Università della Calabria

Dipartimento di Economia Statistica e Finanza

Corso di Laurea in **Data Science per le Strategie Aziendali**

Sviluppo di una Rete Bayesiana

Relatore

Prof. Paolo Carmelo Cozzucoli

Candidati

Pierfrancesco Lindia **256641**
Francesco De Nisi **269762**

Anno Accademico 2025 / 2026

Indice

1 Introduzione

Le reti bayesiane sono modelli grafici probabilistici che consentono di rappresentare in modo compatto ed interpretabile le relazioni di dipendenza tra un insieme di variabili aleatorie. Esse si basano su un grafo diretto aciclico (DAG), in cui i nodi rappresentano le variabili del dominio e gli archi diretti codificano dipendenze probabilistiche dirette. La struttura del DAG non ha solo un ruolo descrittivo, ma induce precise assunzioni di indipendenza condizionale tra le variabili, che permettono di fattorizzare la distribuzione di probabilità congiunta in un prodotto di distribuzioni locali. Questa proprietà rende le reti bayesiane particolarmente adatte a modellare sistemi complessi caratterizzati da incertezza, elevato numero di variabili e relazioni non deterministiche. In questo progetto le reti bayesiane vengono applicate allo studio della malattia cardiaca utilizzando il dataset Heart Disease – Cleveland. L'obiettivo principale è modellare la variabile target che indica la presenza o assenza di patologia cardiaca, integrando informazioni demografiche, cliniche, funzionali e diagnostiche. Poiché l'implementazione considerata richiede variabili discrete, le variabili continue del dataset sono state discretizzate e le variabili categoriali trattate come fattori. L'intero processo di modellazione è strutturato in più fasi: definizione del dominio e della variabile target, apprendimento della struttura del DAG sotto vincoli clinicamente motivati, validazione delle assunzioni di indipendenza condizionale, stima delle CPT e, infine, inferenza probabilistica. Questo approccio consente di ottenere un modello che sia al tempo stesso tecnicamente corretto, interpretabile dal punto di vista clinico e utilizzabile per analisi inferenziali.

2 Dataset e Variabile Target

Il dataset utilizzato per la costruzione della rete bayesiana è il *Heart Disease – Cleveland Dataset*, reso disponibile dall'UCI Machine Learning Repository. Il dataset contiene osservazioni relative a pazienti sottoposti a una serie di valutazioni cliniche e diagnostiche finalizzate alla rilevazione di patologie cardiache. Ciascuna osservazione rappresenta un singolo paziente ed è descritta da un insieme di variabili che includono caratteristiche demografiche, parametri clinici di base, risultati di test da sforzo ed esami diagnostici più avanzati.

Dal punto di vista modellistico, l'insieme delle variabili è stato interpretato come il dominio informativo di una rete bayesiana orientata alla diagnosi. In particolare, è stata individuata una variabile target che rappresenta l'esito finale del processo diagnostico: la presenza o assenza di malattia cardiaca. Nel dataset originale tale informazione è codificata tramite una variabile discreta multi-livello che indica il grado di severità della patologia; ai fini di questo progetto, tale variabile è stata trasformata in una variabile binaria, distinguendo tra soggetti sani e soggetti affetti da patologia cardiaca. Questa scelta consente di formulare il problema in termini di classificazione probabilistica e di interpretare direttamente le distribuzioni posteriori come probabilità di malattia.

Le variabili esplicative coprono diverse dimensioni del quadro clinico del paziente:

- **Variabili demografiche:** descrivono caratteristiche intrinseche non modificabili, come età e sesso.

- **Variabili cliniche di base:** includono parametri osservabili a riposo o in fase preliminare, quali pressione arteriosa, colesterolo sierico e risultati dell'elettrocardiogramma a riposo.
- **Variabili legate allo sforzo:** riguardano la risposta all'esercizio fisico e le alterazioni elettrocardiografiche indotte dallo sforzo.
- **Variabili diagnostiche avanzate:** forniscono informazioni più dirette sull'anatomia e sulla perfusione coronarica.

Questa eterogeneità informativa rende il dataset particolarmente adatto a essere modellato tramite una rete bayesiana, in quanto consente di rappresentare una progressione logica dalle caratteristiche di base del paziente fino all'esito diagnostico finale.

Poiché le implementazioni considerate nel progetto richiedono variabili discrete, i dati sono stati pre-processati come segue:

- le variabili continue presenti nel dataset sono state discretizzate in un numero limitato di classi ordinali;
- le variabili categoriali sono state trattate come fattori;
- i pochi valori mancanti sono stati gestiti mediante una procedura di imputazione semplice e riproducibile, così da garantire la completezza del campione utilizzato per l'apprendimento della rete.

Tali trasformazioni sono state effettuate con l'obiettivo di rendere i dati compatibili con il modello bayesiano discreto, preservando al contempo l'informazione rilevante dal punto di vista clinico.

2.1 Variabile Target

La variabile target del modello, denominata `hd` (*Heart Disease*), rappresenta la presenza o assenza di patologia cardiaca ed è interpretata come esito diagnostico finale. Nel dataset originale l'informazione relativa alla diagnosi è codificata dalla variabile `num`, che assume valori ordinali associati a diversi gradi di severità della patologia. Ai fini di questo progetto, e in linea con l'obiettivo di costruire un modello diagnostico interpretabile in termini di rischio/probabilità di malattia, la variabile `num` è stata trasformata in una variabile binaria `hd` mediante binarizzazione: `hd = no` per `num = 0` (assenza di malattia) e `hd = yes` per `num` ≥ 1 (presenza di malattia). Tale scelta consente di formulare il problema come classificazione probabilistica e di interpretare direttamente le distribuzioni posteriori come probabilità di malattia condizionate alle evidenze cliniche osservate.

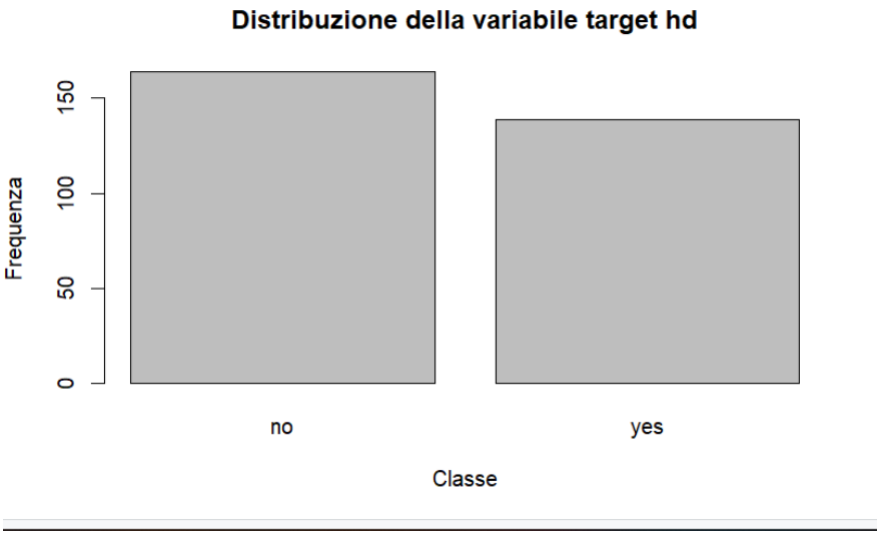


Figura 1: Distribuzione della variabile target `hd`

La Figura 1 mostra la distribuzione della variabile target `hd` nel campione considerato. Si osserva una leggera prevalenza di soggetti senza malattia cardiaca rispetto ai soggetti affetti, ma lo sbilanciamento tra le due classi risulta contenuto. Questo aspetto è rilevante dal punto di vista modellistico: in assenza di un forte sbilanciamento, la probabilità *a priori* di `hd` non è dominata da una singola classe e costituisce quindi un riferimento informativo significativo per valutare l'impatto delle evidenze in fase di inferenza. In particolare, nelle query inferenziali verrà analizzato come l'introduzione di informazioni cliniche e diagnostiche (ad esempio variabili da sforzo e diagnostica avanzata) modifichi la probabilità *a posteriori* di `hd`, rispetto al valore *a priori* indotto dalla distribuzione del campione.

2.2 Variabili esplicative

Le variabili esplicative del dataset descrivono diverse componenti del profilo clinico del paziente e costituiscono l'insieme di osservabili utilizzate per modellare probabilisticamente l'esito diagnostico rappresentato dalla variabile target *hd*. Il dominio informativo include caratteristiche demografiche, parametri clinici di base rilevati a riposo, indicatori di risposta allo sforzo e risultati di esami diagnostici avanzati. Questa eterogeneità informativa è particolarmente adatta alla modellazione tramite reti bayesiane, poiché consente di rappresentare in modo coerente una progressione logica dell'informazione: dalle caratteristiche intrinseche del paziente e dai fattori di rischio, passando per sintomi e segni clinici, fino a misure funzionali e diagnostiche ad elevata specificità.

