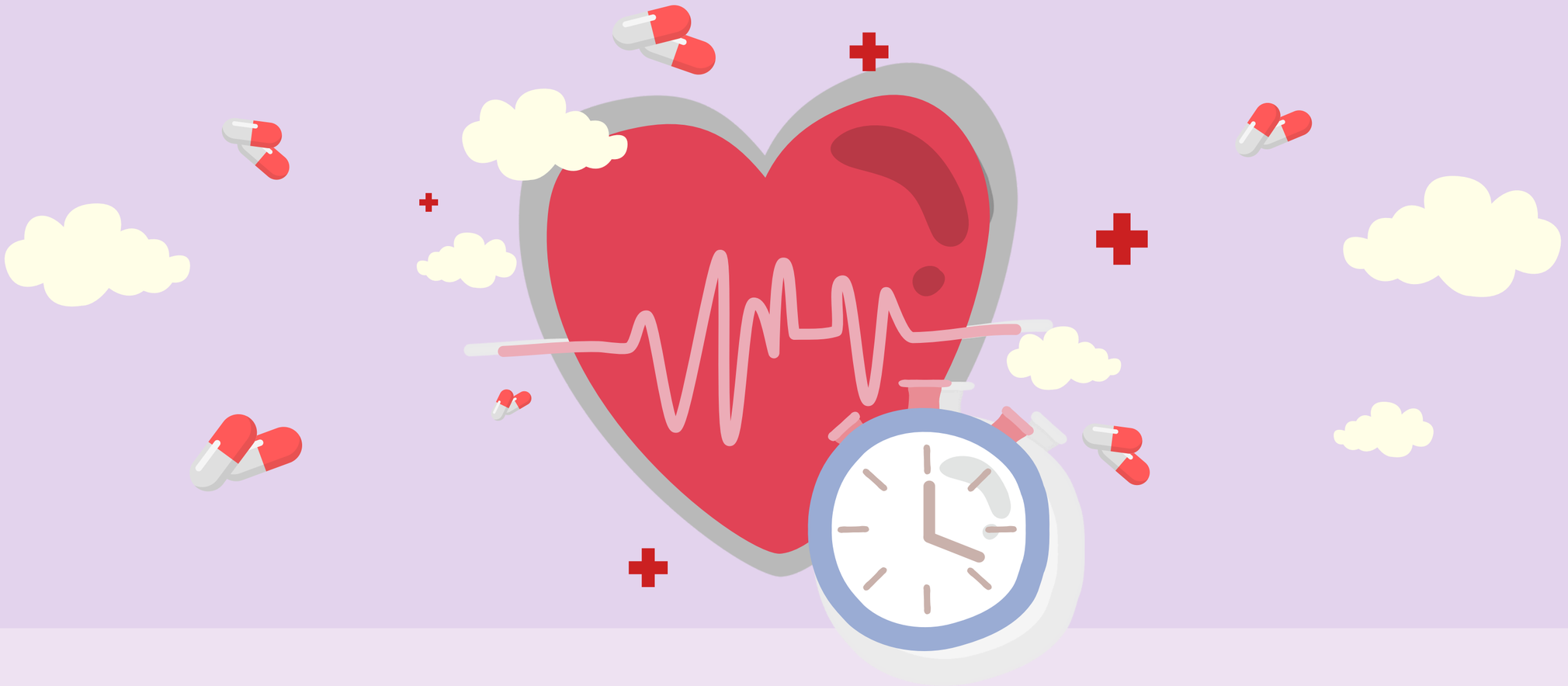


PREDIZIONI DELLE MALATTIE CARDIOVASCOLARI

con scikit-learn

GIULIA GOVERNATORI

IL PUNTO DI PARTENZA

Perché questo progetto?

Le malattie cardiovascolari sono la prima causa di morte nel mondo:

17,9 milioni *di decessi all'anno (OMS)*

Spesso evitabili con una diagnosi precoce...

La domanda:

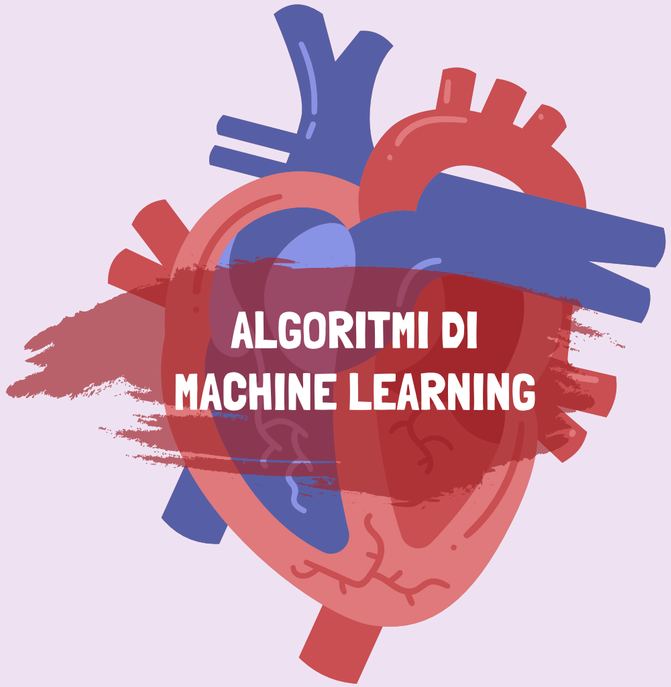
È possibile prevedere il rischio utilizzando i dati clinici?

LA NOSTRA MISSIONE

Obiettivo del progetto

Creare un sistema intelligente che aiuti a identificare i pazienti a rischio

DATI CLINICI DEL PAZIENTE



PROBABILITÀ DI MALATTIA
+ LIVELLO DI RISCHIO

I DATI

Il Dataset UCI Heart Disease

Qualità eccellente: nessun valore mancante!

303 pazienti di 4 ospedali internazionali : → 📋 13 variabili cliniche per paziente

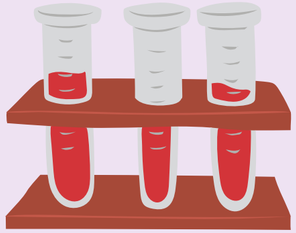
160 persone sane | 137 malate

COSA MISURIAMO?

Le 13 Variabili Cliniche

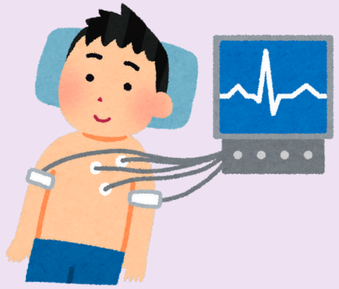
Dati demografici

- età
- sesso



Esami di base

- pressione sanguigna
- colesterolo
- glicemia a digiuno



Tests avanzati

- ECG a riposo
- Frequenza cardiaca massima sotto sforzo
- Tipo di dolore toracico
- Angina da sforzo
- Depressione del segmento ST sotto sforzo (oldpeak)
- Pendenza del segmento ST
- Numero di vasi colorati (angiografia)
- Talassemia (anomalia del sangue)

