Università degli Studi di Milano - Bicocca
Scuola di Scienze
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di laurea magistrale in Informatica

Machine Learning per la Diagnosi del Rischio di Infarto

Dantonio Lorenzo 899759

Raia Lorenzo 894535

Introduzione e Obiettivi

Il presente elaborato si pone l'obiettivo di sviluppare un modello predittivo in grado di identificare la presenza di malattie cardiache in un paziente. L'approccio adottato mira a valorizzare le informazioni cliniche e i parametri fisiologici disponibili a seguito di indagini mediche, trasformando i dati diagnostici in stime probabilistiche relative al rischio cardiovascolare.

Per l'addestramento e la validazione del modello è stato utilizzato uno storico dataset risalente al 1988, che consolida le informazioni provenienti da quattro diverse strutture cliniche. Tale base di dati integra variabili eterogenee, concentrandosi su un sottoinsieme di 14 attributi chiave (rispetto ai 76 originariamente raccolti). Questi spaziano da fattori demografici (età, sesso) a metriche cliniche di performance cardiaca, quali la tipologia di dolore toracico, la pressione sanguigna a riposo, i livelli di colesterolo sierico e i risultati elettrocardiografici. La variabile target oggetto di studio è di natura binaria, e indica l'assenza (0) o la presenza (1) della patologia.

Analisi Esplorativa dei Dati (EDA)

Descrizione del dataset e qualità del dato

Il dataset in esame, originariamente composto da 1025 osservazioni, è stato sottoposto a una fase di pulizia preliminare. La rimozione dei campioni duplicati ha ridotto il corpus a 302 istanze uniche, un passaggio metodologico fondamentale per garantire l'eliminazione di bias sistematici dovuti a misurazioni ripetute in fase di addestramento.

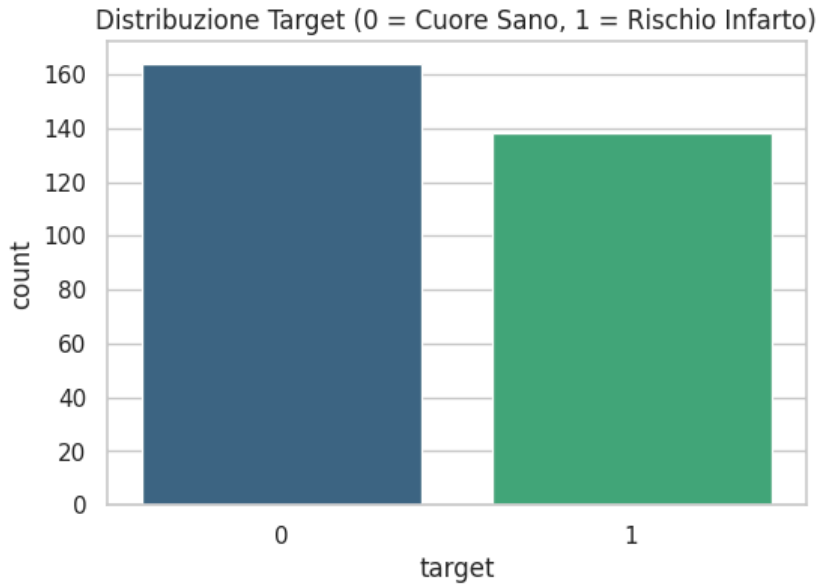
Le 14 feature presenti includono variabili fisiologiche e cliniche di tipo numerico (int64 e float64), quali l'età (age), la pressione arteriosa a riposo (trestbps), il colesterolo sierico (chol) e la depressione del tratto ST (oldpeak).

Un'analisi preliminare della qualità dei dati conferma l'assoluta integrità della struttura: l'esplorazione diagnostica non ha rilevato alcun valore mancante in nessuna delle feature. Questo azzera la necessità di ricorrere a tecniche di imputazione, permettendo ai modelli di apprendere da misurazioni cliniche reali e inalterate.

Analisi della Variabile Target e Coerenza Clinica

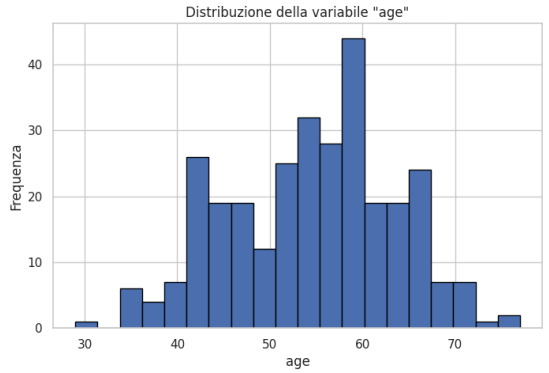
L'obiettivo del modello è la classificazione binaria della variabile target. Durante le prime fasi dell'esplorazione visiva, è emersa un'incongruenza tra le distribuzioni statistiche e la letteratura medica di base: i metadati originali associavano il valore 1 alla presenza della malattia e lo 0 all'assenza, ma i dati fisiologici suggerivano l'esatto opposto. Tale anomalia è stata peraltro confermata da numerose segnalazioni presenti nelle discussioni relative al dataset su Kaggle. Di conseguenza, le etichette del target sono state invertite tramite operazioni di pre-processing.

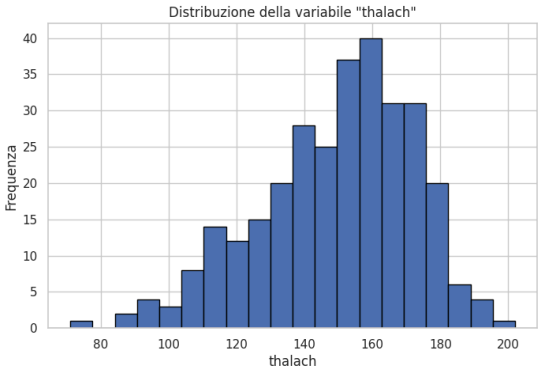
Attualmente, il target è correttamente mappato in 0 = Cuore Sano e 1 = Rischio Infarto. L'analisi della distribuzione rivela un dataset eccezionalmente bilanciato: la classe "Cuore Sano" conta 164 istanze, mentre la classe "Rischio Infarto" ne conta 138. Questo equilibrio strutturale rappresenta un vantaggio metodologico critico, in quanto elimina alla radice il rischio di sbilanciamento delle classi, consentendo l'utilizzo dell'Accuratezza globale come metrica di valutazione primaria, affiancata in modo sinergico a Recall e Precision.

