variabile. Questo tipo di visualizzazione consente di osservare:

- La forma della distribuzione (simmetrica, asimmetrica, normale...);
- La presenza di **outlier** o accumuli sospetti di dati;

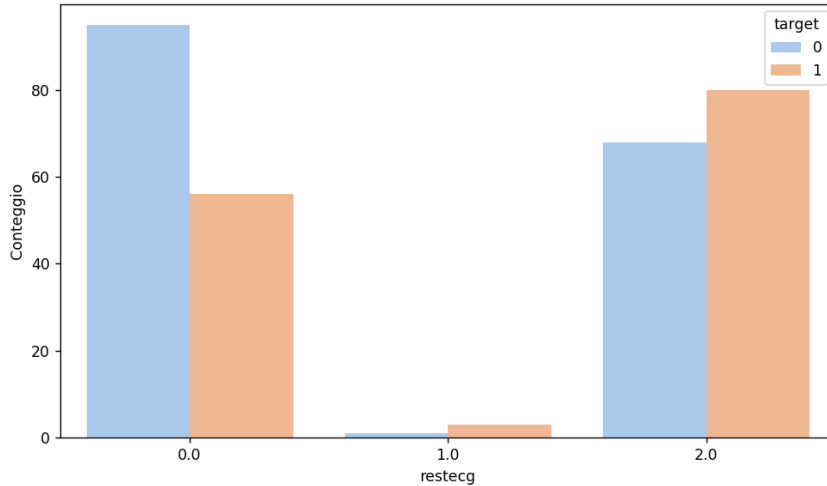
Istogramma	Risultato
<p>Distribuzione della variabile 'age'</p> 	<p>La distribuzione dell'età si presenta simmetrica, con una forma che si avvicina a una distribuzione normale (a campana). I dati si estendono dai 30 ai 75 anni, con una concentrazione maggiore tra i 45 e i 65 anni. Questo riflette la fascia d'età tipica in cui le malattie cardiache tendono a manifestarsi o a essere diagnosticate. Non ci sono Outlier estremi che suggeriscono anomalie nella raccolta delle età; i valori più esterni sono presenti in minor numero ma rientrano nel range atteso.</p>
<p>Distribuzione della variabile 'trestbps'</p> 	<p>La distribuzione della pressione arteriosa a riposo tende ad essere moderatamente asimmetrica a destra (positiva). La maggior parte delle osservazioni si concentra intorno a valori tra 120 e 140 mmHg, che sono intervalli comuni per la pressione sistolica. I valori di "trestbps" vanno da circa 90 mmHg fino a oltre 180 mmHg, con alcuni valori che si estendono fino a 200 mmHg, sebbene rari. La concentrazione maggiore è tra 110 e 150 mmHg. Potenziati Outlier rappresentati come casi di ipertensione grave si notano nelle barre di frequenza molto basse alle estremità superiore della distribuzione (es. oltre 170-180 mmHg).</p>