LA SFIDA INIZIALE

Un problema da risolvere

Il dataset aveva 5 classi di malattia →

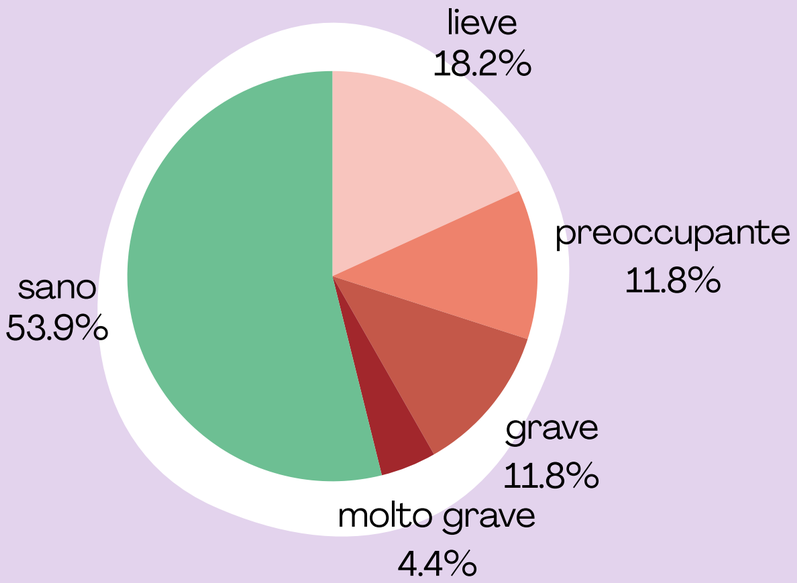
- Classe 0 : Sano (160 pazienti)
- Classi 1-4 : Gravità crescente