Dal punto di vista clinico-semantico, le variabili esplicative possono essere organizzate come segue:

- **Variabili demografiche (*age*, *sex*):** rappresentano fattori a monte, non determinati dalle altre variabili del dominio e potenzialmente associati al rischio cardiovascolare.
- **Variabili cliniche di base:** includono

  - tipo di dolore toracico (*cp*),
  - pressione arteriosa a riposo (*trestbps*),
  - colesterolo sierico (*chol*),
  - indicatore di glicemia a digiuno elevata (*fbs*),
  - risultato dell'elettrocardiogramma a riposo (*restecg*).

  Tali informazioni sono tipicamente disponibili in fase iniziale di valutazione e contribuiscono a delineare lo stato clinico generale.

- **Variabili legate alla risposta allo sforzo:** comprendono
  - frequenza cardiaca massima raggiunta (*thalach*),
  - presenza di angina indotta da sforzo (*exang*),
  - depressione del tratto ST (*oldpeak*),
  - pendenza del tratto ST al picco dello sforzo (*slope*).

  Esse forniscono un'informazione funzionale rilevante nel contesto della sospetta ischemia miocardica.

- **Variabili diagnostiche avanzate (*ca*, *thal*):** forniscono informazioni più direttamente correlate alla patologia, rispettivamente il numero di vasi coronarici maggiori visualizzati e l'esito del test di perfusione miocardica. Queste variabili risultano clinicamente "vicine" alla diagnosi e dotate di alta capacità discriminante.

3 Preprocessing e discretizzazione

Prima dell'apprendimento della rete bayesiana è stata eseguita una fase di preprocessing finalizzata a rendere il dataset coerente, completo e compatibile con la modellazione tramite reti bayesiane discrete. Poiché i dati provengono da una sorgente eterogenea e includono variabili misurate su scale differenti (continue, binarie e categoriali codificate numericamente), è stato necessario definire un flusso di trasformazioni riproducibile che separi chiaramente: (i) la gestione dei valori mancanti, (ii) la trasformazione delle variabili nel formato richiesto dagli algoritmi, e (iii) la discretizzazione delle variabili continue.

Il primo passaggio ha riguardato il trattamento dei valori mancanti. Nel dataset originale alcuni attributi possono assumere il valore "?"; tali occorrenze sono state ricondotte a valori mancanti e successivamente gestite mediante imputazione, così da evitare sia la perdita di osservazioni (che ridurrebbe ulteriormente un campione già moderato) sia l'introduzione di procedure più complesse non richieste dall'obiettivo del progetto. In particolare, è stata adottata un'imputazione deterministica basata sulla moda (valore più frequente) per ciascuna variabile. Questa scelta garantisce semplicità e riproducibilità ed è adeguata in questo contesto poiché i missing sono pochi.

Poiché il modello adottato è una rete bayesiana discreta, è fondamentale che tutte le variabili siano rappresentate come variabili discrete con un insieme finito di stati

ben definiti. Nel dataset Cleveland molte variabili categoriali e binarie sono codificate tramite valori numerici.

Di seguito si riportano le codifiche adottate per alcune variabili esplicative del dataset, distinguendo tra codifica originale e rappresentazione finale utilizzata nel modello (rete bayesiana discreta).

- ***sex***
  Codici: 0 = femmina, 1 = maschio.
  Codifica finale: fattore binario (categorie, non misure).

- ***cp* (*chest pain type*)**
  Codici: 1 = angina tipica, 2 = angina atipica, 3 = dolore non anginoso, 4 = asintomatico.
  Codifica finale: fattore nominale a 4 stati.

- ***fbs* (*fasting blood sugar*)**
  Codici: 0 = no, 1 = yes (glicemia a digiuno > 120 mg/dl).
  Codifica finale: fattore binario.

- ***restecg* (ECG a riposo)**
  Codici: 0 = normale, 1 = ST–T anomalo, 2 = ipertrofia ventricolare.
  Codifica finale: fattore nominale a 3 stati.

- ***exang* (angina indotta da esercizio)**
  Codici: 0 = no, 1 = yes.
  Codifica finale: fattore binario.

- ***slope* (pendenza ST allo sforzo)**
  Codici: 1 = crescente, 2 = piatta, 3 = decrescente.
  Codifica finale: fattore (categorie cliniche).

- ***thal* (test di perfusione miocardica)**
  Codici: 3 = normale, 6 = difetto fisso, 7 = difetto reversibile.
  Codifica finale: fattore nominale.

- ***ca* (numero di vasi coronarici maggiori)**
  Valori: $0, 1, 2, 3$.
  Codifica finale: fattore discreto a 4 stati.

Alcune variabili originariamente continue sono state trasformate mediante discretizzazione in 3 classi definite tramite quantili.

- ***age* (età del paziente)**
  Significato clinico: età anagrafica del paziente (in anni).
  Tipo originale: variabile continua.
  Trasformazione: discretizzazione in 3 classi tramite quantili.
  Codifica finale: fattore ordinale (età bassa, media, alta).

- ***trestbps* (pressione arteriosa a riposo)**
  Significato clinico: pressione arteriosa sistolica misurata a riposo.
  Tipo originale: variabile continua.
  Trasformazione: discretizzazione in 3 classi tramite quantili.
  Codifica finale: fattore ordinale (pressione bassa, media, alta).

- ***chol* (colesterolo sierico)**
  Significato clinico: livello di colesterolo totale nel sangue.
  Tipo originale: variabile continua.
  Trasformazione: discretizzazione in 3 classi tramite quantili.
  Codifica finale: fattore ordinale (basso, medio, alto).

- ***thalach* (frequenza cardiaca massima raggiunta)**
  Significato clinico: massima frequenza cardiaca raggiunta durante il test da sforzo.
  Tipo originale: variabile continua.
  Trasformazione: discretizzazione in 3 classi tramite quantili.
  Codifica finale: fattore ordinale (bassa, media, alta).

- ***oldpeak* (depressione del tratto ST)**
  Significato clinico: entità della depressione del tratto ST indotta dall'esercizio.
  Tipo originale: variabile continua.
  Trasformazione: discretizzazione in 3 classi tramite quantili.
  Codifica finale: fattore ordinale (lieve, moderata, severa).

In sintesi, la codifica adottata distingue chiaramente tra variabili categoriali, trattate come fattori nominali, e variabili originariamente continue, discretizzate in classi ordinali. Questa scelta consente di utilizzare una rete bayesiana interamente discreta, riducendo la complessità parametrica e garantendo al contempo stabilità statistica e interpretabilità clinica delle dipendenze modellate.

4 Blacklist

Nel processo di apprendimento strutturale della rete bayesiana, l'impostazione di una *blacklist* rappresenta un passaggio cruciale per integrare conoscenza di dominio e garantire che il DAG appreso sia coerente con il fenomeno clinico analizzato. In una rete bayesiana, un arco diretto $X \to Y$ indica una dipendenza probabilistica diretta e, in un'ottica interpretativa, può essere letto come un'influenza "a monte" di X su Y. Di conseguenza, consentire archi non plausibili dal punto di vista medico porterebbe a una struttura difficilmente interpretabile e potenzialmente fuorviante.

Per questo motivo, la *blacklist* viene utilizzata come insieme di vincoli *negativi* che:

- impediscono all'algoritmo di Hill-Climbing di esplorare soluzioni strutturali logicamente incompatibili con la sequenza reale del processo diagnostico;

- riducono lo spazio di ricerca;

- favoriscono l'identificazione di relazioni robuste e clinicamente interpretabili.

Nel nostro caso, la variabile target è *hd*, che indica la presenza/assenza di malattia cardiaca ed è interpretata come esito diagnostico finale. Coerentemente con questa impostazione, è stato introdotto il vincolo più importante:

- *hd* non può avere archi uscenti.

Infatti, non è plausibile che la diagnosi finale "causi" retroattivamente età, sesso, sintomi o risultati dei test. Questo vincolo rende la rete compatibile con l'obiettivo del progetto (spiegare/predire l'esito a partire dagli attributi osservabili) e stabilizza l'apprendimento strutturale.