Analisi delle correlazioni Feature-Target

Grafico a barre	Risultato															
<p>Distribuzione di 'sex' rispetto al target</p>  <table data-bbox="95 329 917 790"><tr><th>sex</th><th>target 0</th><th>target 1</th></tr><tr><td>0.0</td><td>72</td><td>25</td></tr><tr><td>1.0</td><td>92</td><td>115</td></tr></table>	sex	target 0	target 1	0.0	72	25	1.0	92	115	<p>Ricordando la mappatura standard:</p> <ul style="list-style-type: none">• Sex = 1, Uomo• Sex = 0, Donna <p>Questa visualizzazione evidenzia una forte correlazione tra la variabile 'sex' e il 'target'. Sembra che, nel contesto di questo dataset, il sesso maschile sia associato a una maggiore incidenza di malattie cardiache.</p>						
sex	target 0	target 1														
0.0	72	25														
1.0	92	115														
<p>Distribuzione di 'cp' rispetto al target</p>  <table data-bbox="95 846 917 1299"><tr><th>cp</th><th>target 0</th><th>target 1</th></tr><tr><td>1.0</td><td>16</td><td>7</td></tr><tr><td>2.0</td><td>41</td><td>9</td></tr><tr><td>3.0</td><td>68</td><td>18</td></tr><tr><td>4.0</td><td>39</td><td>105</td></tr></table>	cp	target 0	target 1	1.0	16	7	2.0	41	9	3.0	68	18	4.0	39	105	<p>Ricordando la mappatura standard:</p> <ul style="list-style-type: none">• CP = 1.0 (Angina Tipica)• CP = 2.0 (Angina Atipica)• CP = 3.0 (Dolore non Anginoso)• CP = 4.0 (Asintomatico) <p>Questo suggerisce che non tutti i dolori toracici sono ugualmente indicativi di malattia, e anzi, l'assenza di sintomi (Asintomatico o CP=4) può essere un indicatore più critico di quanto atteso intuitivamente. La Feature “CP” è un predittore estremamente importante.</p>
cp	target 0	target 1														
1.0	16	7														
2.0	41	9														
3.0	68	18														
4.0	39	105														
<p>Distribuzione di 'fbs' rispetto al target</p>  <table data-bbox="95 1355 917 1924"><tr><th>fbs</th><th>target 0</th><th>target 1</th></tr><tr><td>0.0</td><td>140</td><td>117</td></tr><tr><td>1.0</td><td>23</td><td>22</td></tr></table>	fbs	target 0	target 1	0.0	140	117	1.0	23	22	<p>Ricordando la mappatura standard:</p> <ul style="list-style-type: none">• fbs = 0.0 (Glicemia a digiuno \leq 120 mg/dl) - non a rischio diabete;• fbs = 1 (Glicemia a digiuno $>$ 120 mg/dl) - a rischio diabete. <p>A differenza di altri feature, “FBS” mostra una correlazione meno marcata. Sebbene una glicemia a digiuno elevata sia un fattore di rischio clinico, in questo dataset, non si mostra una netta prevalenza di malattia cardiaca nei pazienti. Questo suggerisce che fbs potrebbe non essere uno dei predittori più forti.</p>						
fbs	target 0	target 1														
0.0	140	117														
1.0	23	22														

Distribuzione di 'restecg' rispetto al target

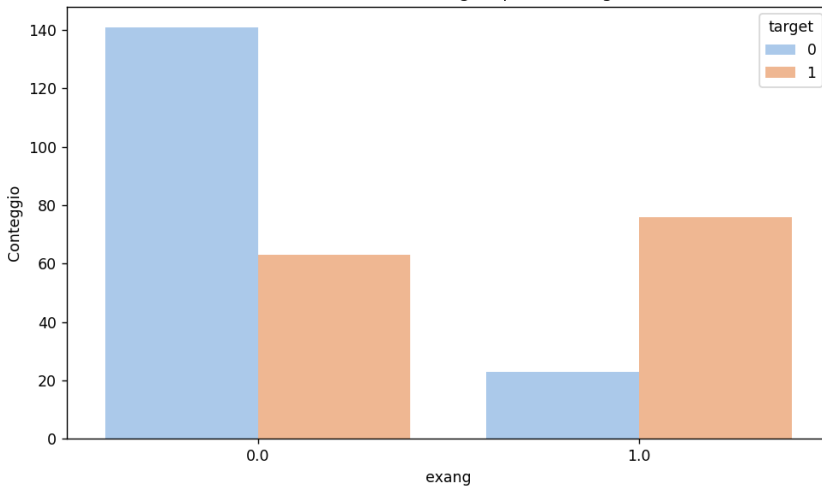


Ricordando la mappatura standard:

- restecg = 0 (Normale)
- restecg = 1 (Anormale)
- restecg = 2 (Ipertrofia ventricolare)

Per restecg = 0, La maggior parte dei pazienti con un ECG a riposo normale **non presenta malattia cardiaca**. Per restecg = 1, dove a causa del numero molto ridotto di osservazioni in questa categoria, è difficile trarre conclusioni. Per il restecg = 2, mostra una forte associazione, ergo tale Feature è un predittore utile.