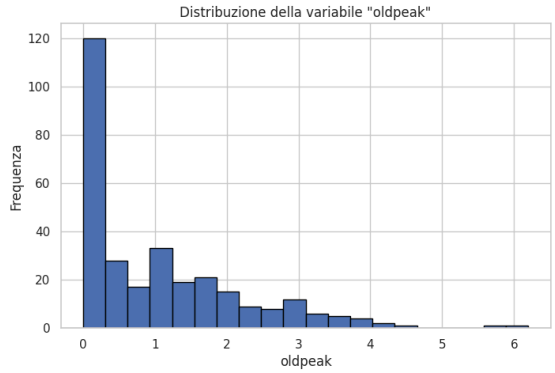


Distribuzione e Analisi Bivariata delle Feature Continue

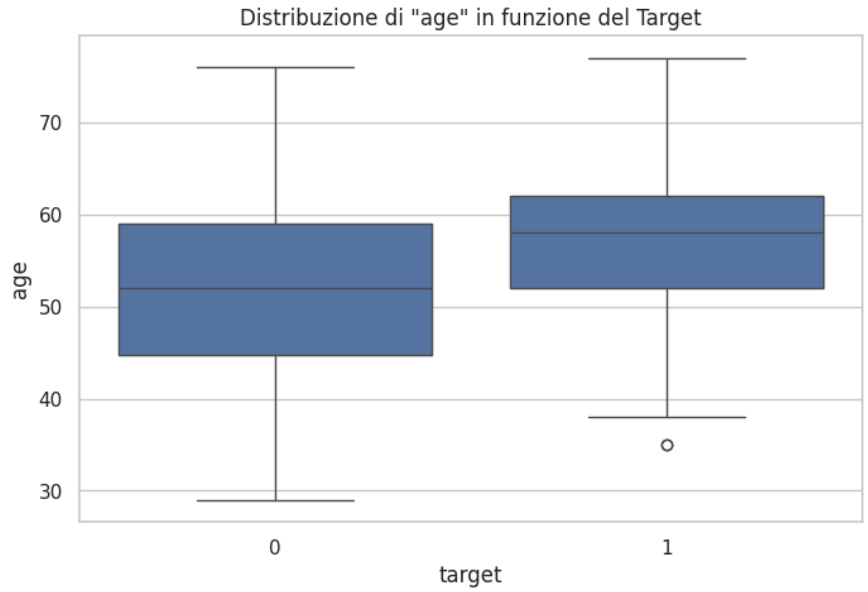
L'analisi visiva tramite istogrammi rivela distribuzioni eterogenee che rispecchiano le dinamiche fisiologiche della popolazione studiata. Variabili come l'età e la pressione arteriosa mostrano un andamento approssimativamente normale. La frequenza cardiaca massima (thalach) presenta una netta asimmetria negativa (left-skewed), mentre la variabile oldpeak evidenzia un decadimento quasi esponenziale, con la stragrande maggioranza dei valori concentrati in prossimità dello zero.



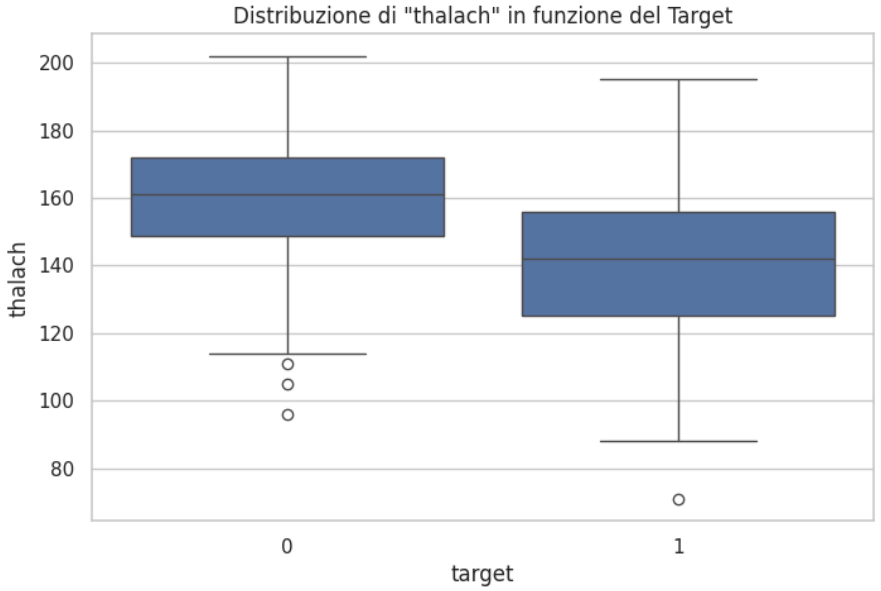




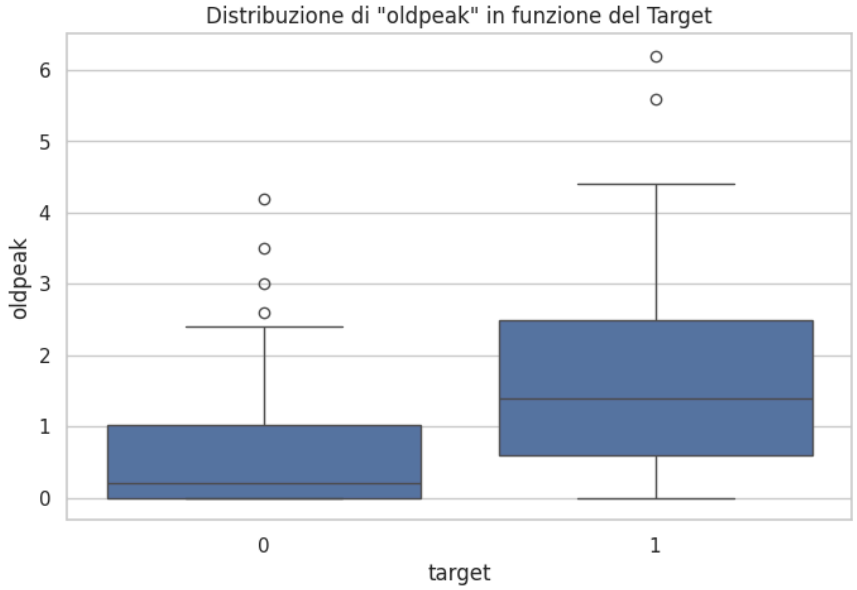
Incrociando queste distribuzioni con la variabile target tramite Boxplot, emergono pattern di notevole rilevanza predittiva:
Età (age): La classe a rischio presenta un'età mediana significativamente superiore (circa 58 anni) rispetto alla classe sana (circa 52 anni), confermando la natura degenerativa del rischio cardiovascolare.