Oltre al vincolo sulla target, la *blacklist* è stata costruita seguendo una logica a livelli (*tiers*) che riflette la progressione temporale e clinica delle informazioni disponibili. In particolare, le variabili sono state organizzate come segue:

- **Livello 1 – Demografia:** *age*, *sex* (caratteristiche intrinseche del paziente, a monte del processo diagnostico).
- **Livello 2 – Clinica di base a riposo:** *cp*, *trestbps*, *chol*, *fbs*, *restecg*.
- **Livello 3 – Risposta allo sforzo / stress test:** *thalach*, *exang*, *oldpeak*, *slope*.
- **Livello 4 – Diagnostica avanzata:** *ca*, *thal* (esami tipicamente eseguiti in fasi successive e altamente informativi per la diagnosi).

In base a questa stratificazione, sono stati esclusi tutti gli archi che implicherebbero una causalità inversa, ad esempio impedendo che variabili diagnostiche avanzate influenzino caratteristiche demografiche o che l'esito diagnostico influenzi qualunque variabile antecedente. Questa scelta non impone a priori un'unica struttura "corretta", ma definisce un insieme di configurazioni ammissibili che restano statisticamente apprendibili dai dati e, al tempo stesso, clinicamente interpretabili.

In tal modo, il DAG appreso dall'algoritmo di Hill-Climbing risulta più facilmente giustificabile, poiché le dipendenze dirette individuate vengono cercate all'interno di un quadro logico coerente con il processo diagnostico e con l'interpretazione probabilistica del modello.

5 Apprendimento automatico della struttura

Una volta definiti i vincoli strutturali tramite la *blacklist*, si è proceduto all'apprendimento automatico della struttura del DAG mediante l'algoritmo di Hill-Climbing (HC). Tale algoritmo implementa una strategia di ricerca euristica che esplora lo spazio dei grafi diretti aciclici a partire da una struttura iniziale priva di archi, valutando iterativamente modifiche locali alla struttura e selezionando, a ogni passo, la configurazione che massimizza un criterio di score prefissato.

In particolare, a ogni iterazione l'algoritmo considera le seguenti operazioni locali:

- aggiunta di un arco;
- rimozione di un arco;
- inversione di un arco.

La procedura termina quando nessuna modifica locale consente di migliorare ulteriormente il valore dello score, individuando così una struttura localmente ottimale.

5.1 Criteri di score considerati

L'apprendimento strutturale è stato condotto utilizzando tre differenti criteri di valutazione: AIC, BIC e BDeu. Tali score differiscono per il modo in cui bilanciano la capacità descrittiva del modello e la complessità strutturale:

- **AIC:** tende a favorire strutture più ricche di dipendenze, privilegiando l'adattamento ai dati;
- **BIC:** introduce una penalizzazione più severa per il numero di parametri, producendo modelli generalmente più parsimoniosi;
- **BDeu:** adotta un approccio bayesiano e integra una regolarizzazione tramite la dimensione campionaria equivalente (*iss*), utile per attenuare l'influenza di dipendenze deboli o instabili, aspetto particolarmente rilevante in presenza di variabili discretizzate e di un campione di dimensione moderata.

5.2 Confronto tra le strutture apprese e scelta del modello finale

Il confronto tra le strutture apprese mediante i diversi score ha evidenziato che:

- AIC produce un DAG più denso, caratterizzato da un numero maggiore di archi;
- BIC restituisce una struttura più semplice e frammentata;
- BDeu (*iss*=10) rappresenta un compromesso equilibrato tra le due soluzioni, mantenendo un livello adeguato di complessità e preservando la stabilità delle relazioni apprese.

Alla luce di tali considerazioni, il DAG stimato con BDeu è stato selezionato come struttura finale del modello, in quanto maggiormente interpretabile e coerente con l'obiettivo diagnostico del progetto.

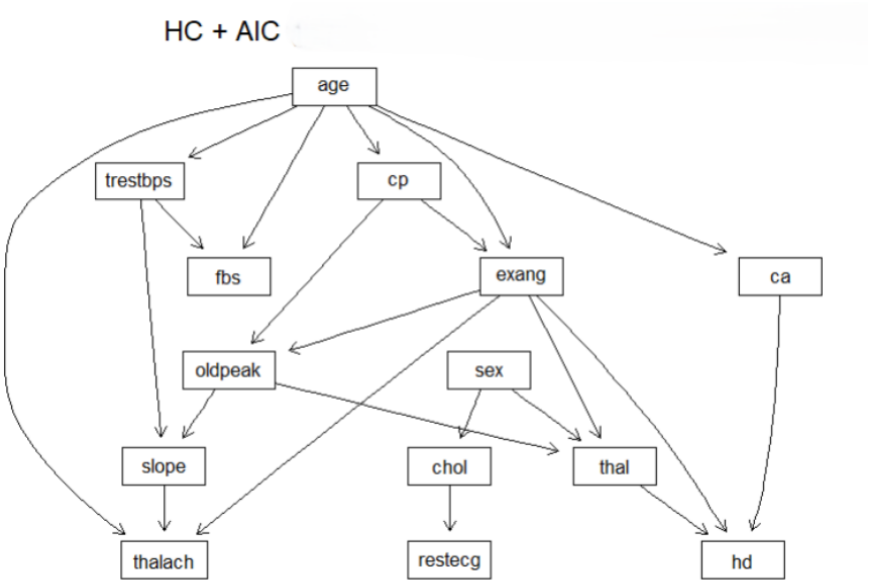
La struttura risultante mostra che la variabile target *hd* è direttamente influenzata da un insieme ristretto di variabili funzionali e diagnostiche, mentre i fattori demografici e clinici di base agiscono prevalentemente in modo indiretto. Tale organizzazione gerarchica delle dipendenze, emersa automaticamente dai dati sotto i vincoli imposti, conferma la capacità dell'approccio adottato di individuare relazioni probabilistiche significative senza sacrificare l'interpretabilità del modello.

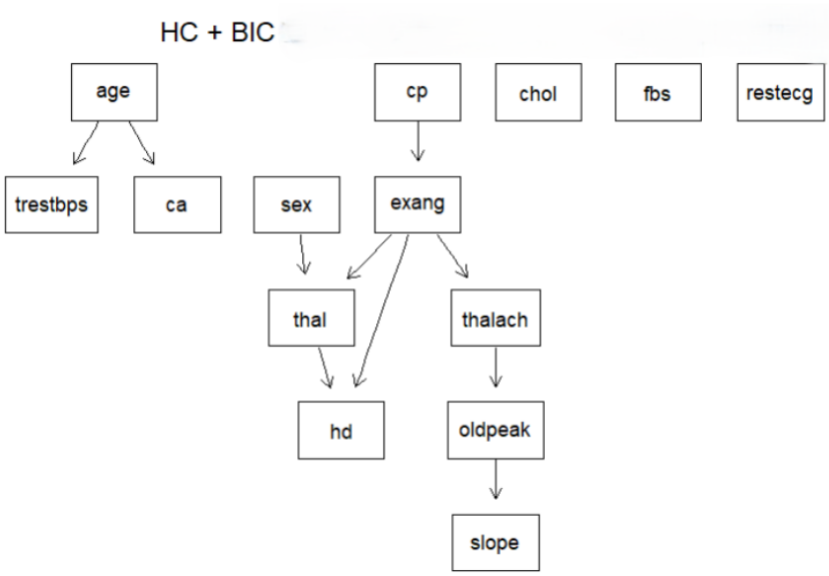
Tabella 1: Confronto dei criteri di score per i DAG appresi con Hill-Climbing.

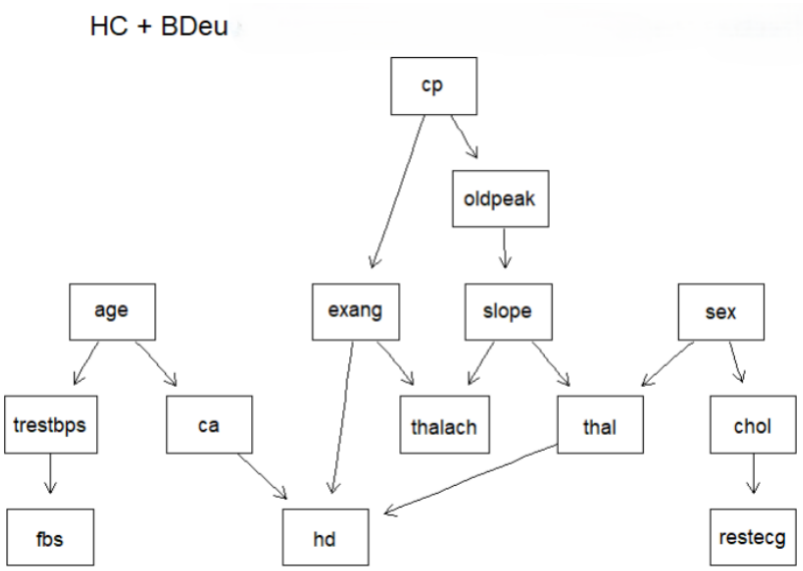
DAG	**AIC**	**BIC**	**BDeu (iss=10)**
HC + AIC	-3566.165	-3892.973	-3706.308
HC + BIC	-3612.662	-3724.074	-3682.788
HC + BDeu	-3575.917	-3761.603	-3666.298

Si osserva che i valori di AIC, BIC e BDeu non sono direttamente confrontabili tra loro in termini assoluti, poiché derivano da criteri di scoring differenti e presentano scale e penalizzazioni della complessità non omogenee. Il confronto risulta invece

significativo all'interno di ciascun criterio, valutando quali strutture ottengono il valore migliore secondo lo stesso score, e utilizzando la variazione tra criteri come analisi di sensibilità sulla robustezza della struttura appresa.