Distribuzione di 'exang' rispetto al target

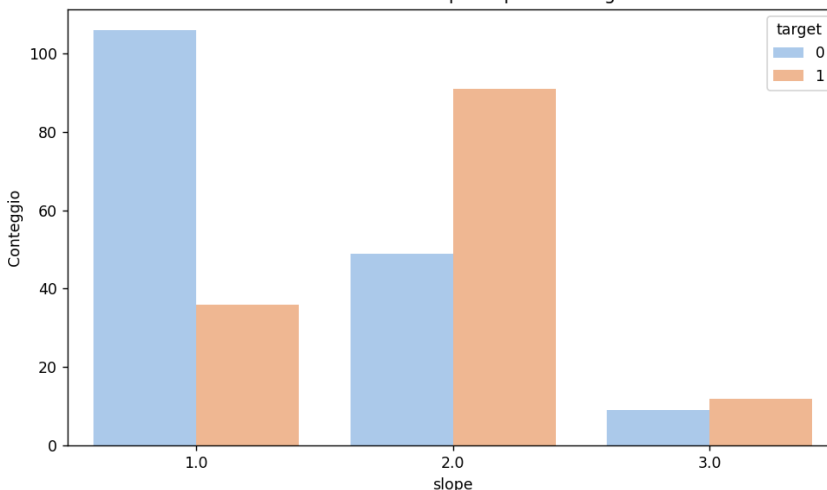


Ricordando la mappatura standard:

- exang = 0 (No Angina da sforzo)
- exang = 1 (Angina da sforzo presente)

La feature exang è un **predittore molto forte e clinicamente rilevante** per la presenza di malattia cardiaca. La presenza di angina da sforzo è fortemente associata a un aumentato rischio di malattia. Questo la rende una delle feature più significative da considerare per il modello predittivo.

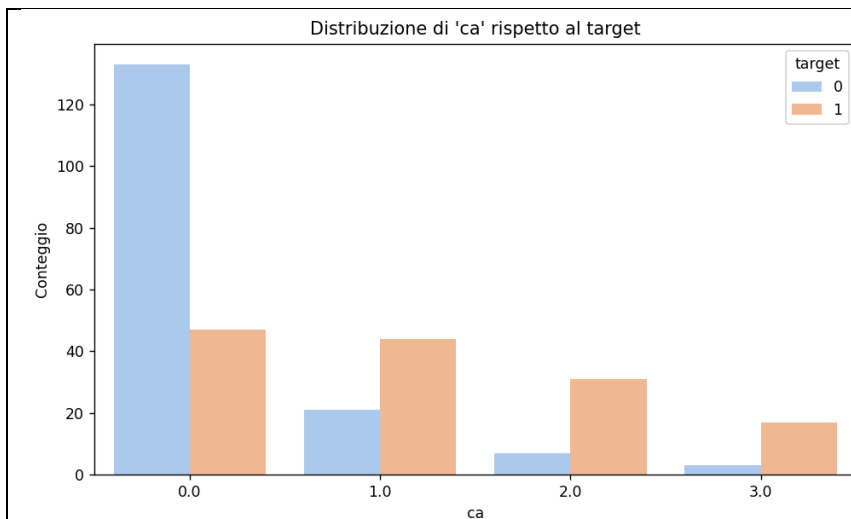
Distribuzione di 'slope' rispetto al target



Ricordando la mappatura standard:

- slope = 1 (Upsloping crescente)
- slope = 2 (Flat)
- slope = 3 (Downsloping decrescente)

