DOCUMENTAZIONE CASO DI STUDIO

INGEGNERIA DELLA CONOSCENZA

SISTEMA INTELLIGENTE PER PREDIZIONE DI PROBLEMI CARDIOVASCOLARI

Gruppo di lavoro:

- **Michele Di Leo, 777750, m.dileo62@studenti.uniba.it**

Repository Github: https://github.com/Michelez25/ICON25_26.git

A.A. 2025/26

INDICE

1. INTRODUZIONE

Le malattie cardiovascolari rappresentano una delle principali cause di mortalità a livello globale e costituiscono una delle sfide più rilevanti per i sistemi sanitari moderni. Esse sono influenzate da una combinazione di fattori di rischio clinici, comportamentali e genetici, come i livelli di colesterolo, la pressione arteriosa, il fumo e la familiarità.

L'impiego di algoritmi intelligenti e tecniche di analisi dei dati consente di gestire e interpretare efficacemente tali informazioni complesse, offrendo un valido supporto alla prevenzione e alla diagnosi precoce. In questo contesto, il progetto si propone di sviluppare un sistema per la previsione del rischio di malattie cardiovascolari, con l'obiettivo di individuare tempestivamente i pazienti maggiormente esposti.

Il progetto mira a sviluppare un sistema per la previsione del rischio di malattie cardiovascolari e supportare la diagnosi medica identificando precocemente i pazienti a rischio.

Argomenti di interesse

Gli argomenti affrontati in questo progetto sono:

1. **Approccio Statistico (Machine Learning):** il modello viene addestrato attraverso un dataset in input per poi classificare i pazienti basandosi su dati storici;
2. **Approccio Logico (Sistema Esperto):** attraverso algoritmi logici, il modello fornisce una valutazione conclusiva.

---

2. APPRENDIMENTO SUPERVISIONATO

Il cuore dell'analisi predittiva è l'Apprendimento Supervisionato. Il modello viene addestrato su un dataset etichettato dove la variabile target (chd) è nota, imparando a mappare gli input (fattori di rischio) all'output (presenza o assenza di malattia).

2.1 Strumenti Utilizzati

Il progetto è stato sviluppato utilizzando il linguaggio Python e le seguenti librerie:

- **Pandas:** Per la manipolazione e l'analisi dei dati tabulari.
- **Scikit-Learn:** Per l'implementazione degli algoritmi di Machine Learning e le metriche di valutazione.
- **Matplotlib:** Per la visualizzazione dei risultati e dei grafici di ottimizzazione.
- **PySwip & SWI-Prolog:** Per l'integrazione della logica simbolica del Sistema Esperto.

2.2 Dataset

Per questo progetto è stato analizzato un campione di 462 casi studio estratti dal dataset del 1983 di Rousseauw in Sud Africa:

https://www.kaggle.com/datasets/waalbannyantu8dre/south-african-heart-disease-dataset/data

Le features principali sono:

- **sbp:** Pressione sanguigna sistolica;
- **tobacco:** Consumo cumulativo di tabacco (kg);
- **ldl:** Livello di colesterolo a bassa densità;
- **adiposity:** Indice di adiposità;
- **famhist:** Storia familiare di malattie cardiache (Presente/Assente);
- **typea:** Comportamento di tipo A;
- **obesity:** Indice di massa corporea: 30<x<34.9→ obesità lieve, 35<x<39.9→ obesità moderata, x>40→ obesità grave;
- **alcohol:** Consumo corrente di alcol;
- **age:** Età del paziente;
- **Target (chd):** 1 se presente malattia coronarica, 0 se assente;

2.2.1 Preprocessing del dataset

La fase di preprocessing (preprocessing_CHD.py) è iniziata con la rimozione della colonna ind (indice del paziente), in quanto non apporta valore predittivo. È stata verificata l'assenza di valori nulli e duplicati critici.