Dalle strutture mostrate in Figura, emerge come l'algoritmo di Hill-Climbing, applicato con differenti criteri di score, produca DAG con livelli di complessità e densità differenti. In particolare:

- l'utilizzo dello score AIC tende a generare una struttura più ricca di archi, privilegiando l'adattamento ai dati;
- il criterio BIC conduce a un grafo più parsimonioso, caratterizzato da un numero ridotto di dipendenze dirette.

La struttura appresa mediante HC con score BDeu rappresenta un compromesso efficace tra le due soluzioni, mantenendo un livello di complessità intermedio e preservando al contempo la stabilità delle relazioni probabilistiche apprese. Tale modello risulta inoltre particolarmente coerente con l'impostazione bayesiana del problema e con la natura discreta delle variabili considerate. Alla luce di queste considerazioni, il DAG ottenuto con HC+BDeu è stato selezionato come struttura finale del modello, in quanto maggiormente interpretabile e adeguato all'obiettivo diagnostico del progetto.

6 Analisi delle relazioni tra variabili ottenute

La struttura della rete bayesiana appresa mediante algoritmo Hill-Climbing con score BDeu, sotto i vincoli imposti dalla *blacklist*, evidenzia relazioni complessivamente coerenti con il dominio clinico della malattia cardiaca.

Blocco demografico e clinica di base Nel blocco demografico e di clinica di base, l'età risulta influenzare direttamente:

- la pressione arteriosa a riposo (*trestbps*);
- il numero di vasi coronarici coinvolti (*ca*);

riflettendo il ruolo dell'invecchiamento come fattore di rischio per ipertensione e aterosclerosi coronarica. La relazione tra pressione arteriosa e glicemia a digiuno (*trestbps* $\rightarrow$ *fbs*), pur non rappresentando una causalità diretta, è interpretabile come dipendenza statistica mediata da fattori metabolici condivisi, risultando quindi accettabile nel contesto del modello.

Sottografo sintomatologia e risposta allo sforzo Un sottografo particolarmente informativo riguarda la sintomatologia e la risposta allo sforzo. In particolare, il tipo di dolore toracico (*cp*) influenza:

- la presenza di angina indotta da esercizio (*exang*);
- l'entità della depressione del tratto ST (*oldpeak*);

delineando una catena fisiopatologica plausibile che prosegue con l'effetto di *oldpeak* sulla pendenza del tratto ST (*slope*) e, indirettamente, sulla frequenza cardiaca massima raggiunta (*thalach*). La presenza di angina da sforzo limita infatti la capacità di sostenere un elevato carico fisico, mentre le alterazioni elettrocardiografiche allo sforzo risultano coerentemente associate a una risposta cronotropa ridotta.

Diagnostica avanzata ed esito Per quanto riguarda la diagnostica avanzata, il numero di vasi coronarici visibili (*ca*) e il risultato del test di perfusione miocardica (*thal*) influenzano direttamente la variabile target *hd*, rappresentante la presenza di malattia cardiaca, in accordo con il loro ruolo di indicatori strutturali e funzionali diretti della patologia. Anche la frequenza cardiaca massima raggiunta (*thalach*) contribuisce alla previsione dell'esito, suggerendo che una ridotta capacità funzionale sia associata a una maggiore probabilità di malattia.

Associazioni metaboliche e ulteriori dipendenze Nel blocco metabolico e demografico emerge inoltre una relazione tra sesso e colesterolo sierico (*sex* $\rightarrow$ *chol*), coerente con le note differenze di profilo lipidico tra uomini e donne. Il legame tra colesterolo ed ECG a riposo (*chol* $\rightarrow$ *restecg*), pur più debole dal punto di vista causale, può essere interpretato come associazione indiretta legata a effetti cardiovascolari cronici.

Nel complesso, il DAG appreso presenta una struttura interpretabile, priva di relazioni logicamente errate e coerente con i vincoli imposti. La rete distingue chiaramente fattori demografici, variabili cliniche intermedie, diagnostica avanzata ed esito finale, fornendo una base solida per eventuali raffinamenti strutturali e per la successiva stima delle distribuzioni di probabilità condizionate.

7 Indipendenze implicate dal DAG e Test sui dati

L'interpretazione qualitativa degli archi non è sufficiente, da sola, a validare una struttura appresa. In una rete bayesiana, la struttura del DAG implica un insieme di relazioni di indipendenza condizionale tra variabili, derivabili tramite il criterio di *d-separation*. Una validazione strutturale coerente con l'impostazione teorica delle BN

consiste quindi nel verificare se alcune indipendenze condizionali rilevanti, implicate dal DAG, risultino compatibili con i dati osservati.

Operativamente, per ciascuna relazione selezionata si procede in due passaggi:

- **Verifica sul grafo:** si controlla che l'indipendenza sia effettivamente implicata dalla *d-separation* nel DAG;
- **Verifica sui dati:** si esegue un test di indipendenza condizionale sui dati osservati, qui basato sul test χ^2 (Pearson), poiché le variabili sono discrete.

In questo modo, l'obiettivo non è testare la "significatività" dei singoli archi, bensì controllare la coerenza tra le assunzioni globali del DAG e le dipendenze empiriche osservate.

Di seguito vengono riportate alcune verifiche mirate, selezionate perché clinicamente significative e direttamente connesse alle catene principali del DAG.

Verifica 1 — Mediazione nella catena sintomo/sforzo → esito Il DAG suggerisce che l'associazione tra *cp* (dolore toracico) e *hd* (malattia cardiaca) sia mediata dalle variabili di risposta allo sforzo, implicando l'indipendenza condizionale:

$$cp \perp hd \mid \{exang, oldpeak, thalach\}.$$

La *d-separation* sul grafo risulta verificata; il test χ^2 condizionato sui dati non evidenzia una dipendenza significativa (p-value = 0.124). Questo risultato indica che, fissate le variabili intermedie, l'informazione residua contenuta in *cp* non aggiunge evidenza statisticamente rilevante per *hd*, risultando coerente con l'ipotesi di mediazione implicata dal DAG.

Verifica 2 — Separazione metabolismo ↔ risposta allo sforzo condizionando sulla demografia In base al DAG, l'eventuale associazione tra *chol* (colesterolo) ed *exang* (angina da sforzo) è spiegata da fattori demografici, implicando:

$$chol \perp exang \mid \{age, sex\}.$$

La *d-separation* è verificata e il test χ^2 condizionato non mostra dipendenza significativa (p-value = 0.760). Pertanto, i dati risultano compatibili con l'assenza di un collegamento diretto tra metabolismo lipidico e angina da sforzo una volta fissati età e sesso, supportando la struttura appresa.

Verifica 3 — Anatomia coronarica e capacità funzionale condizionando sull'esito Il DAG implica che l'associazione tra *ca* (numero di vasi) e *thalach* (frequenza cardiaca massima) sia mediata dal quadro clinico complessivo rappresentato da *hd*, implicando:

$$ca \perp thalach \mid \{hd\}.$$

La *d-separation* è verificata e il test χ^2 condizionato non evidenzia dipendenza significativa (p-value = 0.214). Questo risultato è coerente con l'idea che la relazione tra anatomia coronarica e performance allo sforzo venga assorbita, nel modello, dall'informazione diagnostica complessiva legata a *hd*.

Le verifiche condotte non forniscono evidenze di incoerenza tra le indipendenze condizionali implicate dal DAG e le dipendenze osservate nei dati. Pur non costituendo una prova di correttezza causale della struttura (che non è l'obiettivo del progetto), questi risultati supportano la plausibilità statistica del DAG appreso sotto vincoli clinici e ne giustificano l'utilizzo come struttura baseline per la fase successiva di stima delle CPT e inferenza probabilistica.

Tabella 2: Verifiche di indipendenza condizionale implicate dal DAG appreso.