❌ **PROBLEMA:**

Classi troppo sbilanciate e solo 303 osservazioni

↓

Modello che non apprende dalle classi inferiori



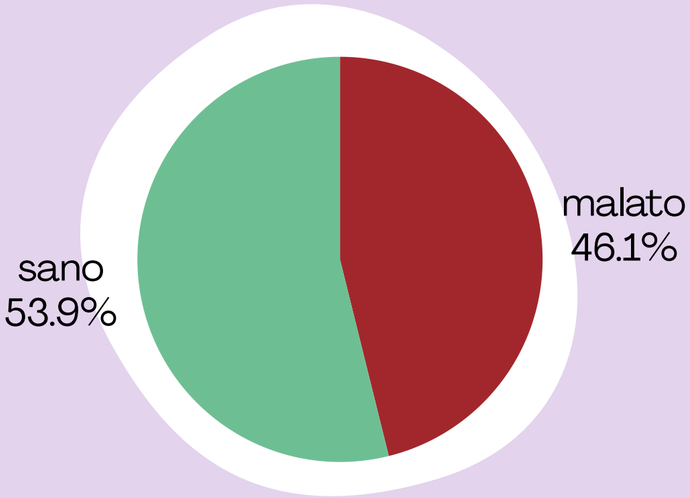
→

✅ **SOLUZIONE:**

Trasformazione in problema binario

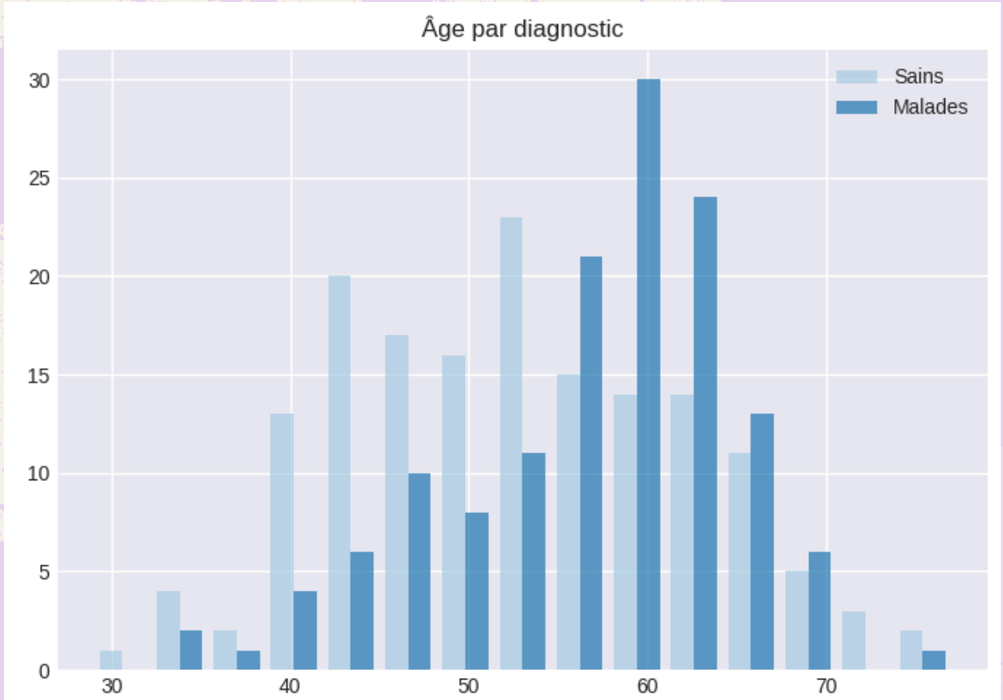
↓

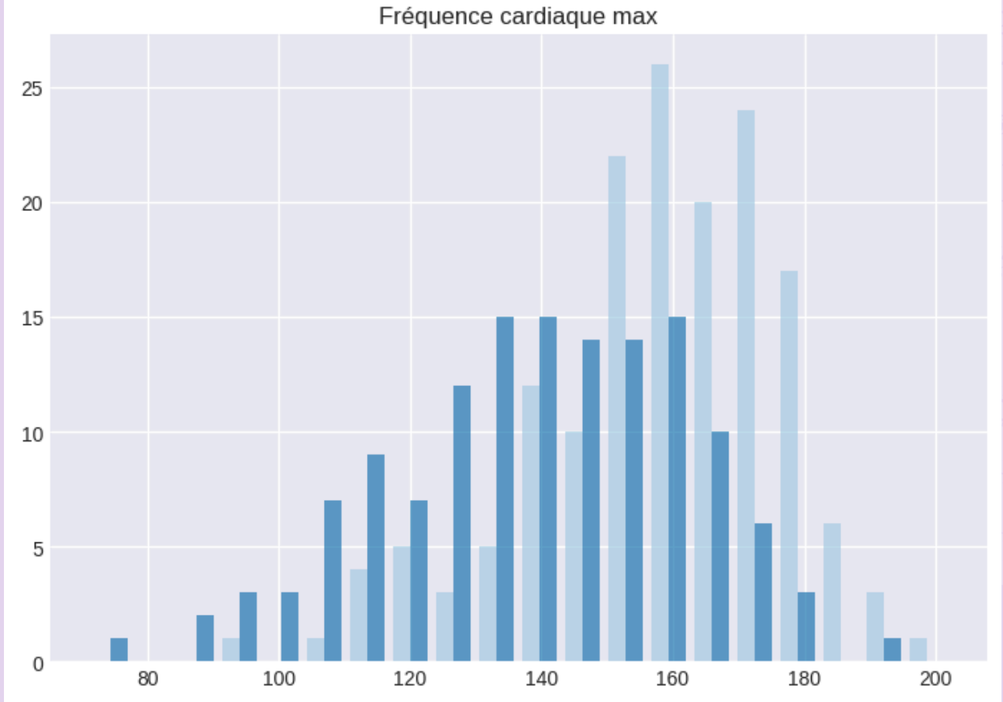
Sano vs Malato → classi equilibrate

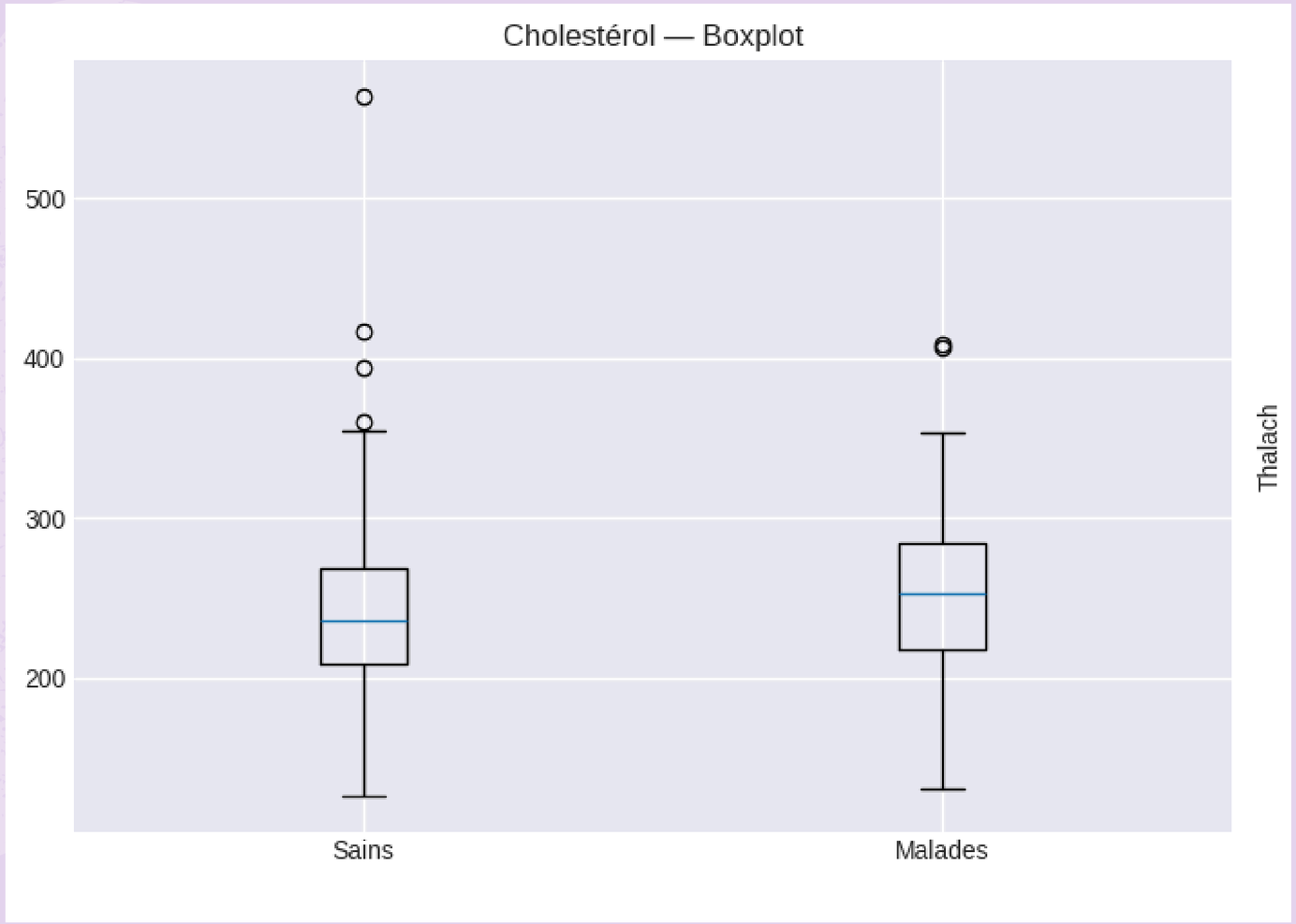


EXPLORARE I DATI

Risultati chiave dell'analisi esplorativa







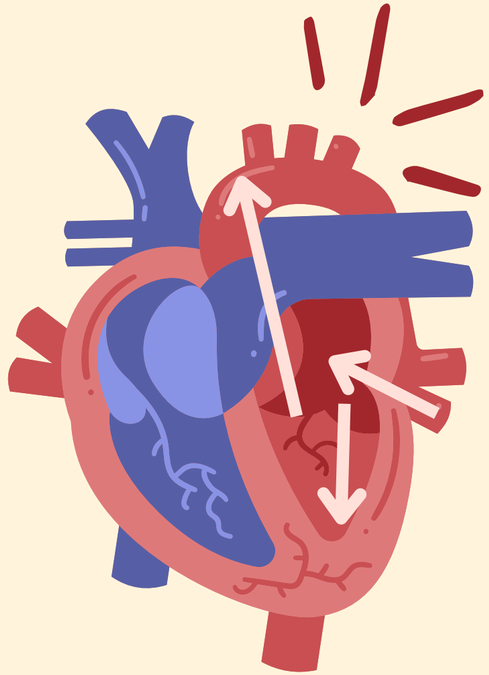
PREPARARE I DATI
Il Preprocessing

STEP 1:
Train-Test Split + Cross-Validation

La struttura a 3 livelli:

📦 Set di dati completo (303 pazienti)
- 🎓 **80% Training Set** (237 pazienti)
  - Diviso in **5** "**fold**" per la **Cross-Validation**
  - → Fold 1, 2, 3, 4, 5 si alternano come validazione
  - → Utilizzato per scegliere i migliori iperparametri
- 🧪 **20% Test Set** (60 pazienti)

→ MAI toccato fino alla valutazione finale!



PREPARARE I DATI
Il Preprocessing

STEP 1:
Train-Test Split + Cross-Validation

Come funziona la validazione incrociata a 5 pieghe (5-Fold Cross-Validation):

Training Set → diviso in 5 parti uguali
Val → Training set usato come Validation

Round 1: [**Val**] [**Train**] [**Train**] [**Train**] [**Train**]
Round 2: [**Train**] [**Val**] [**Train**] [**Train**] [**Train**]
Round 3: [**Train**] [**Train**] [**Val**] [**Train**] [**Train**]
Round 4: [**Train**] [**Train**] [**Train**] [**Val**] [**Train**]
Round 5: [**Train**] [**Train**] [**Train**] [**Train**] [**Val**]

Media dei 5 risultati = Prestazione stimata

Perché questo evita il **DATA SNOOPING**:

✅ Test set = visto una sola volta alla fine
✅ Cross-Validation = “test” sempre sul “training set”
✅ Decisioni basate unicamente sul Training Set
✅ Stima onesta sulla capacità di generalizzazione

🔒 Stratificazione: le proporzioni sani/malati (54 %/46 %) sono mantenute in ogni Fold

PREPARARE I DATI
Il Preprocessing

STEP 2:

Standardizzazione dei numeri

→ Tutte le variabili sulla stessa scala

STEP 3:

Codifica delle categorie

→ Trasformare il testo (delle classi) in numeri comprensibili dalla macchina (0/1)

PRIMO MODELLO – RANDOM FOREST

La foresta che impara

Che cos'è un Random Forest ?

Imagina 100 medici che danno la loro opinione:

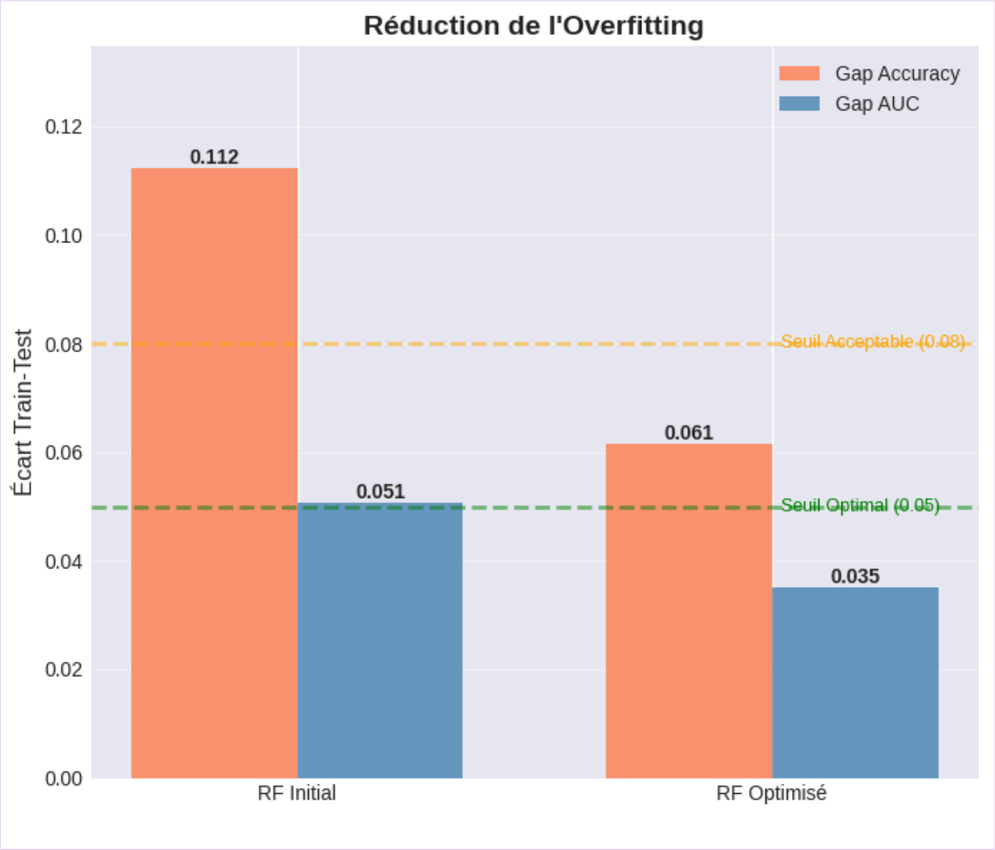
- Ogni medico esamina il paziente da un punto di vista diverso.
- Poi votano insieme per la diagnosi finale.
- Vince la maggioranza!