Frequenza Cardiaca (thalach): Si osserva una capacità discriminante eccellente. I pazienti in salute mantengono regimi di frequenza cardiaca sotto sforzo nettamente superiori (mediana a 160 bpm) rispetto ai soggetti patologici, i cui valori risultano visibilmente ridotti.



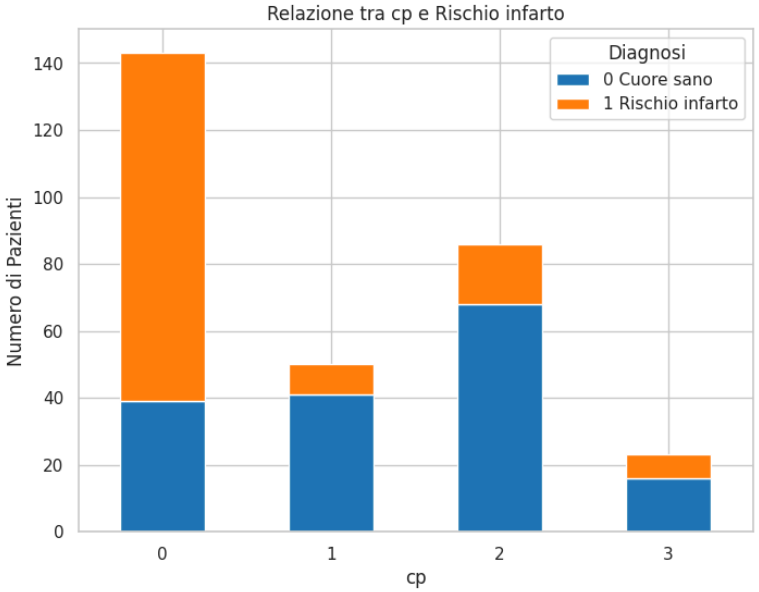
Depressione ST (oldpeak): I pazienti con esito patologico mostrano valori mediani marcatamente più elevati (circa 1.4) rispetto alla coorte di controllo, i cui valori risultano schiacciati verso lo zero, confermando questa anomalia elettrocardiografica come un forte segnale di ischemia.



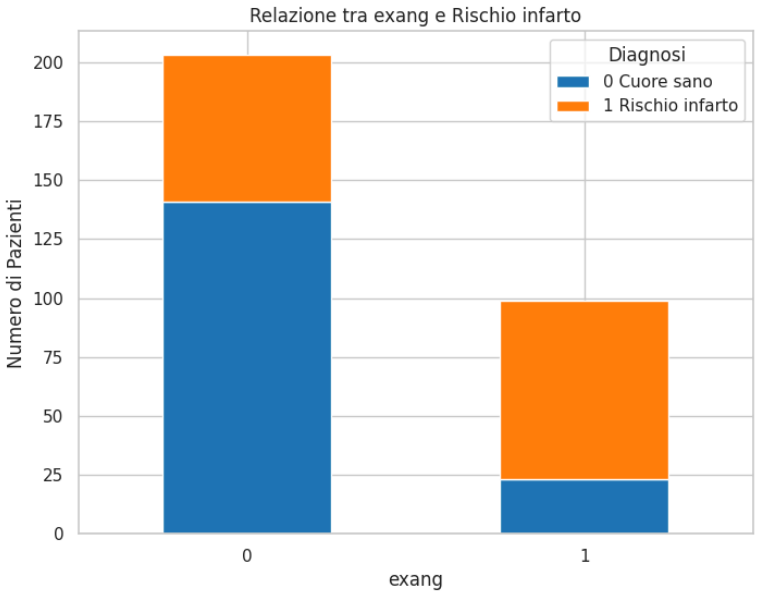
Analisi delle Feature Categoriche Predittive

L'analisi bivariata condotta sulle variabili categoriche, supportata da grafici a barre sovrapposte, ha permesso di isolare i predittori clinici più significativi, evidenziando variazioni drastiche della probabilità diagnostica in base a specifici fattori isolati:

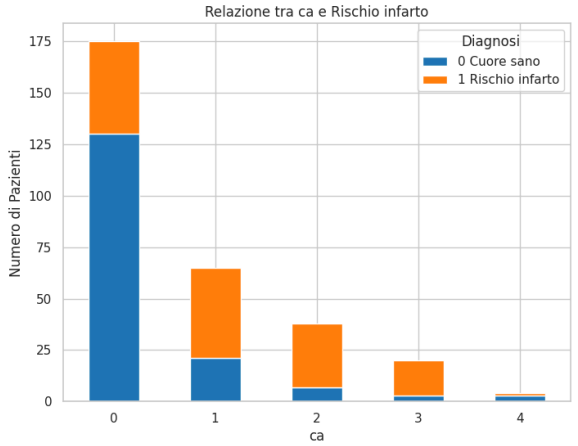
Dolore Toracico (cp): La tipologia di dolore toracico 0 è massicciamente associata alla presenza di rischio cardiovascolare. Al contrario, i valori 1, 2 e 3 si relazionano prevalentemente a un quadro clinico sano.

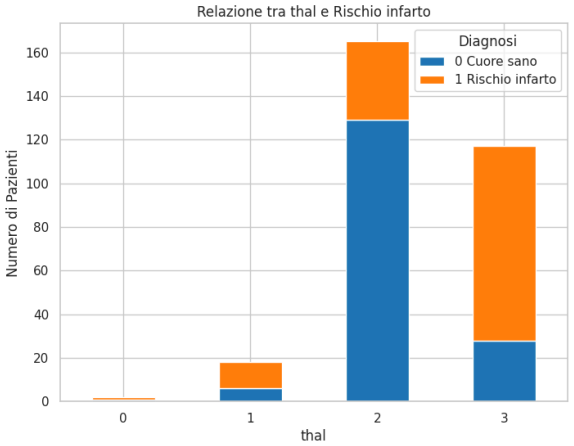


Angina Indotta dall'Esercizio (exang): Mostra una robusta correlazione positiva con la patologia. Nei pazienti in cui lo sforzo induce angina (exang=1), la proporzione di casi a rischio sovrasta nettamente i casi sani.