Verifica	**Indipendenza testata**	$\boldsymbol{\chi^2}$	**df**	p**-value**
1	$cp \perp hd \mid \{exang, oldpeak, thalach\}$	66.15	54	0.124
2	$chol \perp exang \mid \{age, sex\}$	8.31	12	0.760
3	$ca \perp thalach \mid \{hd\}$	15.53	12	0.214

Le indipendenze condizionali implicate dalla struttura del DAG sono state individuate mediante il criterio di d-separation e successivamente verificate sui dati tramite test di indipendenza condizionale basati sul test χ^2 di Pearson, utilizzando le funzioni dsep() e ci.test() della libreria `bnlearn`. Poiché tutte le variabili del modello sono discrete, il test χ^2 risulta appropriato per valutare la compatibilità tra le assunzioni strutturali del DAG e le dipendenze empiriche osservate. Nel complesso, le verifiche effettuate non evidenziano violazioni delle principali indipendenze condizionali implicate dal DAG appreso. Ciò supporta la plausibilità statistica della struttura selezionata, che viene pertanto assunta come struttura di riferimento per la successiva fase di stima dei parametri e inferenza probabilistica.

7.1 Markov Blanket della variabile target e riduzione del modello

In una rete bayesiana, la *Markov blanket* di un nodo X è l'insieme minimo di variabili che rende X indipendente da tutte le altre variabili del modello, ossia

$$X \perp\!\!\!\perp Rest \mid MB(X).$$

In generale, la Markov blanket è definita come

$$MB(X) = Pa(X) \cup Ch(X) \cup Pa\big(Ch(X)\big),$$

dove $Pa(X)$ denota l'insieme dei genitori e $Ch(X)$ l'insieme dei figli di X. Poiché l'obiettivo del progetto è l'inferenza diagnostica sulla variabile target *hd*, la Markov blanket fornisce un criterio teorico naturale per valutare una possibile riduzione del modello, mantenendo soltanto le variabili strettamente informative per la diagnosi.

Nel DAG appreso, la Markov blanket di *hd* coincide con l'insieme dei suoi genitori,

$$MB(hd) = \{exang,\ ca,\ thal\},$$

e *hd* non presenta figli, in coerenza con il vincolo strutturale imposto (assenza di archi uscenti dalla variabile target). Di conseguenza, una volta condizionato su $\{exang,\ ca,\ thal\}$, le variabili esterne alla Markov blanket non dovrebbero apportare informazione aggiuntiva su *hd* ai fini dell'inferenza.

Sulla base di questa osservazione, è stata valutata la possibilità di rimuovere alcune variabili appartenenti a sottografi laterali del DAG, in particolare *trestbps*, *fbs*, *chol* e *restecg*. Dal punto di vista strutturale, il DAG implica che tali variabili siano *d-separated* da *hd* una volta condizionato sull'insieme {*exang*, *ca*, *thal*}; tale implicazione è stata verificata formalmente tramite il criterio di *d-separation*.

La struttura del modello implica inoltre che *thalach* non dovrebbe fornire informazione aggiuntiva su *hd* una volta fissate le variabili della Markov blanket {*exang*, *ca*, *thal*}. Per valutare questa implicazione sul piano empirico, è stato eseguito un test di indipendenza condizionale basato sul test χ^2 di Pearson, verificando l'indipendenza tra *thalach* e *hd* condizionatamente a {*exang*, *ca*, *thal*}. Il test conferma pienamente l'ipotesi di indipendenza condizionale:

$$\chi^2 = 37.67, \qquad df = 48, \qquad p\text{-value} = 0.858.$$

Il valore di p molto elevato indica assenza di evidenza statistica contro l'ipotesi di indipendenza condizionale; in altre parole, nei dati non emerge alcuna dipendenza residua tra *thalach* e *hd* una volta condizionato su *exang*, *ca* e *thal*.

Per verificare che questa separazione sia coerente anche con i dati, sono stati eseguiti test di indipendenza condizionale basati sul test χ^2 di Pearson, valutando l'indipendenza tra ciascuna variabile candidata alla rimozione e *hd* condizionatamente a {*exang*, *ca*, *thal*}. In tutti i casi, il test non ha evidenziato dipendenze statisticamente significative, con p-value ampiamente superiori ai livelli convenzionali. Pur tenendo presente che un risultato "non significativo" non costituisce una dimostrazione formale di indipendenza (specie in presenza di configurazioni rare e celle sparse), l'evidenza empirica risulta compatibile con l'ipotesi che, fissata la Markov blanket di *hd*, le variabili *trestbps*, *fbs*, *chol* e *restecg* non aggiungano informazione rilevante per la diagnosi. Tali variabili risultano quindi candidate alla rimozione in un modello ridotto, a condizione che non si intenda utilizzarle come evidenza nelle query inferenziali.

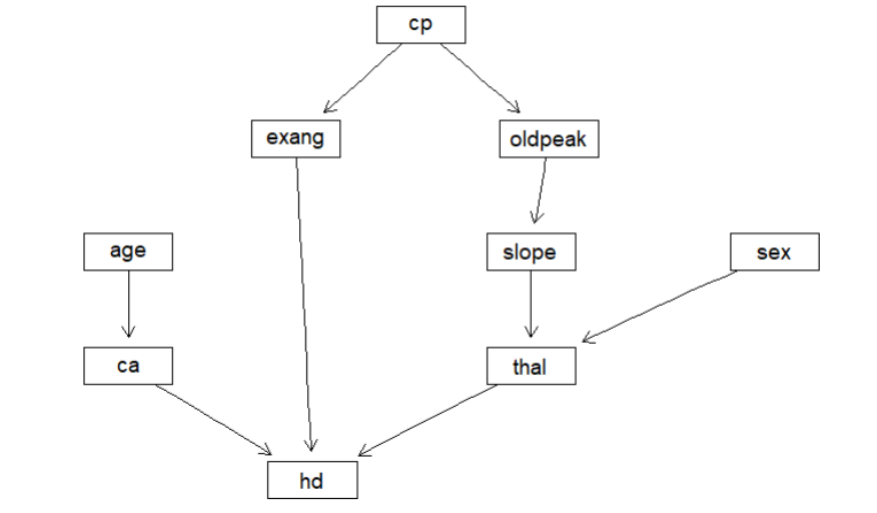


Figura 2: Enter Caption

La Figura mostra la struttura del modello ridotto ottenuta eliminando i sottografi laterali *trestbps–fbs* , *sex–chol–restecg* e *thalach*, mantenendo invariati i cammini rilevanti per l'inferenza diagnostica su *hd*. In particolare, il grafo preserva la catena

di risposta allo sforzo (*cp, oldpeak, slope, thalach, exang*) e i predittori diagnostici diretti della target (*exang, ca, thal*), che coincidono con la Markov blanket di *hd*. La riduzione è pertanto motivata sia teoricamente (Markov blanket) sia empiricamente (*d-separation* e test χ^2), ed è utilizzata come struttura alternativa per analisi successive, sotto l'ipotesi di non introdurre evidenze sulle variabili rimosse.

8 Stima delle distribuzioni di probabilità condizionate (CPT)

Una volta fissata la struttura della rete bayesiana, il passo successivo consiste nella stima dei parametri del modello, ovvero delle distribuzioni di probabilità condizionate (*Conditional Probability Tables*, CPT) associate a ciascun nodo del grafo. In una rete bayesiana discreta, ogni nodo è descritto da una distribuzione di probabilità condizionata ai propri genitori nel DAG, e l'insieme di tutte le CPT, insieme alla struttura del grafo, definisce completamente la distribuzione di probabilabilità congiunta sul dominio delle variabili.

Dal punto di vista metodologico, la stima delle CPT viene effettuata solo dopo aver definito e validato la struttura del DAG. Questa separazione tra apprendimento strutturale e apprendimento parametrico è fondamentale:

- la struttura codifica le ipotesi di dipendenza e indipendenza condizionale tra le variabili;
- i parametri quantificano l'intensità di tali relazioni sulla base dei dati osservati.

Una volta che la struttura è stata selezionata e ritenuta plausibile sia dal punto di vista clinico sia statistico, la stima dei parametri può essere condotta mantenendo il grafo fisso, evitando di confondere scelte strutturali con adattamenti locali ai dati.

8.1 CPT Nodi Radice

Nel modello ridotto, i nodi *age*, *sex* e *cp* non presentano genitori nel DAG e sono pertanto descritti da distribuzioni di probabilità marginali. Tali CPT rappresentano le probabilità *a priori* delle rispettive variabili e forniscono una descrizione sintetica della composizione del campione, costituendo il punto di partenza del modello prima dell'osservazione di qualunque evidenza. Le distribuzioni a priori stimate per i nodi radice sono riportate in Tabella 3.