Risultato iniziale:

- ✅ Precision sul test : 86,7%
- ❌ Ma... se lo ricorda troppo bene... (***overfitting***)

OTTIMIZZAZIONE

Overfitting della Random Forest: un problema da risolvere



Il problema: il modello Random Forest impara "a memoria" invece di capire.

Le soluzioni applicate:

1. 🌳Limitare la profondità degli alberi per evitare strutture troppo complesse.
2. 🌿Richiedere più osservazioni in ogni foglia per impedire l'apprendimento di casi isolati.
3. 🎲Introdurre una maggiore casualità nella scelta delle variabili per diversificare gli alberi.
4. 🌲Aumentare il numero totale di alberi per stabilizzare le previsioni mediante media.
5. 🧪Richiedere un numero maggiore di esempi prima di dividere un nodo per ridurre separazioni inutili.
6. ⚓Attivare il ribilanciamento automatico delle classi per limitare l'influenza di una classe dominante.

Risultato dopo ottimizzazione:

- **Precision :** 85% (simile)
- **Overfitting :** ridotto piu del 30% ! ✨

SECONDO MODELLO – LA REGRESSIONE LOGISTICA

L'approccio lineare

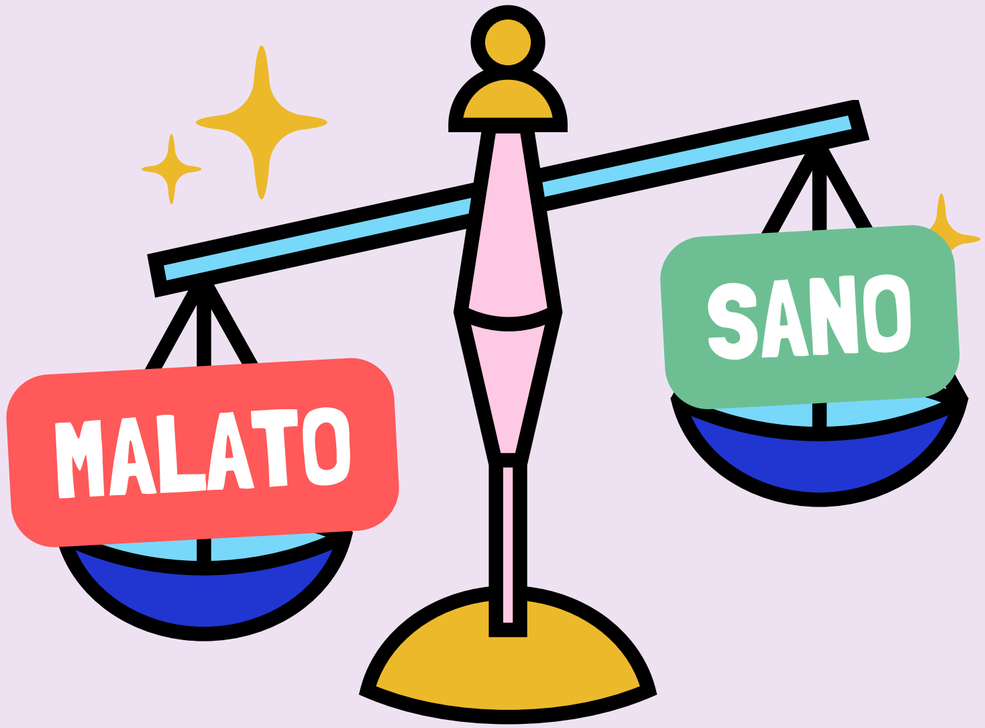
Un modello più semplice ma potente.

Immaginate una bilancia che pesa l'importanza di ciascun sintomo:

- Dolore toracico → +3 punti di rischio
- Età avanzata → +2 punti
- Colesterolo basso → -1 punto
- Totale > soglia = malattia

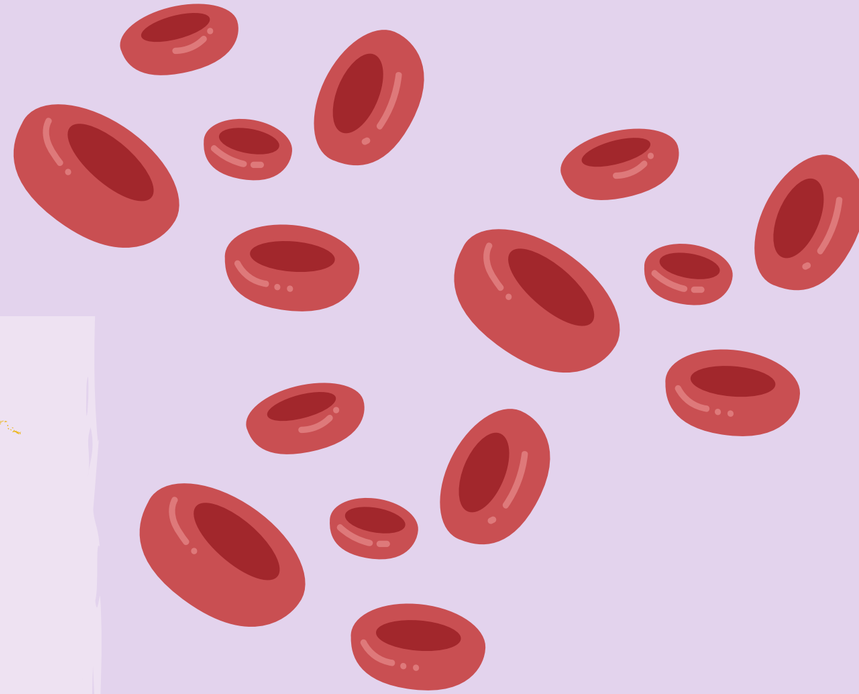
Risultato:

✅ Precisione del test: 86,7%
✅ Zero overfitting!
✅ Facilmente interpretabile
•



IL MEGLIO DEI DUE MONDI

Stacking Ensemble : combinare i due modelli!



Livello 1 : Random Forest fa previsioni + Regressione Logistica fa previsioni

Livello 2 : Un terzo modello (Stacking) impara dai 2 e decide

È come chiedere consiglio a:

- Un esperto intuitivo (Random Forest)
- Un analista metodico (Regressione logistica)
- Un direttore che sintetizza entrambi (Meta-modello)

IL CONFRONTO FINALE

Chi vince?