Il numero di vasi sanguigni occlusi agisce da forte discriminante: un valore ca=0 è fortemente indicativo di un cuore sano, mentre la presenza di occlusioni (ca>0) sposta radicalmente la probabilità verso la classe di rischio. Parallelamente, il test della talassemia con valore 2 è associato alla salute cardiaca, mentre il valore 3 indica un'alta probabilità di patologia.





Decision Tree

Scelte Progettuali

Rappresentazione dei Dati e Pre-processing

Un passaggio fondamentale nella definizione dell'architettura del modello ha riguardato il trattamento delle variabili categoriche non ordinali. L'approccio adottato nel pre-processing è stato l'applicazione di una codifica One-Hot Encoding sulle variabili nominali cp (tipo di dolore toracico), restecg (risultati elettrocardiografici a riposo), slope (pendenza del segmento ST) e thal (talassemia).

Questa trasformazione, implementata tramite la funzione *pd.get_dummies*, si è resa necessaria per evitare che l'algoritmo interpretasse erroneamente questi identificativi numerici come valori scalari dotati di un ordine gerarchico intrinseco. Sebbene questa tecnica aumenti la dimensionalità dello spazio delle feature, l'uso del parametro *drop_first=True* ha permesso di limitare il fenomeno della multi-collinearità, garantendo una scissione più pulita ed efficiente ai nodi dell'albero.

Gestione del Target e Split del Dataset

Per l'addestramento e la validazione del modello, il dataset è stato suddiviso in un set di Training (70%) e un set di Test (30%).
Data l'importanza clinica della classificazione, è stato utilizzato il parametro *stratify=y* per garantire che la medesima proporzione bilanciata tra pazienti sani e pazienti a rischio venisse rigorosamente mantenuta in entrambi i sottoinsiemi, evitando alterazioni stocastiche nella distribuzione delle classi.

Definizione dell'Architettura e Strategie di Regolarizzazione

Nella configurazione degli iperparametri del DecisionTreeClassifier, la scelta del criterio di scissione è ricaduta sull'Entropia. Questo approccio mira a massimizzare l'Information Gain ad ogni step, selezionando le feature cliniche che riducono maggiormente l'incertezza diagnostica tra i due esiti possibili.

Al fine di scongiurare fenomeni di overfitting sono state imposte rigorose strategie di regolarizzazione strutturale:

- Limitazione della profondità: Questo vincolo stringente opera come meccanismo di pre-pruning. Forza il modello a selezionare esclusivamente le feature dotate di maggior potere discriminante globale nei primissimi livelli gerarchici, impedendo la frammentazione eccessiva del dataset e garantendo un albero decisionale altamente interpretabile dal punto di vista medico.

- Vincolo di densità per lo split: L'impostazione di una soglia minima di 5 osservazioni per consentire la biforcazione di un nodo interno impedisce all'algoritmo di creare regole estremamente specifiche basate su pochissimi pazienti, favorendo la generalizzazione del modello.

Validazione Stabile e Cost Complexity Pruning

Per accertare la robustezza del classificatore e garantire che le performance non fossero dipendenti dalla specifica partizione di train/test iniziale, è stata implementata una rigorosa procedura di Repeated Stratified 10-Fold Cross-Validation (ripetuta 10 volte, per un totale di 100 validazioni indipendenti). Questo approccio garantisce una stima dell'Accuratezza estremamente affidabile, accompagnata dal calcolo di un intervallo di confidenza statistico al 95%.

Infine, attraverso l'estrazione del parametro di complessità *ccp_alpha*, il progetto prevede la possibilità di monitorare l'andamento dell'accuratezza su Train e Test al variare della penalità imposta per la complessità dell'albero, consentendo di individuare matematicamente l'indice di potatura che massimizza le performance sul set di validazione.

Discussione dei Risultati

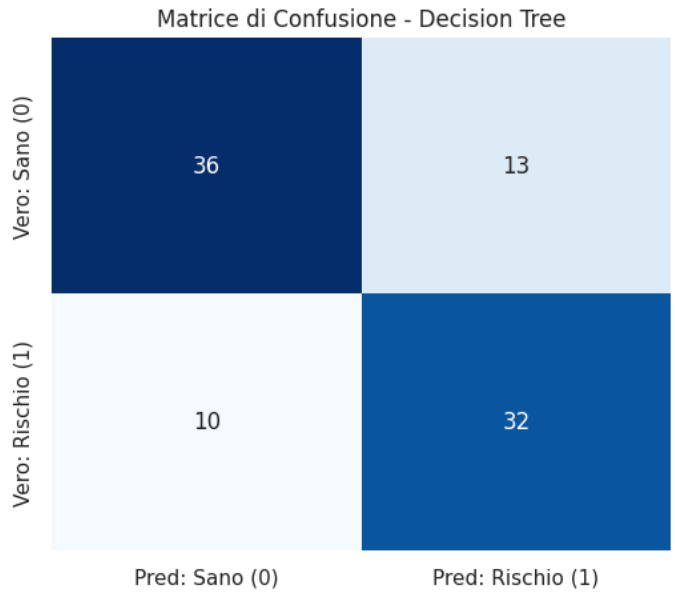
Analisi della Robustezza tramite Cross-Validation

Come già detto, al fine di valutare la stabilità e la capacità di generalizzazione del modello, mitigando il rischio di overfitting legato a una singola partizione dei dati, è stata eseguita una rigorosa procedura di *Repeated Stratified 10-Fold Cross-Validation*. Con 100 validazioni indipendenti eseguite (10 split ripetuti 10 volte), il modello ha registrato un'Accuratezza Media Stabile del 79.5%.

L'intervallo di confidenza al 95% risulta estremamente compatto *0.7822 - 0.8078*, a dimostrazione di una varianza statistica ridotta: le prestazioni del classificatore si mantengono coerenti e omogenee indipendentemente dalla specifica porzione di dataset utilizzata per l'addestramento. Il perfetto bilanciamento delle etichette in questo studio clinico permette di validare strutturalmente l'algoritmo basandosi in prima istanza sull'Accuratezza globale.

