Medical Diagnosis Aid with Data Science and Machine Learning

Luca Granucci, Paolo Walsh

Università di Pisa – A.A. 2024/2025

Introduzione

- Interesse per l'applicazione della **data science** e del **machine learning** in un ambito concreto e ad alto impatto, la **medicina**
- Utilizzo del machine learning per supportare la **diagnosi** precoce di **malattie cardiache**
- Ridurre errori soggettivi e ottenere una prima valutazione **rapida** ed **efficace**

Obiettivi del progetto

- Sviluppare **modelli predittivi** a partire da dati clinici
- Confrontare prestazioni dei modelli scelti con **GPT-4** e un medico in formazione
- Analizzare **l'interpretabilità** dei modelli
- Integrare i modelli in un'**applicazione web**

Dataset UCI Heart Disease

- **910** pazienti da 4 centri clinici (Cleveland, Ungheria, Svizzera, Long Beach)
- Feature cliniche: età, sesso, dolore toracico, colesterolo, ECG, ecc.
- Obiettivo: predizione binaria (malato / non malato)

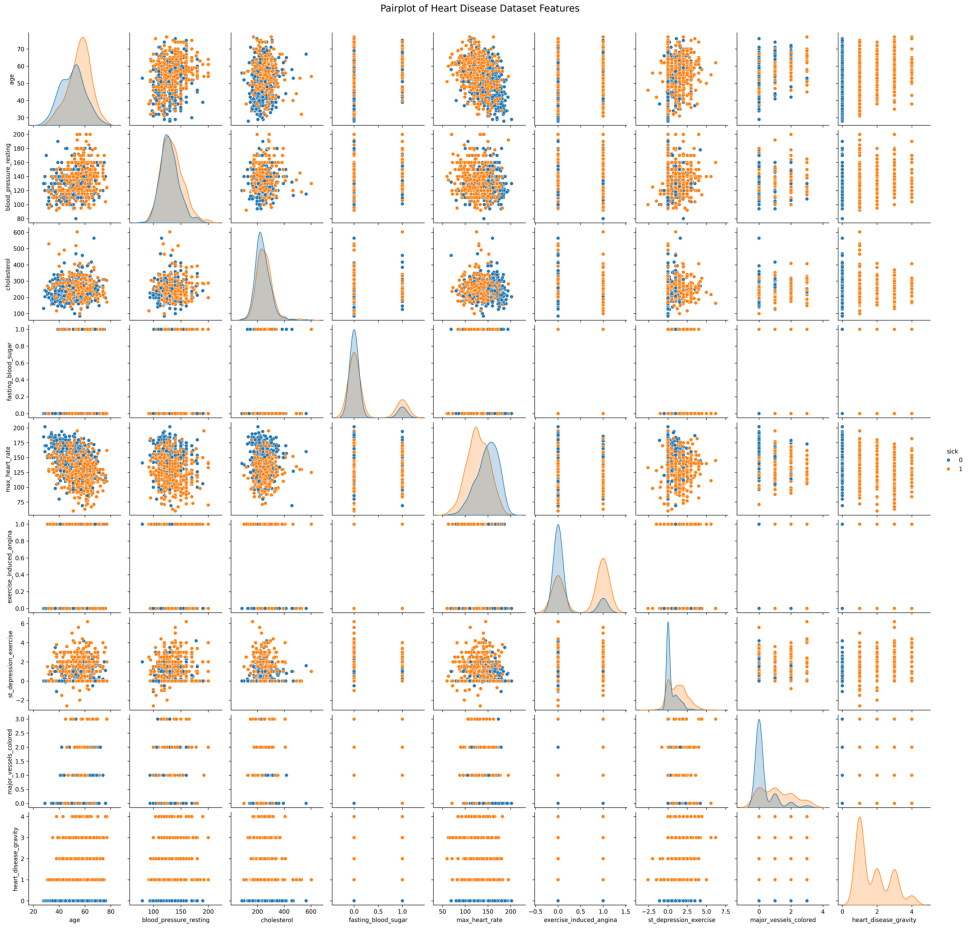
Feature del dataset

Feature	Descrizione
`id`	Identificatore univoco del paziente nel dataset.
`age`	Età del paziente (in anni).
`sex`	Sesso del paziente (0 = femmina, 1 = maschio).
`dataset`	Origine del campione all'interno dei diversi sottodataset UCI.
`chest_pain_type`	Tipo di dolore toracico (es. tipico anginoso, atipico, non anginoso, asintomatico).
`blood_pressure_resting`	Pressione arteriosa a riposo (mm Hg).
`cholesterol`	Colesterolo sierico (mg/dl).
`fasting_blood_sugar`	Glicemia a digiuno > 120 mg/dl (1 = vero, 0 = falso).
`ecg_resting`	Risultati dell'elettrocardiogramma a riposo (0, 1 o 2, che indicano diverse anomalie).
`max_heart_rate`	Frequenza cardiaca massima raggiunta durante il test da sforzo.
`exercise_induced_angina`	Angina indotta da esercizio.
`st_depression_exercise`	Depressione del tratto ST causata dall'esercizio rispetto al riposo.
`st_slope_type`	Inclinazione del segmento ST durante l'esercizio (es. ascendente, piatto, discendente).
`major_vessels_colored`	Numero di vasi principali visualizzati con fluoroscopia colorata (da 0 a 3).
`thal_defect_type`	Tipo di difetto nel test del talio (normale, fisso, reversibile).
`heart_disease_gravity`	Grado di gravità della malattia cardiaca (0 = nessuna, fino a 4 = massima).
`sick`	Etichetta binaria: 1 = presenza di malattia cardiaca, 0 = assenza.