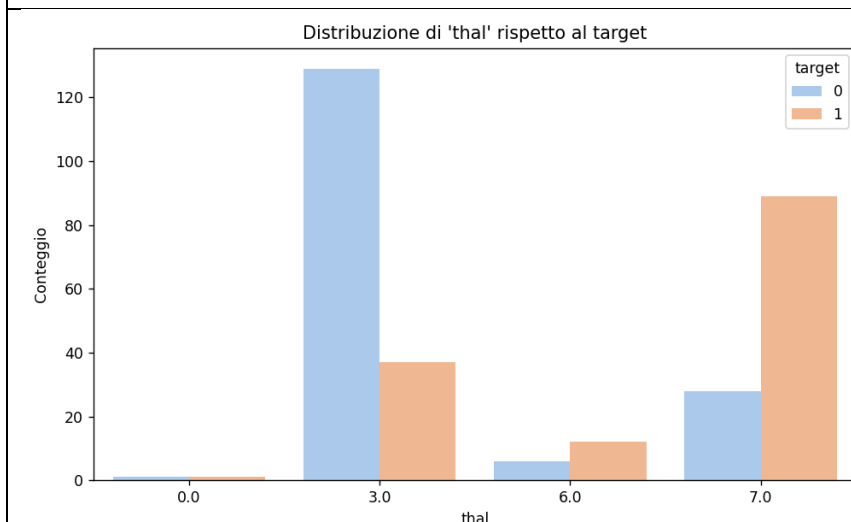
La feature slope è un **predittore molto significativo** per la presenza di malattia cardiaca. Una pendenza del tratto ST piatta (slope=2) o decrescente (slope=3) è fortemente associata a un aumentato rischio di malattia, mentre una pendenza crescente (slope=1) è più comune tra i pazienti sani. Questo la rende una feature di grande valore clinico e predittivo per il modello.



Ricordando la mappatura standard:

- CA = 0 (Nessun vaso colorato)
- CA = 1 (Un vaso colorato)
- CA = 2 (Due vasi colorati)
- CA = 3 (Tre vasi colorati)

La feature ca è un **predittore estremamente forte e uno dei più importanti** per la presenza di malattia cardiaca nel dataset. La sua correlazione positiva con il target è molto evidente e clinicamente significativa, indicando che maggiore è l'estensione dell'ostruzione vascolare, maggiore è la probabilità di malattia. Questa feature avrà un peso elevato nel modello predittivo.



Ricordando la mappatura standard:

- Thal = 0.0
- Thal = 3.0 (Normale)
- Thal = 6.0 (Difetto fisso)
- Thal = 7.0 (Difetto reversibile)

I tipi di Talassemia "difetto fisso" (thal=6.0) e "difetto reversibile" (thal=7.0) sono indicatori molto robusti di malattia. Questo la rende una delle feature più influenti per il modello predittivo, coerente con le sue alte correlazioni osservate nella heatmap. La categoria thal=0.0 è da trattare con cautela a causa della sua scarsa numerosità.