Analisi delle Performance di Classificazione sul Test Set

La valutazione del modello sui dati mai visti dal classificatore (Test Set, 30% del dataset totale) restituisce un'accuratezza complessiva del 75%. Sebbene fisiologicamente inferiore ai valori in Cross-Validation, questo risultato testimonia una buona aderenza del modello ai dati reali.

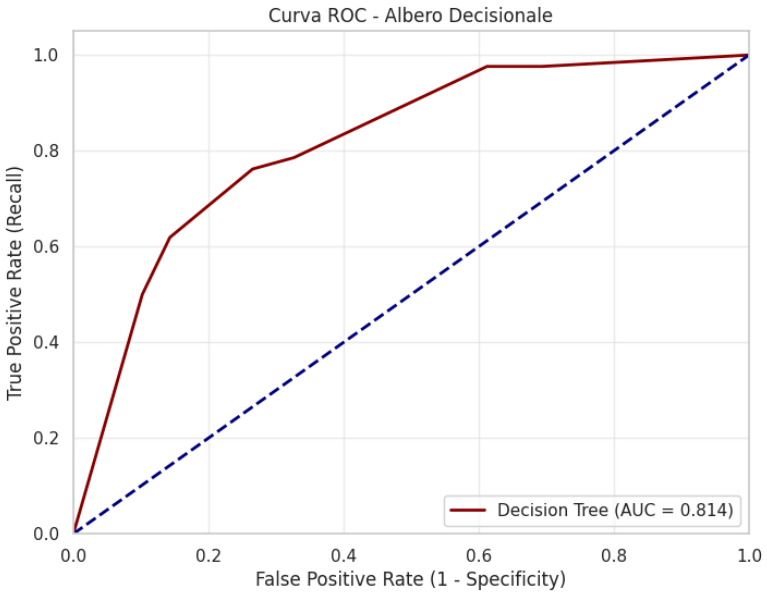


L'analisi del *Classification Report* e della Matrice di Confusione permette di indagare le dinamiche predittive sulle singole classi:

- **Identificazione del Rischio (Classe 1)**: Il modello ha identificato correttamente 32 pazienti patologici su 42 totali, registrando una **Recall dello 0.76**. I falsi negativi ammontano a 10 casi. Dal punto di vista clinico, minimizzare i falsi negativi è imperativo, e la rete ad albero si dimostra sufficientemente sensibile in questa direzione.
- **Identificazione del Cuore Sano (Classe 0)**: Il modello ha classificato correttamente 36 pazienti sani su 49, con una **Recall dello 0.73**. I falsi positivi ammontano a 13.
- **Precisione**: L'algoritmo mostra un buon bilanciamento anche in fase di precisione. Quando predice "Rischio Infarto", la previsione risulta corretta nel 71% dei casi (Precision 0.71), mentre per le diagnosi di "Cuore Sano" l'affidabilità sale al 78%. L'F1-Score bilanciato (0.76 per la classe 0 e 0.74 per la classe 1) conferma l'assenza di bias sistematici verso una specifica etichetta.

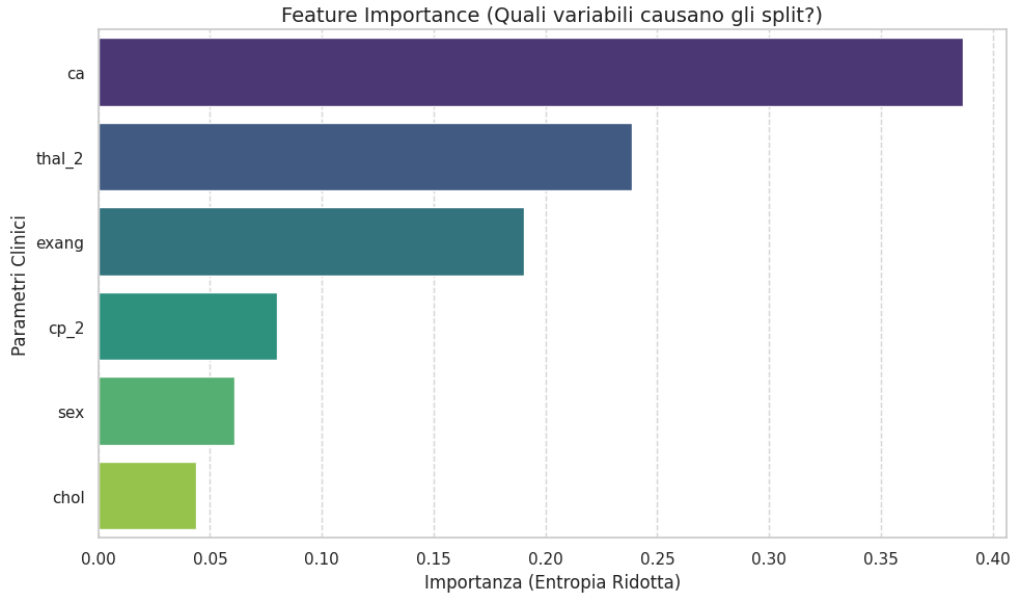
Capacità Discriminante (ROC-AUC)

La curva ROC misura la capacità del modello di separare le due popolazioni al variare delle soglie di decisione. L'area sottesa alla curva ottenuta dal Decision Tree è pari a 0.814. Questo punteggio, distante in modo significativo dal limite del predittore casuale 0.5, conferma una forte capacità discriminante globale, indicando che il classificatore assegnerà, con alta probabilità, uno score di rischio più elevato a un effettivo paziente cardiopatico rispetto a un individuo sano.