Tabella 3: Distribuzioni di probabilità a priori dei nodi radice nel modello ridotto.

Variabile	**Categoria**	**Probabilità**
3**age*	bassa ($[29, 51]$)	0.349
	media ($(51, 59]$)	0.349
	alta ($(59, 77]$)	0.301
2**sex*	femmina (0)	0.326
	maschio (1)	0.674
4**cp*	angina tipica (1)	0.081
	angina atipica (2)	0.168
	dolore non anginoso (3)	0.283
	asintomatico (4)	0.468

Distribuzione a priori di *age* La distribuzione a priori dell'età (*age*), discretizzata in tre classi ordinali, risulta relativamente bilanciata tra le prime due classi, che presentano probabilità pressoché identiche, mentre la classe di età più elevata risulta leggermente meno frequente. Questo andamento riflette una popolazione prevalentemente di età medio-alta, coerente con il contesto clinico del dataset, ma senza una forte concentrazione in una singola classe.

Distribuzione a priori di *sex* Per quanto riguarda la variabile *sex*, la distribuzione marginale evidenzia una prevalenza di soggetti di sesso maschile rispetto a quelli di sesso femminile. Tale sbilanciamento è coerente con la composizione del campione del dataset *Heart Disease – Cleveland* ed è un aspetto rilevante da tenere in considerazione nell'interpretazione delle CPT a valle, in particolare per le variabili diagnostiche e per la variabile target.

Distribuzione a priori di *cp* La variabile *cp*, che descrive il tipo di dolore toracico, presenta una distribuzione fortemente asimmetrica, con una netta prevalenza delle categorie più severe. In particolare:

- la classe corrispondente al dolore toracico asintomatico risulta la più frequente;
- seguono le forme non anginose;
- le forme di angina tipica e atipica sono meno rappresentate.

Questo profilo è coerente con un campione clinico composto da pazienti sottoposti a valutazione diagnostica per sospetta patologia cardiaca e suggerisce che una quota rilevante dei soggetti presenti manifestazioni sintomatologiche non classiche.

8.2 CPT Nodi Intermedi

Dopo aver descritto le distribuzioni a priori dei nodi radice, l'analisi delle distribuzioni di probabilità condizionate prosegue concentrandosi sui nodi intermedi del modello ridotto. Tali variabili svolgono un ruolo di mediazione tra le caratteristiche di base del paziente e l'esito diagnostico finale, e la loro analisi consente di verificare che

le dipendenze strutturali apprese dal DAG si traducano in relazioni probabilistiche coerenti con il dominio clinico.

Per rendere l'esposizione più chiara e sistematica, le CPT dei nodi intermedi vengono analizzate raggruppando le variabili in blocchi funzionali omogenei, ciascuno dei quali rappresenta un aspetto specifico del processo clinico-diagnostico. In particolare, verranno considerati separatamente: (i) il blocco relativo alla sintomatologia e alla risposta allo sforzo, che descrive la catena che va dal dolore toracico alle alterazioni elettrocardiografiche e alla capacità funzionale; (ii) il blocco dell'anatomia coronarica, che media l'effetto dell'età sulla severità della patologia; e (iii) il blocco della diagnostica funzionale avanzata, che integra informazioni elettrocardiografiche e differenze di genere. Per ciascun blocco saranno riportati i valori numerici delle CPT più rilevanti e ne verranno discussi i principali pattern, mantenendo un livello di dettaglio adeguato all'interpretazione clinica del modello.

8.3 Sintomatologia e risposta allo sforzo

L'obiettivo di questa sezione è quantificare, tramite le CPT stimate, la catena di dipendenze emersa nel modello ridotto per quanto riguarda la sintomatologia e le risposte elettrocardiografiche allo sforzo. In particolare, il DAG identifica la seguente sequenza di relazioni probabilistiche:

$$cp \;\rightarrow\; (exang,\, oldpeak) \;\rightarrow\; slope.$$

Le CPT stimate quantificano in modo esplicito la catena di dipendenze del modello ridotto relativa alla sintomatologia e alle alterazioni elettrocardiografiche allo sforzo.

CPT di *exang* condizionata su *cp* In primo luogo, la CPT di *exang* condizionata su *cp* evidenzia che la probabilità di angina indotta da esercizio ($exang = 1$) varia sensibilmente al variare della categoria di dolore toracico: essa risulta relativamente contenuta per $cp = 1$ (circa 0.21) e per $cp = 2$ (circa 0.10), rimane moderata per $cp = 3$ (circa 0.14) e aumenta in modo marcato per $cp = 4$ (circa 0.55). Questo pattern suggerisce che, nel campione considerato, le categorie di dolore toracico si associano a differenti profili di risposta allo sforzo, con una maggiore probabilità di angina indotta nei soggetti classificati come $cp = 4$, coerentemente con la maggiore severità clinica implicata da tale categoria.

Tabella 4: CPT di *exang* condizionata su *cp*: probabilità $P(exang = 1 \mid cp)$.

cp	$\mathbf{P}(exang = \mathbf{1} \mid cp)$
1	0.21
2	0.10
3	0.14
4	0.55

CPT di *oldpeak* condizionata su *cp* In modo analogo, la CPT di *oldpeak* condizionata su *cp* mostra una variazione sistematica della severità delle alterazioni del tratto ST. Per $cp = 2$ la probabilità di osservare valori lievi di *oldpeak* ($[0, 0.1]$) è elevata (circa 0.61) e la classe severa ($(1.4, 6.2]$) è rara (circa 0.03). Al contrario, per $cp = 4$

aumenta la probabilità della classe severa (circa 0.40), mentre la classe lieve scende a circa 0.27. Nel complesso, i risultati indicano che le categorie di dolore toracico più "critiche" sono associate a una maggiore probabilità di alterazioni elettrocardiografiche più marcate allo sforzo, come atteso in un contesto di possibile ischemia.

Tabella 5: CPT di *oldpeak* condizionata su *cp*: distribuzione $P(\textit{oldpeak} \mid \textit{cp})$ per alcune categorie rilevanti.

cp	**Classe *oldpeak***	**Intervallo**	**Probabilità**
2	lieve	$[0, 0.1]$	0.61
2	severa	$(1.4, 6.2]$	0.03
4	lieve	$[0, 0.1]$	0.27
4	severa	$(1.4, 6.2]$	0.40

CPT di *slope* condizionata su *oldpeak* Infine, la CPT di *slope* condizionata su *oldpeak* completa la propagazione dell'informazione lungo la catena: quando *oldpeak* è lieve ($[0, 0.1]$), la probabilità di $\textit{slope} = 1$ è molto elevata (circa 0.82), mentre $\textit{slope} = 3$ è quasi assente (circa 0.02). Al crescere della depressione ST, la distribuzione si sposta verso pendenze più anomale: per *oldpeak* severo ($(1.4, 6.2]$), $\textit{slope} = 2$ diventa la modalità più probabile (circa 0.68) e aumenta sensibilmente anche $\textit{slope} = 3$ (circa 0.17). Questo andamento è coerente con l'interpretazione clinica congiunta delle variabili: depressione ST più elevata implica più frequentemente un andamento del tratto ST non fisiologico.

Tabella 6: CPT di *slope* condizionata su *oldpeak*: probabilità selezionate $P(\textit{slope} = k \mid \textit{oldpeak})$.

Condizione su *oldpeak*	**Esito**	**Probabilità**
$[0, 0.1]$	$P(\textit{slope} = 1 \mid \textit{oldpeak})$	0.82
$[0, 0.1]$	$P(\textit{slope} = 3 \mid \textit{oldpeak})$	0.02
$(1.4, 6.2]$	$P(\textit{slope} = 2 \mid \textit{oldpeak})$	0.68
$(1.4, 6.2]$	$P(\textit{slope} = 3 \mid \textit{oldpeak})$	0.17

Nel complesso, le tre CPT sono consistenti con la struttura del DAG e mostrano che l'informazione sul tipo di dolore toracico si propaga verso le variabili di sforzo/ECG secondo pattern numerici interpretabili dal punto di vista fisiopatologico, fornendo una base quantitativa solida per la successiva analisi delle CPT dei blocchi diagnostici e della variabile target.

8.4 Anatomia Coronarica

La componente di anatomia coronarica del modello ridotto è descritta dalla dipendenza diretta

$$\textit{age} \rightarrow \textit{ca},$$

che quantifica come il numero di vasi coronarici maggiori visualizzati (*ca*) vari al crescere della classe di età (*age*). La CPT stimata $P(\textit{ca} \mid \textit{age})$ (Tabella 7) evidenzia

un andamento coerente con l'interpretazione clinica: nella classe di età più bassa $[29, 51]$ la probabilità di osservare $ca = 0$ è elevata (≈ 0.81), mentre nelle classi di età successive tale probabilità diminuisce progressivamente fino a ≈ 0.53 per $(51, 59]$ e ≈ 0.38 per $(59, 77]$.