Apprendimento supervisionato

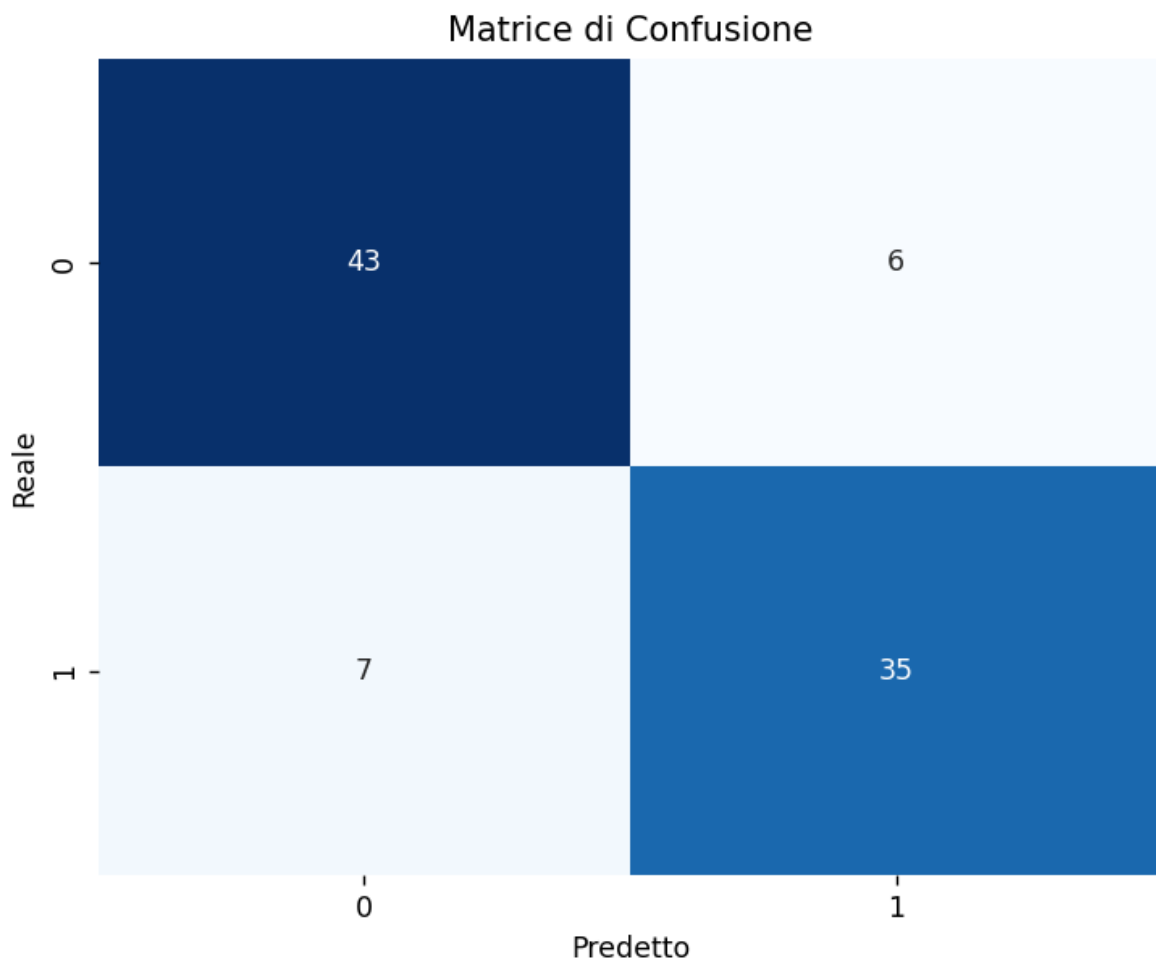
L'apprendimento supervisionato è una tecnica di machine learning in cui un modello viene addestrato su dati etichettati, imparando a predire l'output corretto (ad esempio la presenza o assenza di malattia) a partire dalle caratteristiche di input.

Nel progetto HeartRisk AI, questo approccio è utilizzato per prevedere la presenza di malattie cardiache in base a variabili cliniche e demografiche del paziente, estratte dal dataset Cleveland Heart Disease.