Vincitore per questo projetto: **Stacking**
→ Migliore discriminazione complessiva (**AUC = 0,953**)

Critère	Random Forest	Logistic Regression	Stacking	Gagnant
AUC Test	0.946	0.951	**0.953**	🏆 Stacking
Accuracy	0.867	**0.867**	0.833	🏆 RF/LR
Precision Malade	0.88	0.92	0.88	🏆 LR
Recall Malade	0.82	0.79	0.75	🏆 RF
Overfitting	Modéré	✅ Aucun	Aucun	🏆 LR/Stack
Interprétabilité	Faible	✅ Élevée	Moyenne	🏆 LR

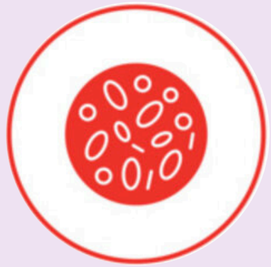
COSA CONTA DAVVERO?

L'importanza delle variabili

Top 5 fattori predittivi (62% del potere predittivo) :

🫀 **Tipo di dolore toracico**

15,8%



🩸 **Talassemia**

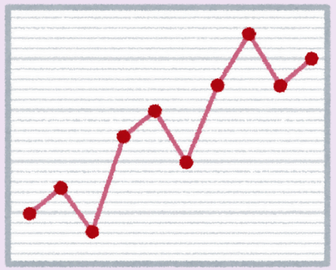
15,0%

💓 **Frequenza cardiaca massima sotto sforzo**

11,1%

🔬 **Vasi sanguigni colorati**

10,4%



📉 **Depressione ST**

9,8%

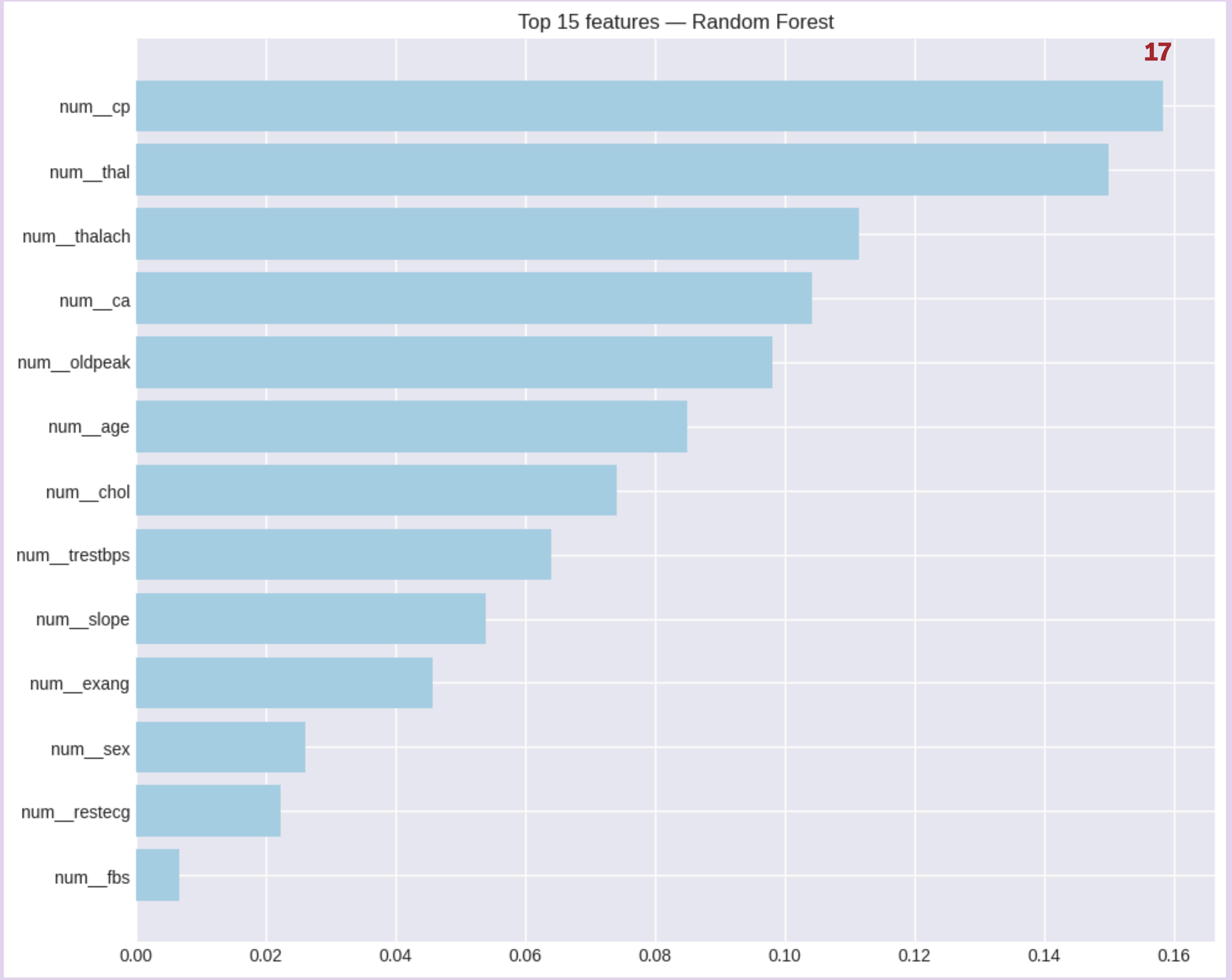
L'analisi dell'importanza delle variabili nel modello Random Forest rivela che cinque caratteristiche dominano la previsione del rischio cardiovascolare.

COSA CONTA DAVVERO?

L'importanza delle variabili

Variables les Plus Prédictives

Variable	Importance (%)
`cp` (type de douleur thoracique)	15,8%
`thal` (thalassémie)	15,0%
`thalach` (fréquence cardiaque maximale)	11,1%
`ca` (nombre de vaisseaux colorés)	10,4%
`oldpeak` (dépression ST)	9,8%



DAL MODELLO ALL'APPLICAZIONE

Il Dashboard What-If

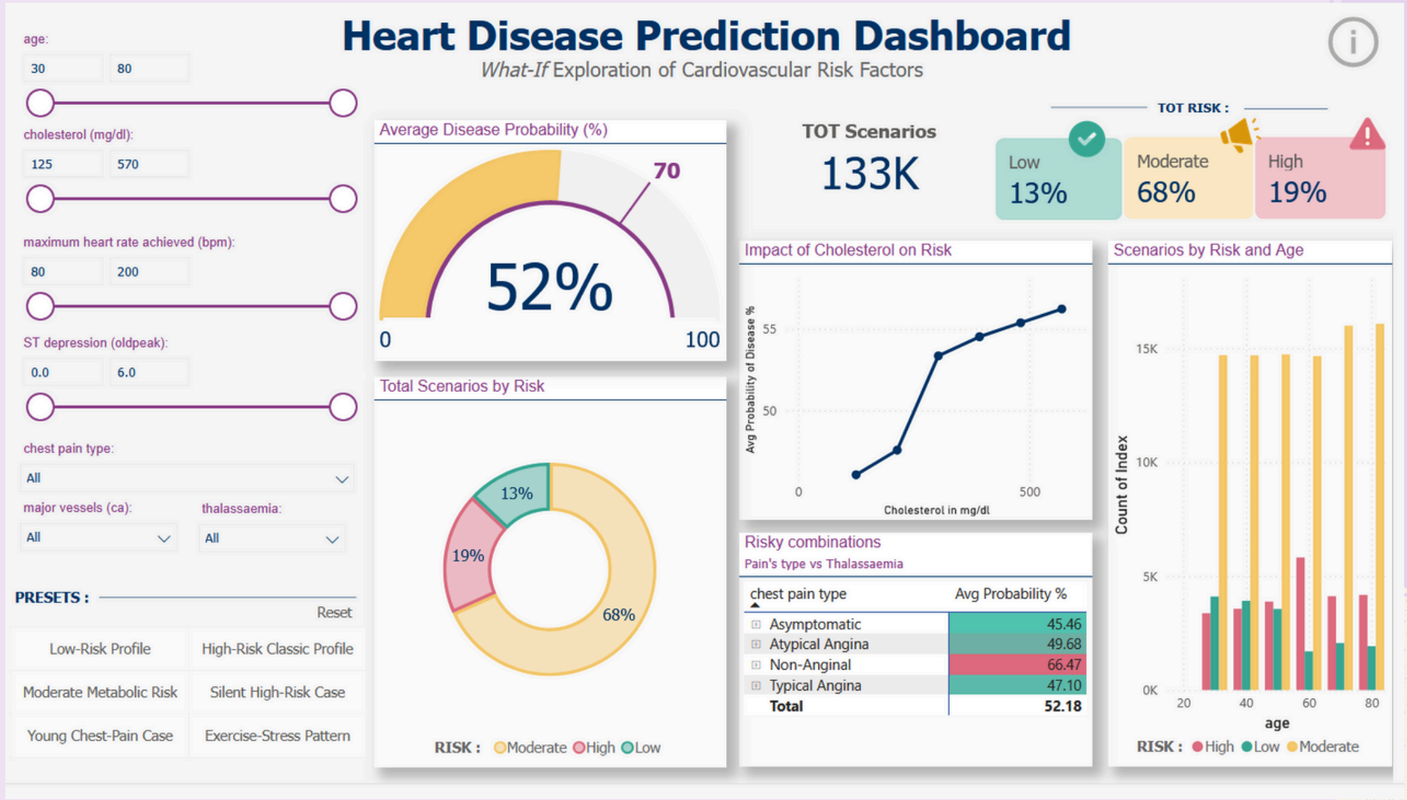
Abbiamo creato 133.056 scenari simulati!