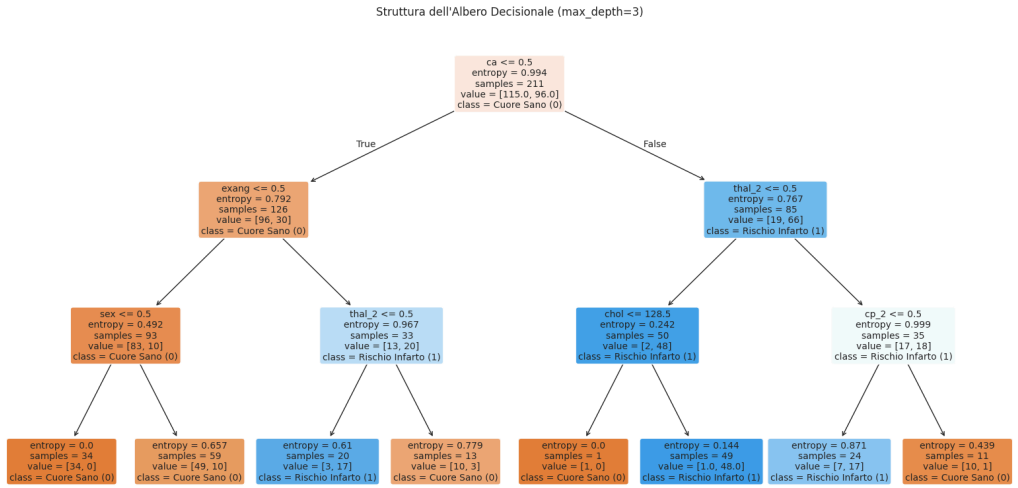


Rilevanza delle Feature e Struttura Diagnostica dell'Albero

L'analisi delle *Feature Importances* espone in modo cristallino i principali vettori causali delle scelte del modello, rivelando quali variabili contribuiscono maggiormente alla riduzione dell'entropia ai nodi. Emerge un pattern strettamente correlato alla letteratura clinica:
Le Occlusioni Strutturali (ca): dominano il processo decisionale, con un'importanza relativa che sfiora lo *0.40*. A seguire, i risultati della Talassemia e la presenza di Angina indotta dall'esercizio agiscono da forti discriminanti secondari.

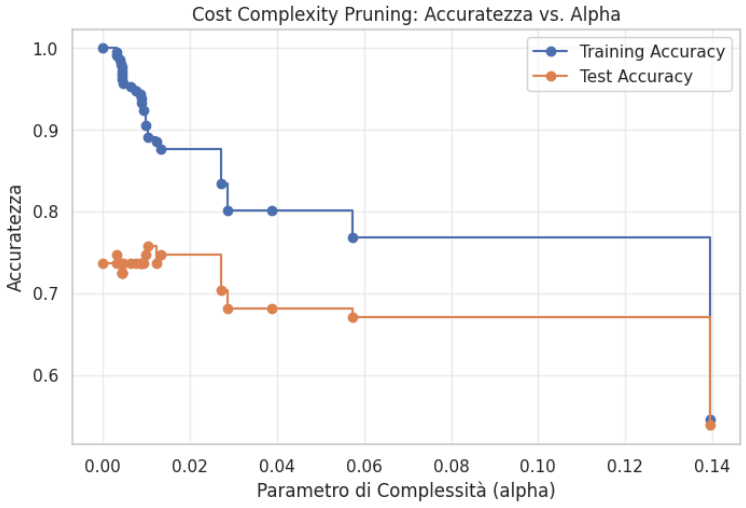


L'ispezione visiva della topologia dell'albero conferma operativamente questa gerarchia predittiva. La primissima scissione logica divide la popolazione sulla base dell'assenza o presenza di vasi parzialmente occlusi, instradando le decisioni verso percorsi clinici molto diversi. In presenza di vasi sani, ad esempio, l'albero scinde immediatamente la popolazione valutando l'angina da sforzo. Questo approccio del tipo *white-box* garantisce ai clinici una totale interpretabilità della stima finale fornita dall'intelligenza artificiale.



Dinamiche di Regolarizzazione

Per prevenire una specializzazione eccessiva sui dati di addestramento (*overfitting*), la struttura arborea è stata sottoposta a un'analisi di sensibilità del parametro di penalizzazione alpha. Il grafico illustra la dicotomia tra l'accuratezza di Training e l'accuratezza di Test al variare della complessità ammessa. All'aumentare della penalità, l'accuratezza di addestramento decresce in modo monotono a gradini, riflettendo la semplificazione della struttura dell'albero. Parallelamente, l'accuratezza sul Test set subisce fluttuazioni positive finché, estraendo empiricamente l'indice di picco, si identifica il *sweet-spot* matematico che massimizza la generalizzazione del modello sui casi clinici reali, prima di incorrere in condizioni di *underfitting*.



Support Vector Machine (SVM)

Scelte Progettuali

Rappresentazione dei Dati e Pre-processing

In modo analogo a quanto strutturato per il Decision Tree, la preparazione del dataset ha richiesto una rigorosa trasformazione delle variabili categoriche nominali (cp, restecg, slope, thal). L'applicazione della tecnica di *One-Hot Encoding* si è resa imperativa per evitare che l'algoritmo SVM, la cui logica si fonda su distanze geometriche e iperpiani, attribuisse un peso matematico errato alle etichette numeriche. L'utilizzo del parametro *drop_first=True* ha ulteriormente raffinato lo spazio delle feature, rimuovendo la ridondanza lineare e prevenendo il problema della multicollinearità. Il dataset così processato è stato suddiviso in set di Training (70%) e Test (30%), avvalendosi del campionamento stratificato (*stratify=y_all*) per preservare il perfetto bilanciamento delle diagnosi in entrambe le partizioni.

Normalizzazione dello Spazio delle Feature

Data la natura intrinsecamente geometrica delle Support Vector Machines, che determinano il confine di decisione calcolando le distanze tra i punti nello spazio vettoriale, l'assenza di una standardizzazione avrebbe compromesso la convergenza del modello. Variabili con ordini di grandezza ampi, come il colesterolo (*chol*) o la frequenza cardiaca massima (*thalach*), avrebbero dominato matematicamente il calcolo del margine rispetto a variabili binarie o su scala ridotta (come *oldpeak*).

A differenza di altri contesti dove si predilige la standardizzazione gaussiana, in questo progetto si è optato per la normalizzazione tramite *MinMaxScaler*. Questa scelta progettuale mappa tutte le variabili continue nell'intervallo [0, 1]. Tale approccio risulta particolarmente elegante ed efficace in presenza di un dataset arricchito da numerose colonne binarie (esito del *One-Hot Encoding*), mantenendo coerenza di scala su tutta la matrice di addestramento e non alterando la sparsità dei dati.

Ottimizzazione degli Iperparametri (Grid Search)

Per individuare la topologia ottimale dell'iperpiano di separazione, è stata implementata un'estesa *Grid Search* volta all'esplorazione combinata di tre iperparametri critici:

- Il parametro di regolarizzazione *C*: Analizzato su una scala logaritmica (0.1, 1, 10, 50, 100) per bilanciare il trade-off tra la massimizzazione del margine (Soft-Margin) e la penalizzazione degli errori di classificazione sul set di addestramento.
- La natura del Kernel: La ricerca ha messo in competizione il kernel lineare (per verificare l'esistenza di un confine di separazione lineare nello spazio dimensionale allargato) e il kernel rbf (Radial Basis Function, per mappare e intercettare eventuali relazioni cliniche complesse e non lineari).
- Il coefficiente Gamma: Parametro chiave per il kernel rbf, testato su valori decrescenti per calibrare il raggio di influenza dei singoli vettori di supporto.