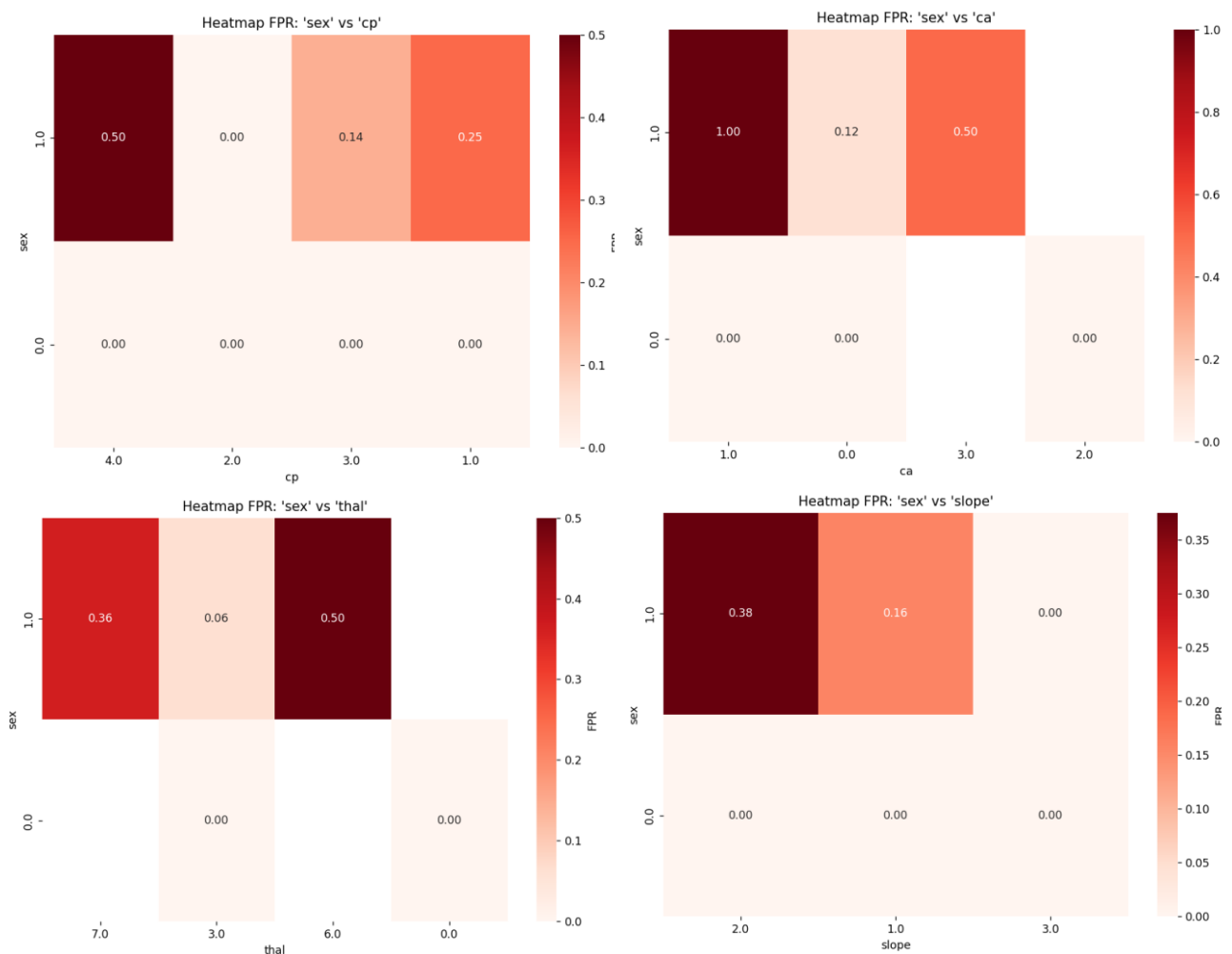
Il dataset è stato suddiviso in due sottoinsiemi distinti: un **training set** per l'addestramento del modello e un **test set** per la valutazione delle sue prestazioni su dati non visti.

Il modello è stato valutato tramite metriche standard come accuratezza, precision, recall e F1-score, oltre alla matrice di confusione, che permette di analizzare più nel dettaglio le prestazioni per le classi positive e negative.

Per l'addestramento, è stata scelta la **Regressione Logistica**, un modello interpretabile e adatto a problemi di classificazione binaria, come il nostro. Sono stati inoltre utilizzati metodi di preprocessing per normalizzare i dati e gestire le variabili categoriche.



Per approfondire la valutazione e il controllo dei bias, sono state create visualizzazioni come la heatmap del False Positive Rate (FPR) incrociando due feature categoriche, utile per identificare possibili disuguaglianze di performance tra sottogruppi.