Come?
Combinando tutte le possibilità:

- *Età: 30-80 anni*
- *Colesterolo: 125-570 mg/dl*
- *Frequenza cardiaca: 80-200 bpm*
- *Altre 4 variabili cliniche*

Risultato:

*Un simulatore interattivo in **Power BI** in cui gli utenti possono esplorare scenari "what if..."*



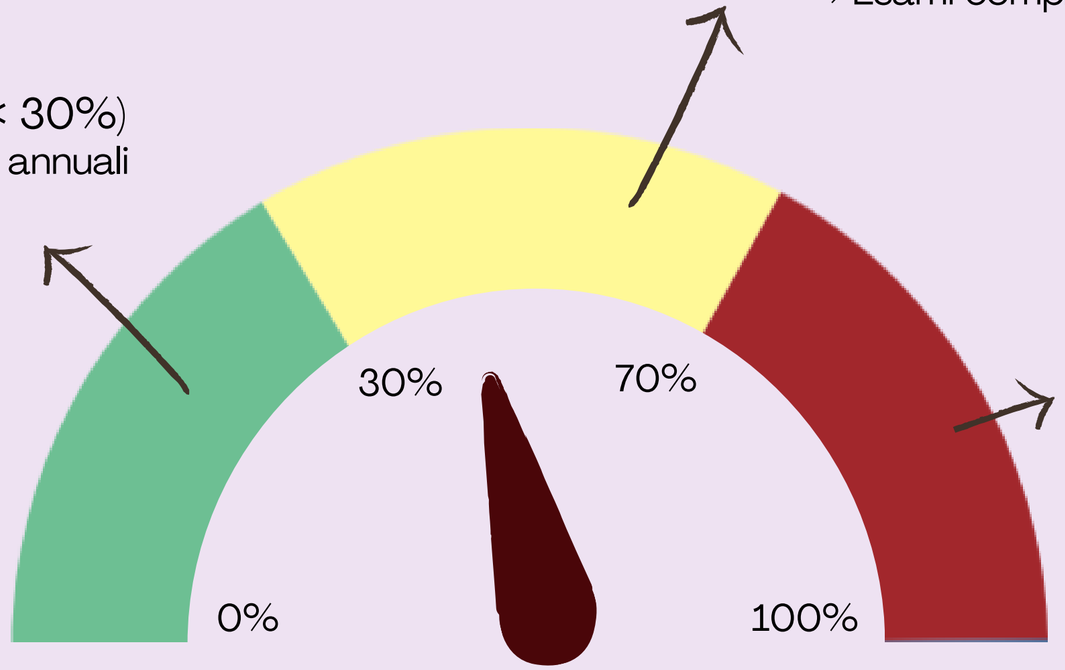
CATEGORIE DI RISCHIO

Comunicare in modo chiaro

Abbiamo tradotto le probabilità in azioni concrete:

Rischio MODERATO (30-70%)
→ Esami complementari consigliati

Rischio BASSO (< 30%)
→ Controlli di routine annuali



Rischio ELEVATO (≥ 70%)
→ Intervento cardiologico immediato

CASO PRATICO

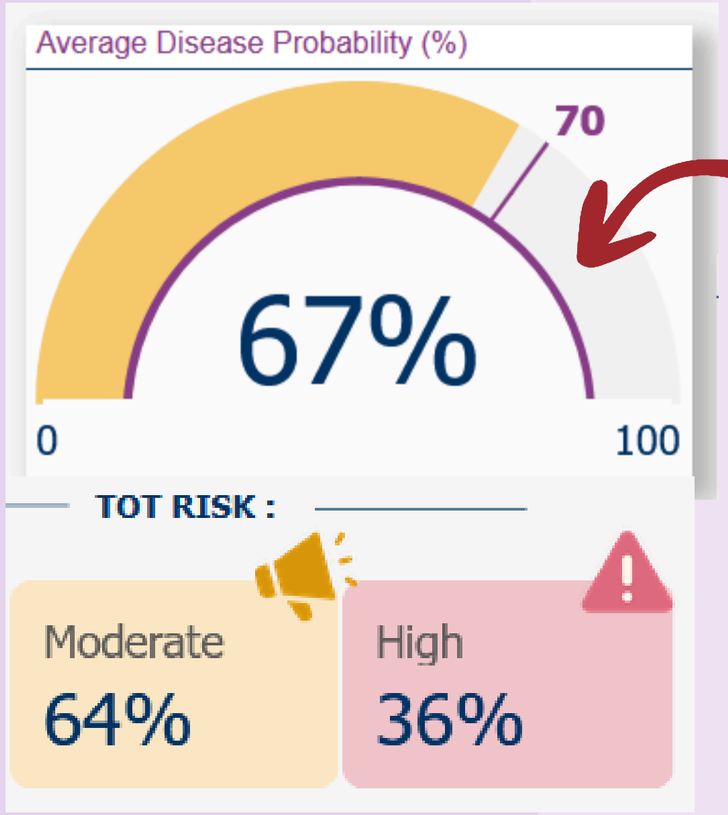
Esempio di utilizzo

Paziente tipo:

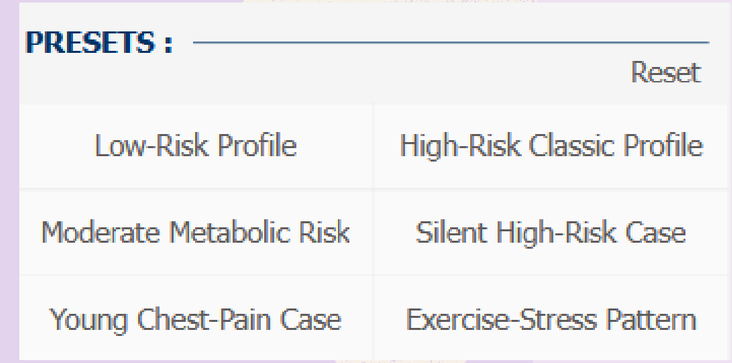
👤 Uomo, **55-65** anni
🩺 Colesterolo **240-450** mg/dl
💓 Frequenza massima **80-120** bpm
😫 Dolore toracico: **tipico anginoso**
🔬 Talassemia **fissa**