2.2.2 Conversione delle Features Categoriche

La colonna famhist conteneva valori testuali ('Present', 'Absent'). Poiché gli algoritmi di ML richiedono input numerici, è stata applicata una codifica binaria:

- Present → 1
- Absent → 0

Per visualizzare la frequenza con cui determinati valori compaiono nel dataset, sono stati creati dei grafici per confrontare visivamente come essi sono distribuiti rispetto alle relative caratteristiche:

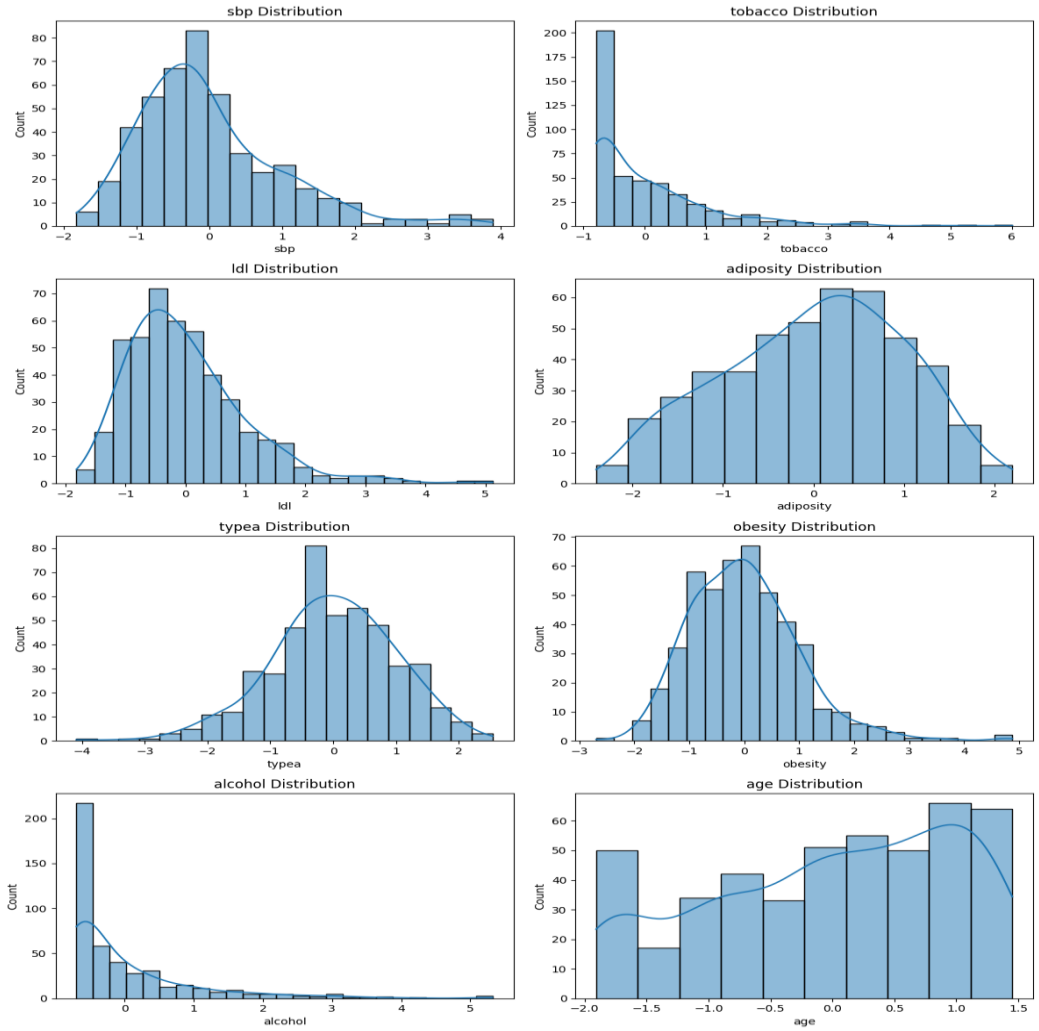


Figura 1. Distribuzione delle variabili nel dataset

2.3 MODEL TRAINING E VALUTAZIONE

Terminata la fase di Preprocessing, si procede con l'addestramento dei modelli predittivi. Per eseguire la classificazione del rischio cardiovascolare di un individuo, sono stati utilizzati diversi algoritmi di apprendimento supervisionato. Tenendo conto della natura clinica dei dati e della necessità di bilanciare prestazioni e interpretabilità, sono stati selezionati tre modelli principali: **Random Forest**, **Decision Tree** e **K-Nearest Neighbors (KNN).** All'interno del file train_val_CHD.py, i modelli sono stati inizialmente testati utilizzando una suddivisione del dataset che prevede l'80% dei dati come training set e il restante 20% come test set.