Parallelamente aumenta la probabilità di coinvolgimento multivasale. In particolare:

- $P(ca = 2 \mid age)$ passa da ≈ 0.04 nella classe più giovane a ≈ 0.12 nella classe intermedia e fino a ≈ 0.24 nella classe più anziana;
- $P(ca = 3 \mid age)$ cresce da ≈ 0.04 fino a ≈ 0.13.

Nel complesso, la CPT supporta l'idea che l'età agisca come fattore demografico a monte associato a una maggiore probabilità di compromissione coronarica, fornendo un collegamento probabilistico interpretabile tra il livello demografico e una variabile diagnostica che contribuisce direttamente alla predizione dell'esito *hd*.

Tabella 7: CPT del nodo *ca* condizionata su *age* nel modello ridotto: $P(ca \mid age)$.

ca	$[29, 51]$	$(51, 59]$	$(59, 77]$
0	0.813	0.529	0.380
1	0.099	0.300	0.253
2	0.044	0.117	0.242
3	0.044	0.053	0.125

8.5 Diagnostica funzionale avanzata

Nel modello ridotto finale, la variabile *thal* (esito del test di perfusione miocardica) è modellata come nodo figlio delle variabili *sex* e *slope*. Di conseguenza, la distribuzione locale associata a *thal* è:

$$P(thal \mid sex, slope).$$

La stima è stata effettuata mediante `bn.fit` con metodo bayesiano (`iss=10`), coerentemente con lo score BDeu utilizzato nella selezione strutturale. La CPT risulta correttamente normalizzata per ogni configurazione di *sex* e *slope*.

Dal punto di vista interpretativo, la CPT quantifica come differenze di genere e pattern elettrocardiografici allo sforzo (riassunti da *slope*) si associno a differenti esiti del test *thal*. In particolare, la probabilità di esito anormale ($thal \neq 3$) risulta pari a ... / ... nelle diverse combinazioni di *sex* e *slope*, mentre la probabilità di difetto reversibile ($thal = 7$) mostra ...

La CPT $P(thal \mid sex, slope)$ quantifica come l'esito del test di perfusione miocardica (*thal*) dipenda congiuntamente dall'andamento del tratto ST sotto sforzo (*slope*) e dal sesso biologico (*sex*). I risultati mostrano pattern coerenti sia con la progressione della severità elettrocardiografica sia con differenze di genere note in ambito cardiovascolare (Tabella 8).

Per $slope = 1$, associato a un andamento del tratto ST più fisiologico, la probabilità di un esito normale del test di perfusione ($thal = 3$) è molto elevata nei soggetti di sesso femminile (≈ 0.96) e rimane maggioritaria anche nei soggetti di sesso maschile

(≈ 0.60), sebbene in questi ultimi emerga una quota non trascurabile di esiti anomali, in particolare di tipo reversibile ($thal = 7$, ≈ 0.38).

All'aumentare della severità delle alterazioni elettrocardiografiche, il profilo di *thal* cambia sensibilmente. Per $slope = 2$, la probabilità di $thal = 3$ diminuisce in entrambi i sessi, passando a circa 0.70 nei soggetti di sesso femminile e a circa 0.26 nei soggetti di sesso maschile, mentre aumenta in modo marcato la probabilità di esiti patologici, in particolare $thal = 7$, che raggiunge valori di circa 0.27 e 0.61 rispettivamente. Un andamento analogo si osserva per $slope = 3$, dove la probabilità di esiti normali scende ulteriormente (≈ 0.53 per $sex = 0$ e ≈ 0.31 per $sex = 1$) e cresce la probabilità di esiti anomali sia reversibili sia fissi.

Nel complesso, la CPT evidenzia che l'esito del test di perfusione miocardica è fortemente modulato dalla severità delle alterazioni elettrocardiografiche allo sforzo e presenta differenze sistematiche tra i sessi. Questo comportamento è coerente con il ruolo di *thal* come variabile diagnostica funzionale avanzata, posta a valle della catena di sforzo/ECG e direttamente coinvolta, insieme a *exang* e *ca*, nella determinazione probabilistica dell'esito clinico finale *hd*.

Tabella 8: CPT del nodo *thal* condizionata su *sex* e *slope*: $P(thal \mid sex, slope)$.

slope	***sex***	***thal*=3**	***thal*=6**	***thal*=7**
1	0	0.957	0.011	0.032
1	1	0.595	0.026	0.378
2	0	0.698	0.033	0.269
2	1	0.264	0.130	0.606
3	0	0.533	0.083	0.383
3	1	0.314	0.201	0.484

8.6 CPT della variabile target

La distribuzione di probabilità condizionata (CPT) stimata per la variabile target *hd*, condizionata sulle variabili *exang*, *ca* e *thal*, consente di quantificare in modo esplicito l'effetto congiunto della risposta allo sforzo, dell'estensione dell'anatomia coronarica e della perfusione miocardica sulla probabilità di presenza di malattia cardiaca. Tali variabili costituiscono la Markov blanket di *hd* nel DAG ridotto finale e rappresentano quindi l'insieme minimo di informazioni sufficienti per l'inferenza diagnostica.

La CPT completa è stata suddivisa in tre tabelle, ciascuna corrispondente a un valore fissato della variabile *thal*, così da rendere più agevole l'interpretazione dei pattern probabilistici associati alle diverse condizioni di perfusione miocardica.

Tabella 9: CPT della variabile target: $P(hd \mid exang, ca, thal = 3)$.

	$exang = 0$		$exang = 1$	
ca	$P(hd = no)$	$P(hd = yes)$	$P(hd = no)$	$P(hd = yes)$
0	0.909	0.091	0.717	0.283
1	0.696	0.304	0.341	0.659
2	0.631	0.369	0.061	0.939
3	0.274	0.726	0.086	0.914

Caso *thal* = 3 (perfusione normale) In presenza di perfusione normale, la probabilità di malattia cardiaca aumenta in modo monotono al crescere del numero di vasi coronarici coinvolti (ca). Anche in assenza di angina da sforzo, valori elevati di ca sono associati a un rischio consistente, mentre la presenza di angina amplifica ulteriormente tale probabilità, in particolare per livelli intermedi e severi di compromissione coronarica.

Tabella 10: CPT della variabile target: $P(hd \mid exang, ca, thal = 6)$.

	$exang = 0$		$exang = 1$	
ca	$P(hd = no)$	$P(hd = yes)$	$P(hd = no)$	$P(hd = yes)$
0	0.962	0.038	0.354	0.646
1	0.147	0.853	0.061	0.939
2	0.061	0.939	0.147	0.853
3	0.147	0.853	0.147	0.853

Caso *thal* = 6 (difetto fisso di perfusione) In presenza di difetti reversibili di perfusione, la probabilità di malattia risulta elevata già per valori moderati di ca. La combinazione di difetto di perfusione e angina da sforzo determina probabilità di hd prossime o superiori al 90%, evidenziando il forte potere discriminante della variabile $thal$ in questa configurazione.

Tabella 11: CPT della variabile target: $P(hd \mid exang, ca, thal = 7)$.

	$exang = 0$		$exang = 1$	
ca	$P(hd = no)$	$P(hd = yes)$	$P(hd = no)$	$P(hd = yes)$
0	0.619	0.381	0.172	0.828
1	0.153	0.847	0.120	0.880
2	0.028	0.972	0.016	0.984
3	0.262	0.738	0.047	0.953

Caso *thal* = 7 (difetto reversibile di perfusione) Nel caso di difetto fisso di perfusione, indicativo di danno miocardico consolidato, la probabilità di malattia cardiaca risulta estremamente elevata per quasi tutte le configurazioni considerate. In particolare, per valori di ca pari a 2 o superiori, la probabilità di hd supera stabilmente il 95%, indipendentemente dalla presenza di angina da sforzo.

Nel complesso, le CPT della variabile target mostrano pattern probabilistici pienamente coerenti con il dominio clinico: la probabilità di malattia cardiaca cresce all'aumentare della severità anatomica (*ca*) e funzionale (*thal*) e viene ulteriormente accentuata dalla presenza di angina indotta da esercizio (*exang*). Tali risultati confermano che il DAG ridotto finale e la stima parametrica adottata sono in grado di catturare in modo efficace le principali determinanti diagnostiche dell'esito clinico.