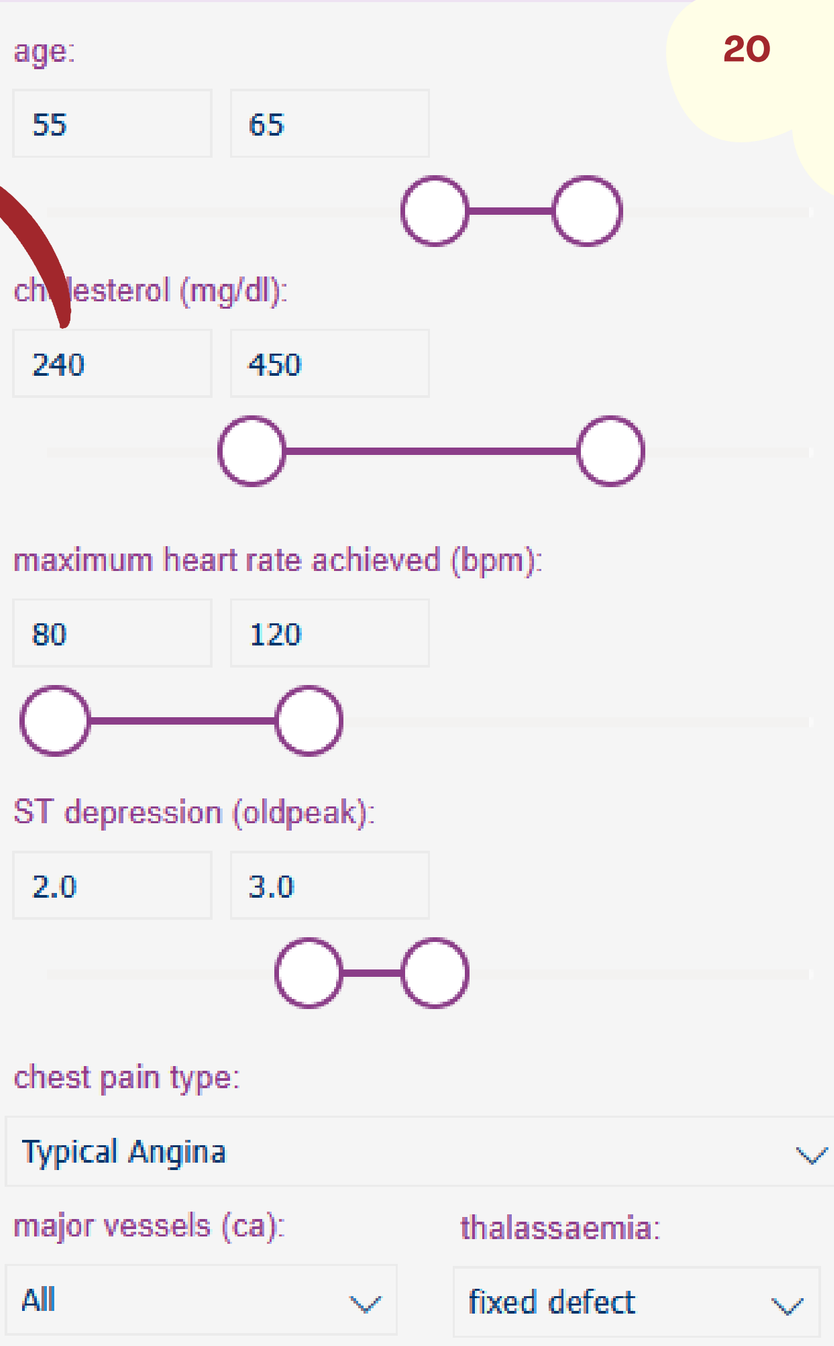
Previsione del modello:

→ **Probabilità di malattia: 67%**
→ Categoria: **RISCHIO MODERATO** 🟡 per il 64%
→ Azione: **Ecocardiogramma + test da sforzo**



Ho anche previsto dei ***presets*** per facilitare l'esplorazione dei dati.





COSA ABBIAMO IMPARATO

Il successo del progetto

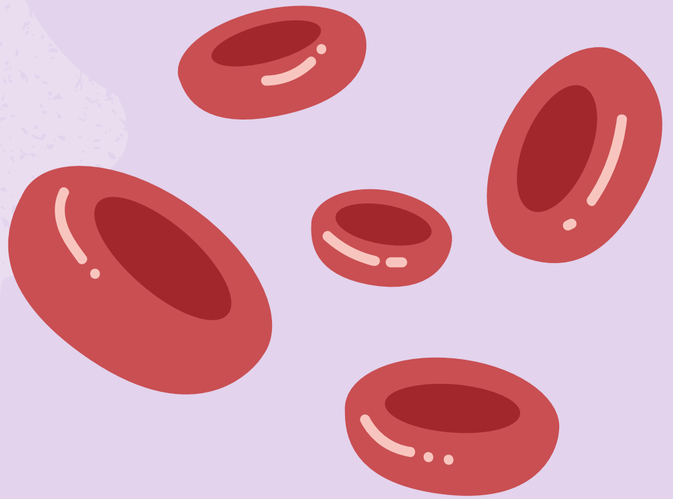
✅ **Prestazioni eccellenti**: AUC 0,953 (quasi perfetto)

✅ **Modello robusto:** nessun overfitting nel modello finale

✅ **Interpretabilità**: sappiamo quali variabili contano

✅ **Applicabilità**: dashboard pronto per l'uso clinico

✅ **Metodologia**: approccio rigoroso e replicabile



LE SFIDE E I LIMITI

Restare onesti

⚠️ *LIMITI DA CONSIDERARE:*

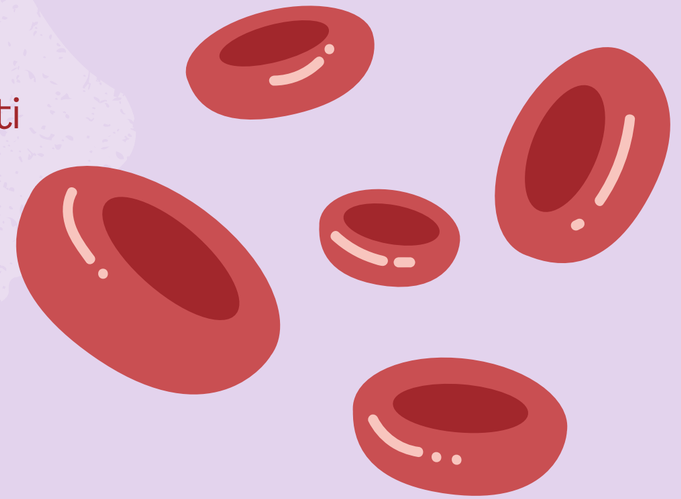
📊 **Dataset piccolo**: 303 pazienti → servono più dati

🌍 **Contesto geografico** : solo 4 ospedali

⚡ **Recall classe Malati** : 75-82% → alcuni malati non diagnosticati

🔄 **Validazione** : Necessità di test su nuovi pazienti reali

Questi limiti non invalidano il progetto, ma indicano i possibili miglioramenti!



IL FUTURO DEL PROGETTO

Prossimi passi

Il modello attuale non può ancora essere utilizzato su pazienti reali.

Questo progetto richiederebbe alcuni miglioramenti.

Cosa possiamo migliorare?