A livello di inizializzazione, il modello è stato istanziato con il parametro *class_weight='balanced'* per garantire che il calcolo del margine non subisse distorsioni neanche da lievissime asimmetrie campionarie, e con *probability=True* per abilitare la trasformazione di Platt, essenziale per estrarre stime probabilistiche del rischio clinico e generare la curva ROC.

Strategia di Validazione: Repeated Stratified Cross-Validation

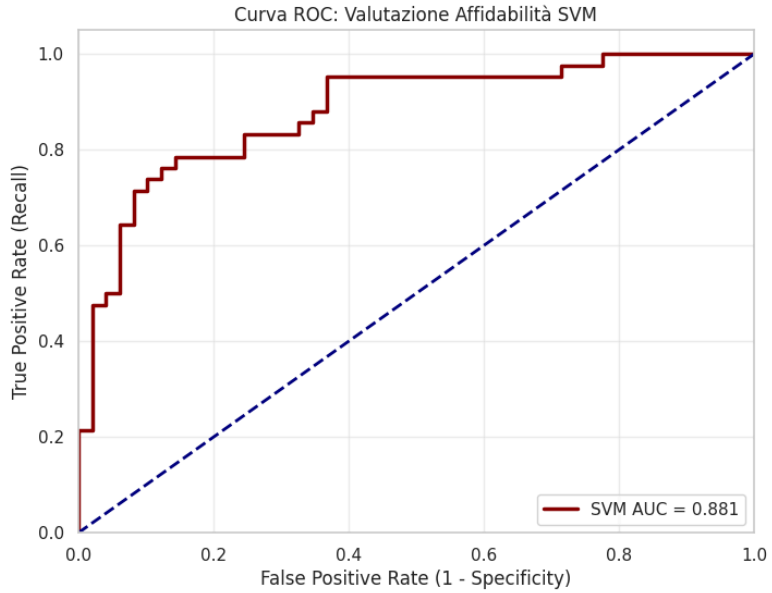
Analogamente a quanto eseguito per il modello ad albero, anche l'addestramento della SVM è stato supportato da una *Repeated Stratified 10-Fold Cross-Validation*. Questo onere computazionale annulla di fatto l'impatto delle fluttuazioni stocastiche derivanti dal partizionamento dei dati, assicurando che gli iperparametri ottimali individuati dalla Grid Search risultino strutturalmente robusti e massimizzino la capacità di generalizzazione diagnostica su nuovi casi clinici.

Discussione dei Risultati

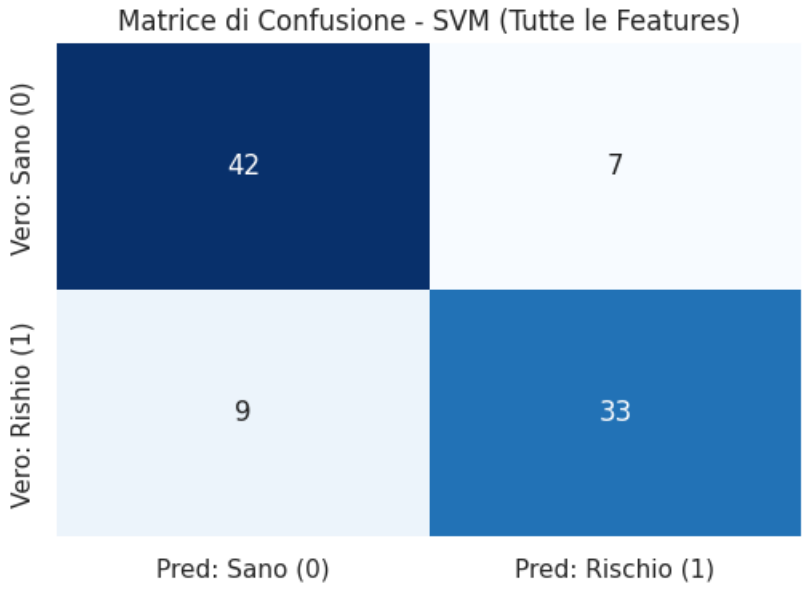
Analisi delle Performance di Classificazione e Capacità Discriminante

L'addestramento del modello SVM, validato attraverso la robusta procedura di *Repeated Stratified 10-Fold Cross-Validation*, ha restituito un'Accuratezza media sul set di training pari all'*83.78%*. Questo dato certifica fin da subito un'ottima convergenza dell'algoritmo in fase di apprendimento, garantendo una solida base per la generalizzazione sui nuovi pazienti.

La trasposizione delle predizioni sul Test Set ha permesso di tracciare la curva ROC. Il modello ha raggiunto un'Area Sottesa alla Curva (AUC) pari a *0.881*. Questo valore indica un'eccellente capacità discriminante globale: l'algoritmo possiede un'elevatissima probabilità di assegnare un punteggio di rischio maggiore a un paziente affetto da reale patologia cardiaca rispetto a un individuo clinico sano.



Matrice di Confusione e Affidabilità Clinica



La valutazione sul Test Set, composto da 91 istanze, restituisce un'accuratezza complessiva dell'82%. A differenza del Decision Tree, la SVM si distingue per un bilanciamento prestazionale straordinario tra le due classi, come si evince dal *Classification Report* e dalla Matrice di Confusione:

- Identificazione del Cuore Sano (Classe 0): Il modello dimostra una spiccata affidabilità nel confermare lo stato di salute dei pazienti, classificando correttamente 42 soggetti sani su 49. Questo si traduce in una Recall dello 0.86 e una Precisione dello 0.82. I falsi positivi (soggetti sani classificati a rischio) sono limitati a soli 7 casi.
- Identificazione del Rischio Infarto (Classe 1): Per quanto concerne i pazienti patologici, la SVM ne individua correttamente 33 su 42, registrando una Recall dello 0.79. Al contempo, la Precisione si mantiene su un elevato 0.82: quando il modello diagnostica un "Rischio Infarto", la probabilità che l'esito clinico sia effettivamente critico è dell'82%.
- Impatto dei Falsi Negativi: Si registrano 9 falsi negativi.
  Sebbene in un contesto clinico la riduzione dei falsi negativi sia un imperativo prioritario, l'F1-Score bilanciato (0.84 per la Classe 0 e 0.80 per la Classe 1) testimonia come il modello abbia ottimizzato matematicamente l'iperpiano di separazione senza introdurre *bias* sistematici verso l'etichetta maggioritaria.