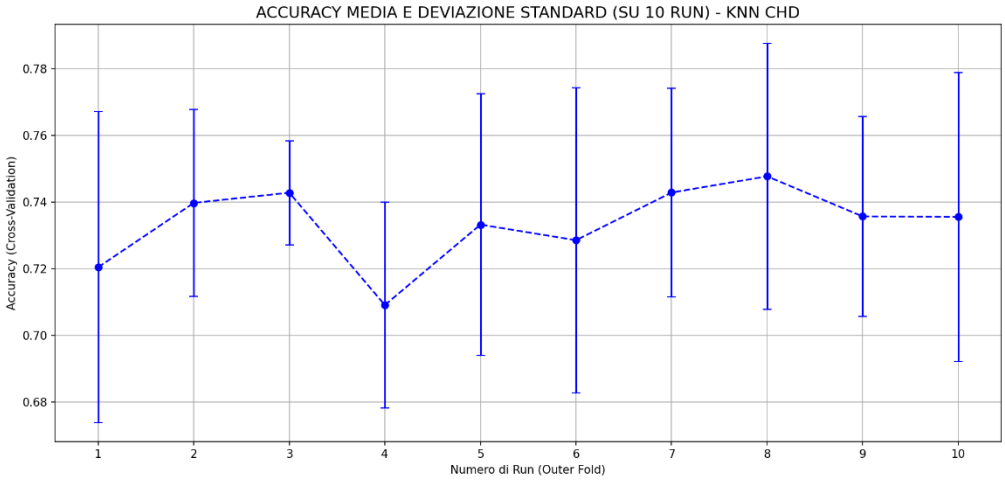
2.3.1 K-Nearest Neighbors (KNN)

Il K-Nearest Neighbors (KNN) è un algoritmo di apprendimento supervisionato il cui funzionamento si basa sul principio di "somiglianza". Per classificare il rischio di un nuovo paziente, l'algoritmo individua i "k" punti (pazienti) più vicini a esso nel set di addestramento e assegna la classe (presenza o assenza di malattia coronarica) più comune tra questi vicini.

Dato che il KNN è estremamente sensibile alla scala dei dati, è stata applicata una standardizzazione delle features per garantire che variabili con unità di misura diverse, come la pressione sanguigna (sbp) e l'età (age), contribuissero equamente al calcolo della distanza.

Per massimizzare le performance del modello ed evitare fenomeni di overfitting (con k troppo bassi) o underfitting (con k troppo alti), è stata condotta una ricerca sistematica dell'iperparametro ottimale n_neighbors:

- Metodologia: Sono state eseguite 10 run indipendenti, ciascuna caratterizzata da una diversa suddivisione casuale del dataset in training e test set (mantenendo la stratificazione della variabile target chd);
- Validazione: In ogni run sono stati testati 30 valori diversi di K (da 1 a 30) utilizzando una Cross-Validation a 5 fold per garantire una stima dell'accuratezza robusta e indipendente dal singolo split dei dati;
- Risultati: Attraverso il calcolo dell'accuratezza media e della deviazione standard su tutte le run, è stato identificato il valore di K che offre il miglior bilanciamento statistico per il dataset clinico in esame.



Il grafico mostra l'andamento dell'accuratezza in Cross-Validation al variare del numero di vicini (K) da 1 a 30. La linea tratteggiata rossa identifica il valore ottimale di K=15, che garantisce il miglior bilanciamento statistico per intercettare i soggetti a rischio.

Tabella riassuntiva delle run (KNN):

Run	CV Accuracy	CV Std
1	0.7205	0.0466
2	0.7398	0.0281
3	0.7428	0.0156
4	0.7091	0.0309
5	0.7332	0.0393
6	0.7286	0.0457
7	0.7429	0.0313
8	0.7477	0.0400
9	0.7357	0.0299
10	0.7355	0.0433

Dall'analisi dei risultati è emerso che i modelli presentano un'accuratezza media che si stabilizza in un range specifico, dimostrando come un valore di K intermedio permetta al classificatore di intercettare correttamente i soggetti a rischio nonostante il rumore statistico tipico dei dati medici.

2.3.2 Decision Tree

Il Decision Tree è un algoritmo di apprendimento supervisionato che opera creando una struttura a flusso, simile a un albero, dove ogni nodo interno decisionale rappresenta una "domanda" su una specifica feature dei dati (come la pressione arteriosa, il colesterolo o l'età) e ogni ramo rappresenta una risposta. Seguendo il percorso dall'alto verso il basso, si arriva a una "foglia" finale che contiene la previsione della classe: nel nostro caso, la presenza (1) o l'assenza (0) di patologie cardiovascolari.