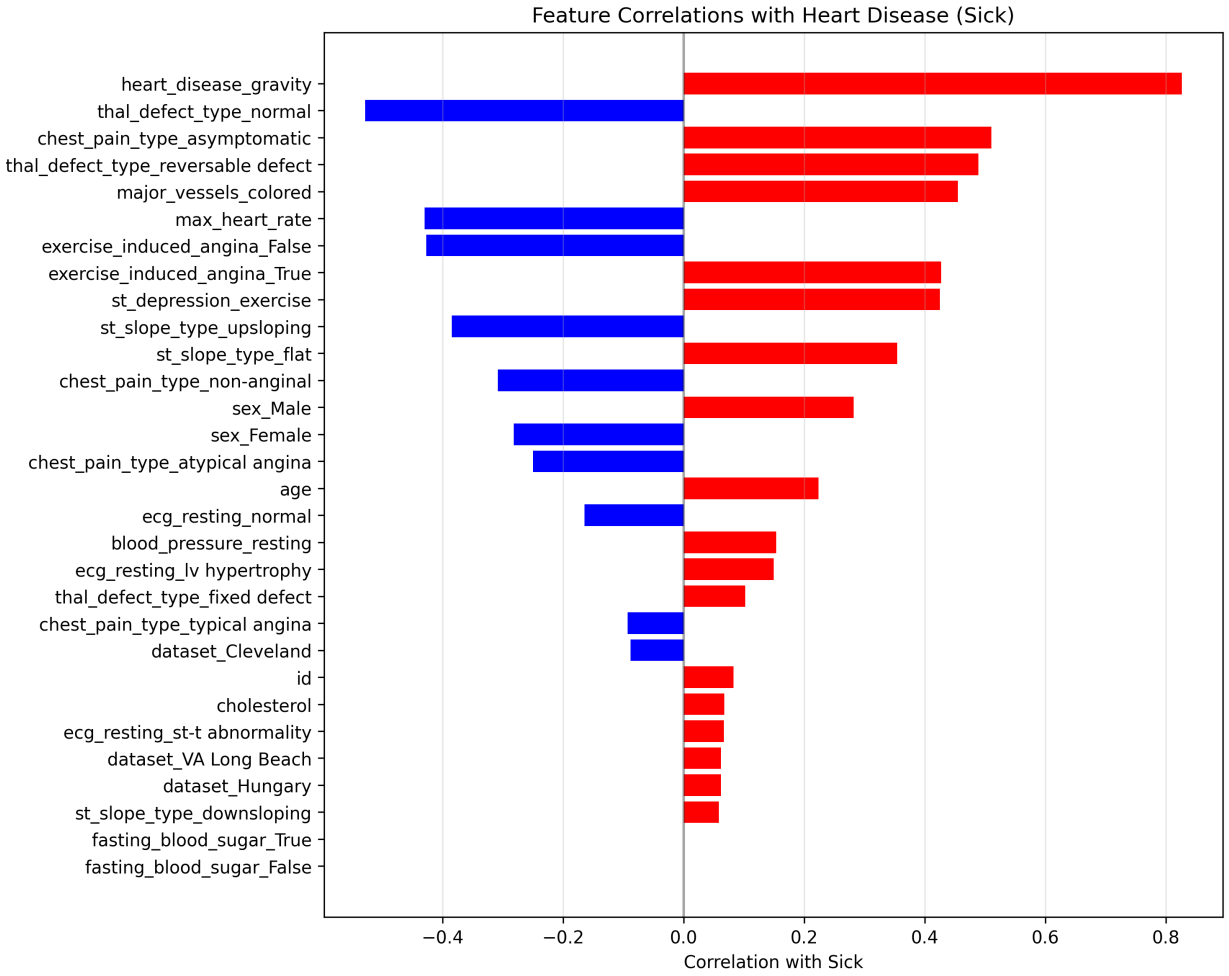
Analisi Esplorativa dei Dati

- Pulizia: rimossi record con valori **anomali/nulli**
- Riduzione a 299 campioni validi
- Individuazione di **bias** all'interno del dataset
- **Correlazioni Pearson** e visualizzazioni esplorative

Piarplot delle Feature



Correlazione con la malattia



Correlazione con la malattia

- La tabella evidenzia le variabili con **correlazione assoluta ≥ 0.3** con la variabile ***sick***