🔬 Più dati: obiettivo 1000+ pazienti

🌍 Diversità geografica: includere più popolazioni

🧬 Nuove variabili: storia familiare, stile di vita, genetica

🤖 Deep Learning: se disponiamo di dati sufficienti (oltre 300)

🔄 Studio prospettico: Convalidare su nuovi pazienti nel tempo

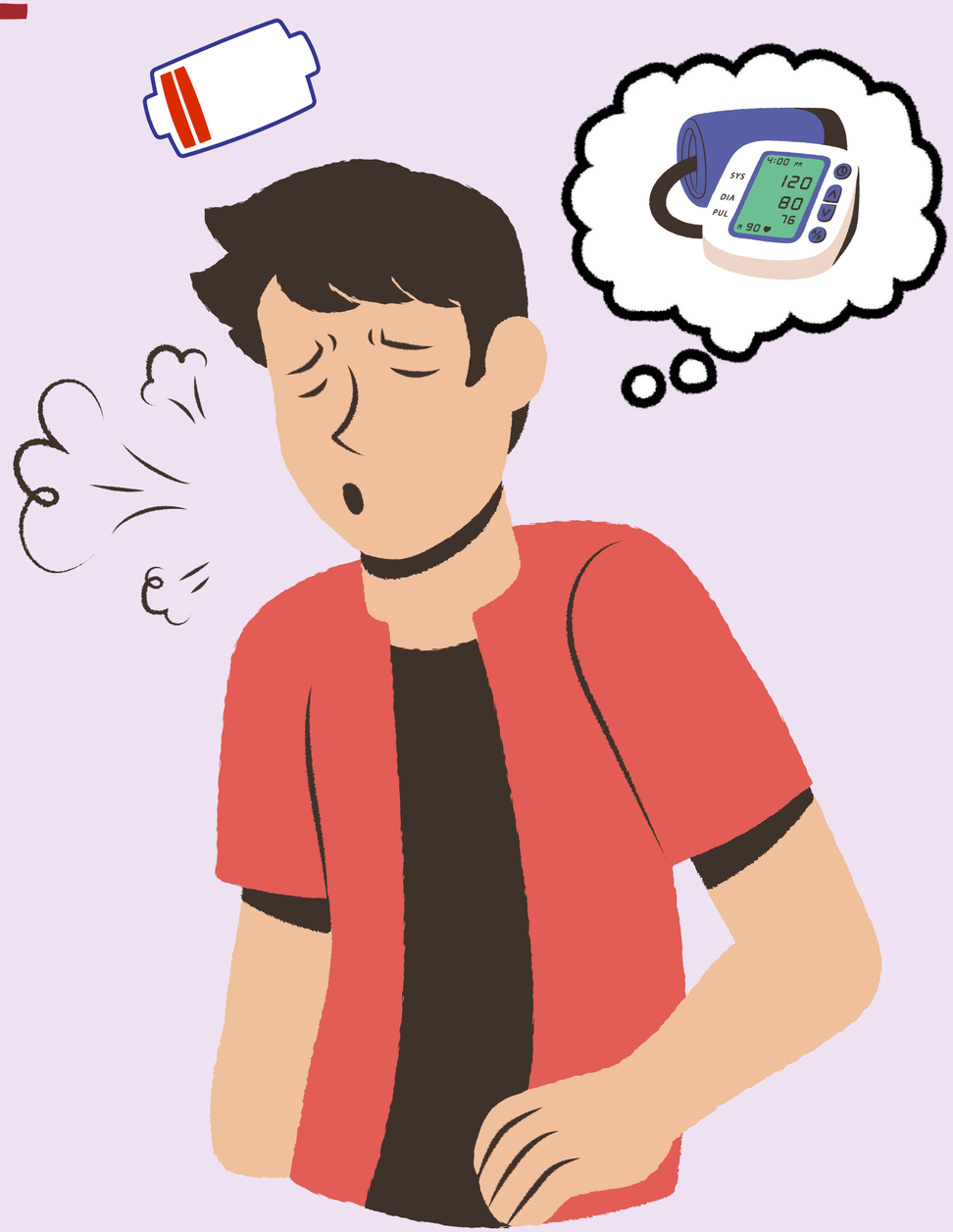
📱 App mobile: Dashboard accessibile ovunque

IMPATTO DEL PROGETTO ATTUALE

A cosa serve questo progetto?

Con il modello attuale, gli utenti possono utilizzare la mia dashboard per:

1. 📱 **Prendere coscienza dei propri rischi**
2. 💪 **Motivarsi a cambiare stile di vita**
3. 🚴 **Vedere l'impatto del colesterolo, dell'esercizio fisico, ecc.**



TECNOLOGIE UTILIZZATE

Gli strumenti del mestiere

Linguaggio : Python

Librerie principali:

- scikit-learn → Modelli ML e metriche
- pandas → Gestione dei dati
- matplotlib/seaborn → Visualizzazioni
- joblib → Salvataggio dei modelli

Tecniche applicate:

- GridSearchCV (ottimizzazione)
- Cross-validation (convalida)
- Ensemble methods (stacking)

Output: Dashboard Power BI

LEZIONI APPRESE

Oltre i numeri

Lezioni tecniche:

- L'AUC è più informativa della precisione per i problemi medici
- La regolarizzazione è fondamentale contro l'overfitting
- I metodi ensemble migliorano la robustezza

Lezioni pratiche:

- Bilanciare prestazioni e interpretabilità
- Pensare all'utente finale (medici/pazienti)
- Essere trasparenti sui limiti

Lezione principale:

L'IA non sostituisce il medico, ma lo supporta nelle sue decisioni

CONCLUSIONI

Il messaggio finale

Siamo partiti da una domanda:

*"**È possibile prevedere le malattie cardiache con i dati?**"*

Siamo giunti a una risposta:

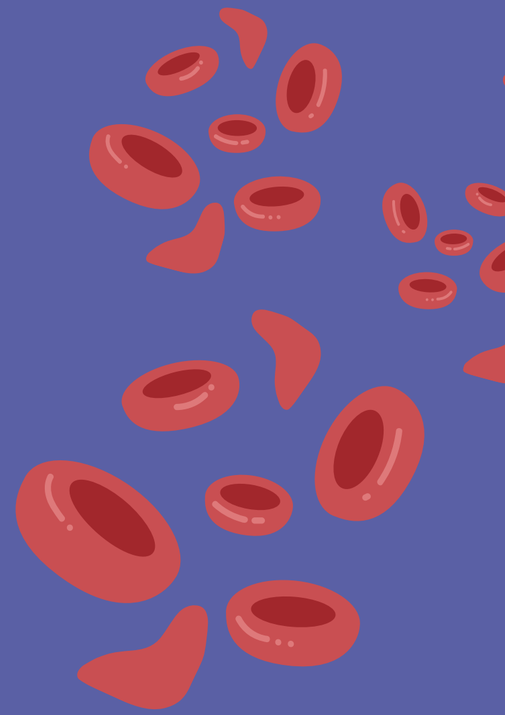
*"**Sì, con una precisione del 95% e uno strumento utilizzabile**"*

Il futuro della medicina è la collaborazione uomo-macchina

GLOSSARIO

Termini chiave spiegati

- **AUC** (**Area Under Curve**)**:** da 0 a 1, misura quanto il modello sia in grado di distinguere tra soggetti sani e malati. 0,95+ = eccellente
- **Overfitting:** quando il modello apprende "a memoria" i dati invece di comprenderne i modelli generali e non è più in grado di generalizzare su nuovi dati.
- **Cross-validation:** testare il modello su diverse parti dei dati per verificarne la robustezza
- **Stacking:** combinare le previsioni di più modelli per ottenerne uno migliore
- **Recall:** % di malati effettivamente identificati dal modello



RINGRAZIAMENTI & CONTATTI

Grazie per l'attenzione!

📊 Progetto: **Previsione delle malattie cardiache**
📅 Data: **9 novembre 2025**

👤 Autrice: **Giulia Governatori**
📧 Email : **giuliagovernatori@hotmail.com**
💥 portfolio : **Giulia-Governatori.alwaysdata.net**
🌐 Linkedin : **@giuliagovernatori-bi-analyst/**

🔗 Dataset : **UCI Machine Learning Repository**
🙏 Credits : Hungarian Institute of Cardiology, University Hospital Zurich, University Hospital Basel, V.A. Medical Center

GRAZIE!