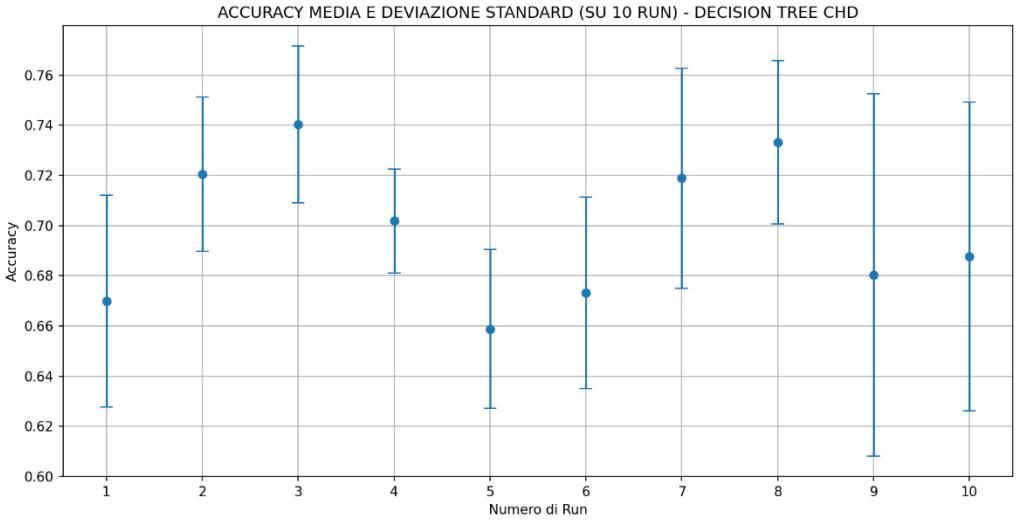
Per ottimizzare le prestazioni di questo modello, è stata adottata una strategia basata su 10 run indipendenti, ciascuna con una suddivisione casuale del dataset. Questo approccio permette di ottenere valutazioni più affidabili e stabili rispetto a una singola esecuzione, riducendo la variabilità dovuta allo split dei dati.

Per ogni run, il dataset è stato suddiviso con stratificazione della variabile target e si è utilizzata la tecnica della RandomizedSearchCV. Questa procedura esegue una ricerca casuale di combinazioni di iperparametri, valutandone l'efficacia mediante una Cross-Validation a 5 fold.

La procedura di tuning ha coinvolto i seguenti iperparametri principali:

- **max_depth**: profondità massima dell'albero per controllare l'overfitting, testata tra i valori [None, 3, 5, 10, 15, 20];
- **min_samples_split**: numero minimo di campioni necessari per suddividere un nodo, scelto tra i valori [2, 5, 10, 15];
- **min_samples_leaf**: numero minimo di campioni richiesti in una foglia, scelto tra i valori [1, 2, 4, 6].

In ciascuna run sono state valutate 15 combinazioni casuali di iperparametri per individuare quella in grado di garantire la migliore accuratezza media in Cross-Validation.



Il grafico riportato visualizza graficamente i risultati del tuning effettuato. Esso dimostra che il Decision Tree, se opportunamente ottimizzato, raggiunge buoni livelli di accuratezza. La lieve variabilità nei valori della deviazione standard tra le diverse run suggerisce comunque che il modello mantenga una certa sensibilità alla specifica configurazione dei dati clinici forniti in input.

Tabella riassuntiva delle run (Decision Tree):

Run	CV Accuracy	CV Std
1	0.6699	0.0422
2	0.7205	0.0309
3	0.7404	0.0312
4	0.7019	0.0207
5	0.6587	0.0317
6	0.6732	0.0381
7	0.7189	0.0438
8	0.7333	0.0326
9	0.6803	0.0722
10	0.6877	0.0616

I risultati delle 10 esecuzioni sono stati sintetizzati nella tabella soprastante che raccoglie per ogni run l'accuracy media ottenuta in Cross-Validation e la relativa deviazione standard. Dalla tabella si può osservare come l'accuracy media si mantenga su valori stabili, a testimonianza della robustezza del modello Decision Tree applicato al dataset CHD.