- Alcune feature comunemente ritenute rilevanti come il **colesterolo** risultano **meno correlate** rispetto ad altre come il **sesso**
- Ciò è probabilmente dovuto alla **ridotta dimensione del dataset** e alla presenza di **bias strutturali**

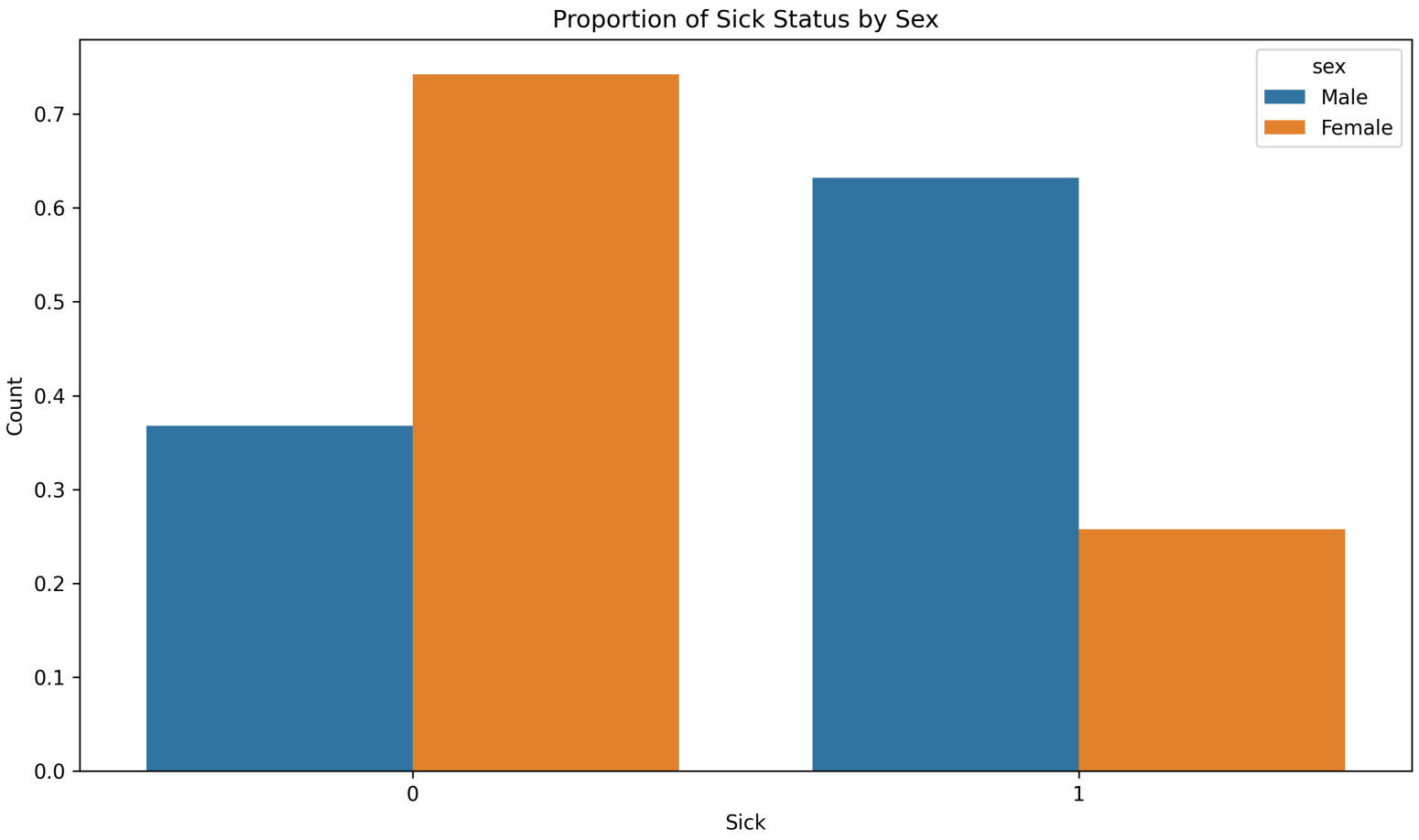
Feature	Correlazione con `sick`
heart_disease_gravity[a]	0.78
chest_pain_type	-0.46
thal_defect_type	0.46
exercise_induced_angina	0.46
major_vessels_colored	0.46
st_slope_type	0.44
max_heart_rate	-0.39