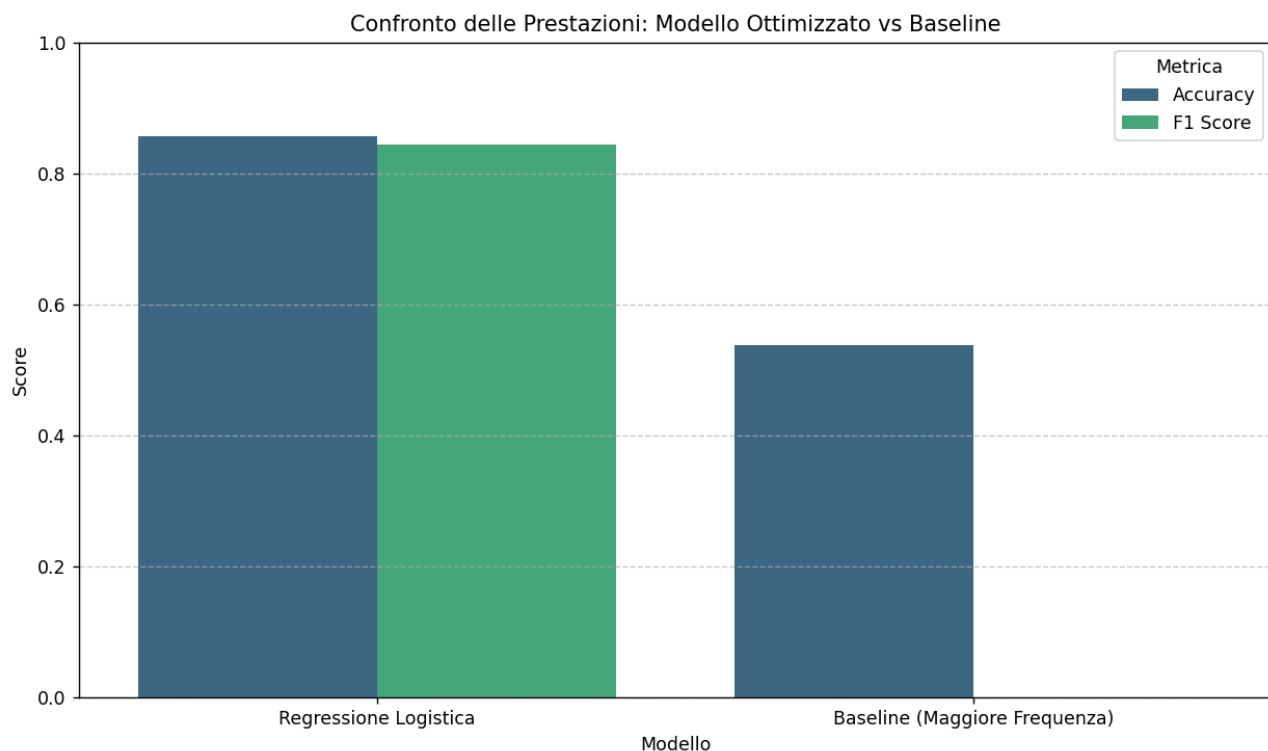
Dall'analisi delle heatmap del False Positive Rate (FPR) incrociando la feature sex (sesso biologico) con diverse altre feature categoriche clinicamente rilevanti (cp, ca, thal, slope), emerge un **chiaro e consistente bias di genere nell'accuratezza del modello**.

In particolare, per sex=0 (Donne), il FPR si attesta a un valore di **0.00 per quasi tutte le categorie** delle varie feature categoriche analizzate. Questo indica che il modello è **quasi perfetto nel non generare falsi positivi** per la popolazione femminile, ovvero non tende a classificare erroneamente donne sane come affette da malattia cardiaca.

Al contrario, per sex=1 (Uomini), il modello presenta un **FPR significativamente più alto** per diverse combinazioni di sottogruppi, suggerendo una maggiore propensione a commettere falsi positivi su pazienti di sesso maschile sani. Questa disparità nelle prestazioni è un aspetto cruciale da considerare per l'equità del modello in un contesto clinico."

Confronto tra modelli

Analizzando le metriche di ogni modello è possibile classificare i modelli in un grafico.



Per valutare l'efficacia del modello di **Regressione Logistica** sviluppato, le sue prestazioni sono state confrontate con quelle di un modello **baseline**, che in questo contesto è stato definito come un classificatore che predice sempre la classe maggioritaria presente nel dataset. Questo confronto è fondamentale per dimostrare il valore aggiunto del nostro modello rispetto a una previsione casuale o non informata.