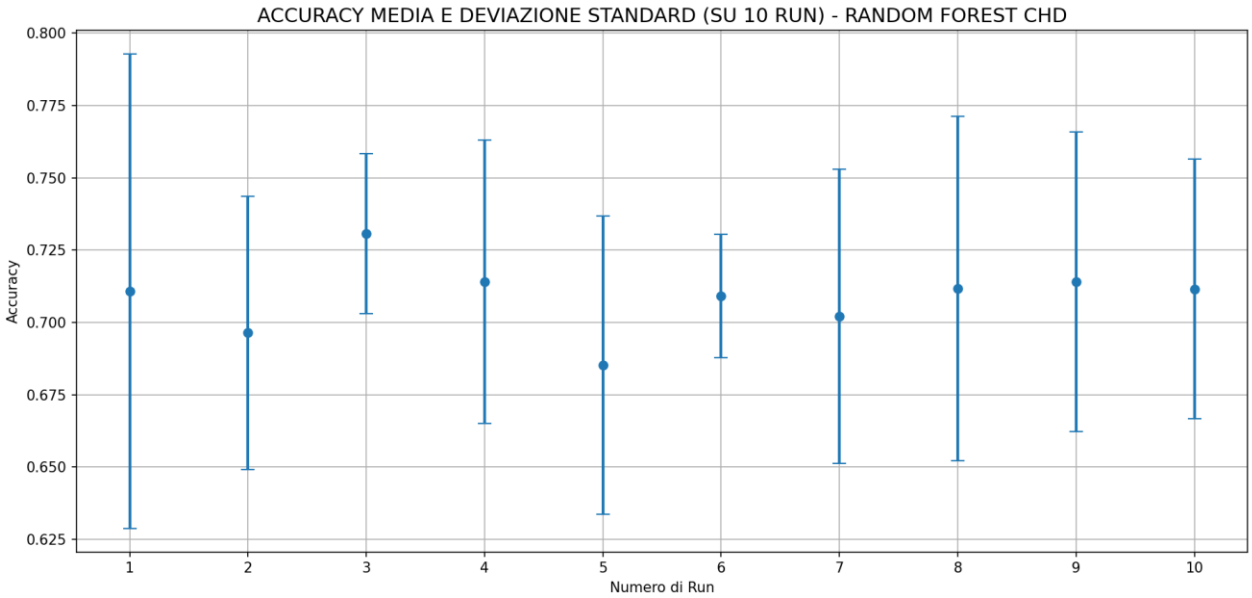
2.3.3 Random Forest

Il Random Forest è un algoritmo di apprendimento supervisionato di tipo ensemble che costruisce una moltitudine di alberi decisionali durante la fase di addestramento e aggrega i loro risultati per produrre una previsione finale. Questo approccio è noto per la sua robustezza e la sua capacità di ridurre l'overfitting rispetto a un singolo albero decisionale, gestendo meglio la varianza dei dati clinici.

Anche per questo modello, è stata adottata una metodologia basata sull'esecuzione di 10 run indipendenti, ciascuna con una diversa suddivisione casuale del dataset in training e test set (con stratificazione sulla variabile target chd). La ricerca degli iperparametri ottimali è stata eseguita mediante RandomizedSearchCV, testando 15 combinazioni casuali valutate con una Cross-Validation a 5 fold.

L'elenco degli iperparametri testati include:

- **n_estimators**: numero di alberi nella foresta, campionato casualmente tra 50 e 200;
- **max_depth**: profondità massima dell'albero per controllare l'overfitting, scelta tra i valori [None, 5, 10, 15];
- **min_samples_split**: numero minimo di campioni necessari per suddividere un nodo, scelto tra [2, 5, 10];
- **min_samples_leaf**: numero minimo di campioni richiesti in una foglia, scelto tra i valori [1, 2, 4].



Il grafico evidenzia che le performance si mantengono estremamente coerenti lungo le diverse run, con accuratezze medie stabili che testimoniano un'eccellente capacità di

generalizzazione del modello nonostante la complessità e il "rumore" statistico intrinseco dei fattori di rischio clinici forniti in input.

Tabella riassuntiva delle run (Random Forest):

Run	CV Accuracy	CV Std
1	0.7108	0.0821
2	0.6964	0.0472
3	0.7307	0.0277
4	0.7141	0.0489
5	0.6852	0.0515
6	0.7092	0.0214
7	0.7021	0.0510
8	0.7117	0.0595
9	0.7141	0.0517
10	0.7116	0.0449

L'ottimizzazione dei modelli ha prodotto un'accuratezza tra il 68% e il 73%, con il Random Forest come classificatore più stabile.