[a] Questa colonna rappresenta direttamente il grado di malattia diagnosticato, e quindi è naturalmente molto correlata con la variabile `sick`. Per questo motivo, non viene utilizzata come feature nei modelli predittivi.

Bias presenti nel dataset

- **80% dei campioni sono maschi** → forte squilibrio di genere
- Il modello rischia di **sovra-adattarsi alla classe dominante**, penalizzando la diagnosi femminile
- Anche normalizzando per sesso, persiste uno **sbilanciamento nella distribuzione della malattia**
- I dati provengono da **solo 4 centri clinici → bassa diversità geografica e demografica**

Bias presenti nel dataset



Modelli di Machine Learning

- **Preprocessing**: One-Hot Encoding + StandardScaler
- **Modelli**: K-Nearest Neighbors e Logistic Regression
- **Metriche**: Accuracy, Recall, Precision, F1, ROC AUC
- **Fine-tuning** con GridSearchCV (cv=5) ottimizzando per Recall

Risultati Fine-Tuning dei Modelli

Modello	Accuracy	Recall	Precision	F1-Score	ROC AUC
K-Nearest Neighbors	0.822	0.762	0.842	0.800	0.881
Logistic Regression	0.822	0.738	0.861	0.795	0.928

Tabella 3.2: Prestazioni iniziali dei modelli sul test set.

Modello	Accuracy	Recall	Precision	F1-Score	ROC AUC
K-Nearest Neighbors	0.833	0.762	0.865	0.810	0.881
Logistic Regression	0.833	0.786	0.846	0.815	0.928

Tabella 3.5: Prestazioni dei modelli sul test set post fine-tuning.

GPT-4 vs Studente di Medicina

- **GPT-4** è stato interrogato tramite **l'API di OpenAI** con un **prompt strutturato** che includeva i dati clinici del paziente in esame

- Lo studente di medicina è stato interrogato con **lo stesso prompt** fornito a GPT-4