Diagnostica Tecnica: Iperparametri e Kernel

Un dato estremamente interessante emerso dalla complessa procedura di Grid Search riguarda l'individuazione della configurazione ottimale degli iperparametri: *{'C': 100, 'gamma': 'scale', 'kernel': 'linear'}*.

- Scelta del Kernel Lineare: la selezione del kernel linear suggerisce che, una volta espanso lo spazio dimensionale tramite *One-Hot Encoding* e normalizzate accuratamente le feature continue, il confine di separazione clinica tra un cuore sano e un cuore a rischio può essere definito efficacemente da un iperpiano lineare. Questa evidenza riduce drasticamente il rischio di overfitting che tipicamente affligge i kernel non lineari (come l'RBF) in dataset clinici di dimensioni contenute.
- Parametro di Regolarizzazione (C = 100): L'impostazione di un valore C così elevato indica l'adozione di una strategia di *Hard-Margin*. L'algoritmo impone una severa penalizzazione per gli errori di classificazione sul training set, preferendo un margine geometrico più ristretto pur di classificare correttamente il maggior numero possibile di istanze mediche. Data l'importanza della precisione diagnostica, questa configurazione matematica riflette perfettamente le necessità del dominio di indagine.

Analisi Statistica e Validazione Metodologica

Applicazione del Paired T-test

Al fine di corroborare la validità statistica delle discrepanze prestazionali osservate tra il Decision Tree e la Support Vector Machine, ed escludere che tali variazioni siano imputabili a fluttuazioni stocastiche dipendenti dalla specifica partizione del dataset, si è proceduto all'applicazione di un Paired T-test. Tale rigorosa metodologia ha consentito il confronto diretto delle distribuzioni dei punteggi ROC-AUC emersi da una *Repeated Stratified 10-Fold Cross-Validation*, per un totale di 100 misurazioni indipendenti per ciascun modello. L'adozione del medesimo schema di partizionamento per entrambi gli algoritmi ha garantito la totale omogeneità delle condizioni sperimentali, assicurando che l'analisi della varianza tra le medie dei risultati avvenisse a parità di sottocampionamento.

Discussione della Significatività

L'esecuzione della procedura di validazione ha rilevato una Media AUC pari a 0.8579 per l'Albero Decisionale, superata dalla Media AUC registrata dalla SVM, che si è attestata a 0.8947 (con una ridotta deviazione standard di 0.0655). L'elaborazione della statistica-T (pari a -3.770) ha restituito un *p-value* estremamente basso, pari a 0.000277. Essendo tale valore enormemente inferiore alla soglia critica di significatività fissata a livello canonico (0.05), il test determina il fermo rifiuto dell'ipotesi nulla, la quale ipotizzava l'equivalenza probabilistica e prestazionale tra i due classificatori.

A supporto di tale evidenza, è stata condotta un'ulteriore indagine empirica incrementando iterativamente il numero di ripetizioni all'interno della *10-Fold Cross-Validation*. Si è osservato che, all'aumentare delle ripetizioni, il *p-value* continuava a diminuire in modo progressivo e costante. Dal punto di vista statistico e metodologico, questo comportamento ha un significato ben preciso: all'aumentare delle misurazioni cresce la "potenza" del test e la varianza delle stime si stabilizza, confermando che il divario prestazionale tra i due algoritmi è altamente sistematico. In altre parole, la diminuzione del *p-value* dimostra che il vantaggio della SVM non dipende in alcun modo da una partizione "fortunata" dei dati, ma si ripresenta in modo coerente e costante su qualsiasi possibile combinazione di training e test set.

Tale esito attesta in modo definitivo l'esistenza di una differenza statisticamente significativa tra le performance globali dei modelli; nello specifico, la superiorità della Support Vector Machine in termini di capacità discriminante (ROC-AUC) non è configurabile come un evento fortuito, bensì come un'evidenza strutturale della maggior capacità di questo algoritmo nel separare clinicamente i pazienti sani da quelli a rischio.

Conclusioni

Il presente elaborato si poneva l'obiettivo di sviluppare un ecosistema predittivo di Machine Learning capace di stimare il rischio cardiovascolare, massimizzando l'utilità clinica dei parametri diagnostici a disposizione. Sin dalle prime fasi, l'Analisi Esplorativa dei Dati (EDA) ha rivestito un ruolo critico: non solo ha confermato l'eccellente qualità e bilanciamento del dato, ma ha permesso di scovare e correggere preventivamente una profonda anomalia semantica nei metadati originali, riallineando la classificazione del *Target* alla reale letteratura medica.

Il confronto statistico, condotto e certificato tramite il Paired T-test, ha confermato matematicamente la netta superiorità della Support Vector Machine in termini di capacità discriminante globale. A differenza di altri domini applicativi, la topologia spaziale definita dalla SVM si è dimostrata l'approccio più affidabile per la diagnostica puntuale, garantendo l'Accuratezza più alta registrata sul set di test (82%). La configurazione trovata tramite Grid Search ha esibito un bilanciamento prestazionale eccellente, unendo un'alta Precisione (0.82) a una solida Recall (0.79) nell'individuazione clinica dei pazienti patologici, minimizzando così l'incidenza critica dei falsi negativi senza sovrastimare sistematicamente la malattia.

D'altro canto, l'Albero Decisionale, pur registrando performance globali fisiologicamente inferiori (Accuratezza del 75%), ha mantenuto un utilità ai fini dello studio. La sua intrinseca natura di modello *white-box* ha permesso di mappare gerarchicamente i fattori clinici dominanti offrendo un grado di interpretabilità totale sulle cause alla base della stima.

In sintesi, i modelli sviluppati si sono dimostrati in grado di processare indicatori clinici latenti con estrema precisione. La sinergia tra l'interpretabilità diagnostica offerta dal Decision Tree e la robustezza predittiva certificata dalla Support Vector Machine fornisce un sistema di supporto decisionale completo e affidabile per l'identificazione precoce delle patologie cardiache.