---

3. SISTEMA ESPERTO

Il Sistema Esperto è un programma di Intelligenza Artificiale progettato per simulare il processo decisionale di uno specialista umano in un dominio specifico, operando attraverso conoscenze esplicite e regole logiche. Nel presente progetto, il sistema è stato sviluppato per stimare il livello di rischio cardiovascolare basandosi su fattori clinici e comportamentali consolidati in letteratura, quali familiarità, età, fumo, pressione arteriosa e colesterolo LDL. L'architettura segue il modello classico dei sistemi esperti, articolandosi in tre componenti fondamentali:

1. Knowledge Base (Base di Conoscenza): l'insieme delle regole e delle informazioni del dominio medico;
2. Motore Inferenziale: il componente che applica la logica ai fatti per generare conclusioni;
3. Base dei Fatti: l'insieme dei dati dinamici relativi all'individuo analizzato.

3.1 Knowledge Base

La Knowledge Base è stata progettata seguendo un approccio Top-Down per superare i limiti di una rappresentazione puramente fattuale. La conoscenza è organizzata su tre livelli logici:

1. **Livello di Astrazione**: trasforma i dati numerici grezzi (es. SBP: 150) in concetti clinici simbolici (es. iperteso).
2. **Livello Euristico**: definisce regole mediche complesse e profili di emergenza (es. crisi_ipertensiva).
3. **Livello di Diagnosi (Goal)**: l'obiettivo finale del sistema, che scompone il problema principale in sotto-obiettivi verificabili.

3.1.1 Rappresentazione dei tratti di rischio

Le caratteristiche del paziente vengono caricate dinamicamente nel sistema sotto forma di fatti Prolog, utilizzando la seguente struttura: ha_tratto(Persona, Tratto, Valore).
Dove:

- **Persona:** identificativo univoco dell'utente;
- **Tratto:** variabile clinica;
- **Valore:** stato o misura rilevata per quel tratto.

I tratti considerati nel sistema includono:

- storia familiare di malattie cardiovascolari (famhist);
- consumo di tabacco (tobacco);
- età (age);
- pressione sistolica (sbp);
- livello di colesterolo LDL (ldl).

3.2 Logica e Criticità

Il motore inferenziale opera mediante un processo di **soddisfabilità logica**. Questo metodo permette al sistema di identificare immediatamente i **Profili di Emergenza**:

- **Gestione dei casi limite**: Se un singolo parametro vitale risulta estremamente fuori norma (ad esempio un livello di LDL > 250 mg/dl o una pressione sistolica > 180 mmHg), il sistema attiva una clausola di salvaguardia che assegna il 'Rischio Alto' indipendentemente dalla media degli altri valori.
- **Sicurezza Diagnostica**: Tale meccanismo garantisce che anomalie cliniche gravi non vengano ignorate o sottostimate nel calcolo complessivo, assicurando una valutazione prudenziale tipica del contesto medico .

3.3 Classificazione e Spiegabilità

Una caratteristica distintiva di questo sistema è la capacità di fornire una giustificazione logica per ogni conclusione raggiunta:

- **Trasparenza**: Grazie al predicato di spiegazione, il software non restituisce solo un'etichetta di rischio (Alto, Medio, Basso), ma elenca esplicitamente i tratti clinici e le regole mediche che hanno portato a tale esito .
- **Supporto Decisionale**: Questa funzionalità trasforma il sistema in un reale strumento di supporto alla diagnosi, permettendo al medico o all'utente di verificare il percorso logico seguito dall'algoritmo

---

4. CONCLUSIONI

Il sistema esperto sviluppato fornisce una valutazione interpretabile e trasparente del rischio cardiovascolare, basata su regole esplicite e facilmente comprensibili.

A differenza dei modelli di Machine Learning, il sistema consente di:

- spiegare il motivo di una classificazione;
- modificare manualmente pesi e soglie;
- integrare conoscenza medica esperta in modo diretto.

Questo lo rende particolarmente adatto come strumento di supporto decisionale, affiancabile a modelli statistici o predittivi più complessi.

I modelli presentano un'accuratezza media compresa tra il 69% e il 73%. Sebbene tali valori possano apparire moderati, l'analisi del Decision Tree evidenzia una Recall del 69% sulla classe CHD, dimostrando una buona capacità di intercettare i soggetti effettivamente a rischio. La difficoltà nel raggiungere accuracy superiori è legata alla natura stessa del dataset clinico, caratterizzato da un forte rumore statistico e una sovrapposizione delle distribuzioni dei fattori di rischio come sbp, ldl e tobacco. Proprio per sopperire a questi limiti statistici, è stato integrato il Sistema Esperto, che permette di validare le predizioni attraverso regole logiche trasparenti e spiegabili.