GPT-4 vs Studente di Medicina

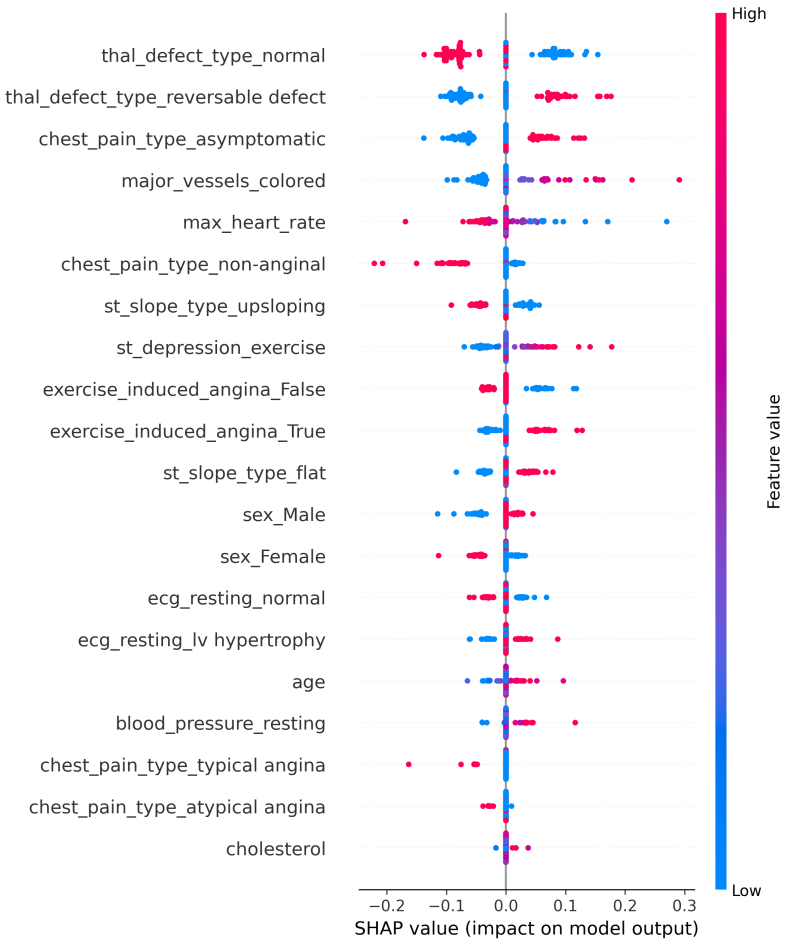
Modello	Accuracy	Recall	Precision	F1-Score
Logistic Regression	0.833	0.786	0.846	0.815
GPT-4	0.645	0.281	0.867	0.424
Studente Medicina	0.699	0.604	0.651	0.795

Tabella 3.7: Performance di Reg Log, GPT-4 e studente di medicina nella diagnosi di malattie cardiache.

Interpretabilità con SHAP

- Utilizzo di **SHapley Additive exPlanations (SHAP)** per interpretare il modello di regressione logistica ottimizzato.
- **SHAP** si basa sulla **teoria dei giochi**.
- Assegna a ogni feature un **valore di importanza** per ogni singola previsione.
- Spiega come ogni feature contribuisce a **modificare la previsione** rispetto al **valore base** (media delle previsioni).

Summary Plot SHAP



Summary Plot SHAP

- Il summary_plot permette di identificare le **feature più influenti** e la **direzione** del loro impatto.
- **Punti rossi** a **destra** indicano che **valori alti della feature aumentano la probabilità** della **classe positiva** (malattia); punti **blu** a **sinistra** indicano **l'opposto**.
- **Differenze rilevanti** rispetto alle **correlazione di Pearson**: la feature *exercise_induced_angina*, **molto correlata** con *sick* (ρ=**0.46**), non risulta tra le più rilevanti nei valori SHAP.
- Il livello di **colesterolo** è considerato **poco importante** dai valori SHAP, probabilmente perché è uniforme nel dataset e **poco correlato** a *sick* (ρ=**0.12**).

Web App – Sviluppo

- **API** con **Flask**:
  - `/models`
  - `/model_list`
  - `/predict`
- **Frontend** interattivo con **Streamlit** per inserimento parametri clinici
- Versionamento con **Git** e repository **GitHub**
- Containerizzazione completa con **Docker**

Conclusioni

- **Modelli classici di machine learning** sono **più efficaci** nella predizione di malattie cardiache rispetto ad approcci alternativi
- **Bias** e **limiti** del dataset richiedono ulteriori dati e validazione
- **Interpretazione** del modello con **SHAP** ha prodotto risultati **parzialmente rilevanti** dal punto di vista clinico ma **condizionati** dalla **limitata dimensione** del dataset.