Il grafico sopra illustra le prestazioni di entrambi i modelli in termini di **Accuratezza** e **F1 Score**.

- **Regressione Logistica Ottimizzata:**
 - **Accuratezza:** Il modello ha raggiunto un'accuratezza di **0.86**. Questo indica che l'86% delle previsioni totali (sia per la presenza che per l'assenza di malattia) sono state corrette.
 - **F1 Score:** Il valore dell'F1 Score per la classe positiva (presenza di malattia) è di **0.84**. Questa metrica, essendo la media armonica di precisione e recall, evidenzia una buona capacità del modello di bilanciare la correttezza delle previsioni positive e la sua completezza nel rilevare tutti i casi positivi. Un valore elevato suggerisce che il modello ha sia pochi falsi positivi che pochi falsi negativi.
- **Baseline (Maggiore Frequenza):**
 - **Accuratezza:** La baseline ha ottenuto un'accuratezza di circa **0.53**. Questo valore è direttamente correlato alla proporzione della classe più frequente nel dataset, indicando che prevedendo sempre quella classe, si ottiene circa il 53% delle previsioni corrette.

- **F1 Score:** L'F1 Score per la baseline è di **0.00**. Poiché la baseline predice sempre la classe maggioritaria (assenza di malattia), non ci saranno mai veri positivi per la classe "malattia", portando la recall e di conseguenza l'F1 Score a zero.

In sintesi, il confronto mostra una **netta superiorità** del modello di Regressione Logistica ottimizzato rispetto alla baseline. Le sue prestazioni significativamente più elevate in entrambe le metriche confermano che il modello ha imparato pattern rilevanti dai dati e può fare predizioni affidabili per la presenza di malattie cardiache.

Sviluppi futuri

Il progetto "HeartRisk AI" ha dimostrato la capacità di predire la presenza di malattie cardiache con un'elevata accuratezza utilizzando un modello di Regressione Logistica. Tuttavia, esistono diverse direzioni per futuri sviluppi che potrebbero ulteriormente migliorare l'efficacia e l'applicabilità del sistema.

- **Raccolta Dati Storici Clinici e Comportamentali:** Integrare il sistema con un database che memorizzi lo storico delle interazioni e dei dati clinici del paziente nel tempo, come:
 - **Dati inseriti:** età, parametri vitali (pressione sanguigna, colesterolo), abitudini di vita (fumo, attività fisica, dieta), familiarità con malattie cardiache, ecc.
 - **Esiti clinici:** evoluzione della malattia, eventi cardiaci (infarti, ictus), risposte a trattamenti o modifiche dello stile di vita.
 - **Feedback dell'utente:** se il sistema fosse interattivo, feedback sull'efficacia delle raccomandazioni o sulla percezione del proprio stato di salute.
- **Arricchimento del dataset:** originale integrando informazioni e dati esterni. L'obiettivo è ottenere predizioni più accurate e una comprensione più approfondita delle relazioni complesse tra le feature e il rischio di malattie cardiache.

Un approccio possibile potrebbe essere effettuato tramite:

- **Dati Socioeconomici:** Livello di istruzione, reddito, occupazione, accesso ai servizi sanitari e alimentari, che possono influenzare stili di vita e accesso a cure.
- **Dati Ambientali:** Inquinamento atmosferico (particolato, ozono), esposizione a rumore, temperatura ambiente, qualità dell'acqua, che sono noti fattori di rischio cardiovascolare.
- **Dati Genetici:** Predisposizione genetica a condizioni come ipertensione, ipercolesterolemia, o diabete, varianti genetiche associate al metabolismo dei lipidi o alla risposta infiammatoria.
- **Dati Comportamentali Estesi:** Livelli di stress cronico, qualità e durata del sonno, utilizzo di sostanze (es. consumo di tabacco e alcool dettagliato, non solo binario).
- **Dati Omics:** Integrazione di dati proteomici, metabolomici o microbiomici per una comprensione più dettagliata dei meccanismi biologici sottostanti.