9 Inferenza

Una volta fissata la struttura del DAG e stimate le CPT, la rete bayesiana può essere utilizzata per effettuare inferenza diagnostica, ossia per calcolare probabilità a posteriori del tipo $P(hd \mid E = e)$, dove $E = e$ rappresenta un insieme di evidenze cliniche osservate.

L'inferenza è stata eseguita con il pacchetto `gRain`. Il modello parametrizzato (`bn.fit`) è stato convertito in un oggetto `grain` tramite `as.grain()` e successivamente *compilato* con `compile()`, che costruisce una rappresentazione basata su *junction tree* e consente inferenza esatta mediante propagazione delle credenze.

9.1 Evidenza e interrogazione del modello: `setEvidence()` e `querygrain()`

Nel workflow di `gRain` è fondamentale distinguere tra:

- **Inserimento dell'evidenza (`setEvidence`)**: impone che una o più variabili assumano stati osservati (ad es. $exang = 1$), ottenendo un modello condizionato su tali osservazioni;
- **Interrogazione (`querygrain`)**: calcola le distribuzioni marginali o posteriori delle variabili di interesse (ad es. $P(hd)$ oppure $P(hd \mid E)$) sul modello corrente, con o senza evidenza.

In termini probabilistici, `setEvidence()` realizza il condizionamento su, mentre `querygrain()` restituisce le quantità richieste (marginali/posteriori) dopo la propagazione dell'evidenza nel junction tree.

Una volta fissata la struttura del DAG e stimate le CPT con approccio bayesiano, la rete bayesiana discreta risulta completamente specificata e può essere utilizzata per rispondere a query probabilistiche di tipo diagnostico. In questa sezione si esegue inferenza esatta sulla rete mediante propagazione delle credenze.

9.2 Query di baseline: distribuzione marginale della variabile target

Il primo passo inferenziale consiste nel calcolo della distribuzione marginale della variabile target *hd* in assenza di qualunque evidenza osservata:

$$P(hd).$$

Questa distribuzione rappresenta la probabilità *a priori* di malattia cardiaca indotta dal modello (e quindi dal campione e dalla regolarizzazione bayesiana), e costituisce il riferimento di confronto per le successive probabilità *a posteriori* ottenute introducendo evidenze diagnostiche.

In particolare, nella fase successiva verrà analizzato come l'introduzione di evidenza sulle variabili appartenenti alla Markov blanket di *hd*, ossia {*exang*, *ca*, *thal*}, modifichi la probabilità di $hd = yes$ rispetto al valore *a priori*.

Tabella 12: Distribuzione marginale *a priori* della variabile target nel modello ridotto.

Stato	**Probabilità**
$hd = no$	0.534
$hd = yes$	0.466

La Tabella 12 riporta la distribuzione marginale della variabile target in assenza di evidenza. Nel campione considerato la probabilità *a priori* di malattia cardiaca risulta pari a $P(hd = yes) \approx 0.466$, mentre $P(hd = no) \approx 0.534$. Questo valore rappresenta la baseline diagnostica del modello e verrà utilizzato come riferimento per valutare l'impatto delle evidenze inserite nelle query successive.

9.3 Query diagnostiche con evidenza parziale: effetto di *exang*

Per valutare l'impatto di una singola evidenza sulla diagnosi, si considera la distribuzione *a posteriori* della variabile target condizionata sulla presenza/assenza di angina indotta da sforzo:

$$P(hd \mid exang).$$

Poiché *exang* appartiene alla Markov blanket di *hd*, l'evidenza su questa variabile produce un aggiornamento informativo diretto della probabilità di malattia.

Tabella 13: Distribuzione a posteriori di *hd* condizionata su *exang*.

Evidenza	$\mathbf{P}(hd = no \mid \cdot)$	$\mathbf{P}(hd = yes \mid \cdot)$
$exang = 0$	0.631	0.369
$exang = 1$	0.340	0.660

Rispetto alla baseline a priori $P(hd = yes) \approx 0.466$, l'assenza di angina da sforzo ($exang = 0$) riduce la probabilità di malattia a $P(hd = yes \mid exang = 0) \approx 0.369$ (variazione $\Delta \approx -0.097$). Al contrario, la presenza di angina da sforzo ($exang = 1$) aumenta sensibilmente la probabilità di malattia fino a $P(hd = yes \mid exang = 1) \approx 0.660$ (variazione $\Delta \approx +0.194$).

Dal punto di vista clinico, il risultato è coerente con l'interpretazione di *exang* come indicatore funzionale diretto di ischemia indotta dallo sforzo: la sua osservazione modifica in modo marcato la probabilità diagnostica rispetto al rischio di base.

9.4 Query diagnostiche con evidenza completa: scenari su {*exang*, *ca*, *thal*}

La variabile target *hd* ha come genitori nel DAG ridotto finale l'insieme {*exang*, *ca*, *thal*}, che coincide con la sua Markov blanket. Ne segue che, fissata evidenza su tali varia-

bili, la distribuzione di *hd* risulta completamente determinata localmente dalla CPT della target e non dipende da ulteriori variabili del modello. In questa sezione vengono definiti alcuni scenari clinici rappresentativi, con severità crescente, e si calcola la probabilità *a posteriori* di malattia cardiaca:

$$P(hd = yes \mid exang, ca, thal).$$

Tabella 14: Posteriori diagnostiche per scenari clinici definiti sulle variabili della Markov blanket di *hd*.

Scenario	***exang***	***ca***	***thal***	$\mathbf{P}(hd = yes \mid \cdot)$
Baseline (prior)	–	–	–	0.466
S1: quadro rassicurante	0	0	3	0.091
S2: rischio intermedio	1	1	3	0.659
S3: rischio alto	1	2	7	0.984

I risultati mostrano un aggiornamento diagnostico marcato rispetto alla probabilità *a priori* (≈ 0.466). In particolare, nello scenario rassicurante (S1) la combinazione di assenza di angina da sforzo, assenza di vasi coinvolti e perfusione normale riduce la probabilità di malattia a circa 0.091, indicando un'evidenza fortemente negativa. Al contrario, nello scenario ad alta severità (S3) la presenza di angina da sforzo, coinvolgimento multivasale e difetto di perfusione determina una probabilità di malattia prossima a 1 (≈ 0.984), coerentemente con l'elevata specificità diagnostica delle variabili *ca* e *thal*.

9.5 Evidenza su variabili esterne alla Markov blanket: effetto di *age*

Dopo aver analizzato query con evidenza diretta sulla Markov blanket della target, si considera ora un caso di evidenza su una variabile *esterna* alla Markov blanket di *hd*. In particolare, si valuta l'effetto dell'età:

$$P(hd = yes \mid age).$$

Nel DAG ridotto, *age* non è nella Markov blanket di *hd*; pertanto l'evidenza su *age* non aggiorna direttamente la target, ma si propaga lungo i cammini attivi del grafo. In questo caso, l'influenza di *age* sulla diagnosi avviene principalmente attraverso la variabile di anatomia coronarica *ca* (catena *age*→*ca*→*hd*), coerentemente con il Blocco B e con l'interpretazione clinica dell'età come fattore di rischio a monte.

Tabella 15: Probabilità a posteriori di malattia cardiaca condizionata sull'età: $P(hd = yes \mid age)$.

Classe di età (*age*)	$\mathbf{P}(hd = yes \mid age)$
$[29, 51]$	0.378
$(51, 59]$	0.478
$(59, 77]$	0.553

I risultati evidenziano un incremento monotono della probabilità di malattia al crescere della classe di età: dai soggetti più giovani (≈ 0.378) ai soggetti di età intermedia (≈ 0.478) fino alla classe più anziana (≈ 0.553). Dal punto di vista clinico, questo andamento è plausibile poiché l'età aumenta la probabilità di aterosclerosi e coinvolgimento coronarico, fenomeno che nel modello è catturato dalla CPT $P(ca \mid age)$ (Blocco B) e che si riflette sulla target tramite l'arco $ca \to hd$.

10 Conclusioni

In questo progetto è stata sviluppata una rete bayesiana discreta per lo studio della malattia cardiaca (dataset Heart Disease – Cleveland), con l'obiettivo di ottenere un modello diagnostico probabilistico interpretabile e utilizzabile per inferenza. Il workflow ha seguito una pipeline coerente: definizione della variabile target (*hd*), preprocessing dei dati (gestione dei missing, conversione a fattori e discretizzazione delle variabili continue), apprendimento della struttura del DAG, validazione delle assunzioni di indipendenza condizionale, stima delle CPT e inferenza probabilistica.

L'apprendimento strutturale è stato condotto con Hill-Climbing sotto vincoli clinicamente motivati tramite blacklist: (i) *hd* come nodo terminale (assenza di archi uscenti), e (ii) organizzazione per *tiers* (demografia → clinica di base → sforzo → diagnostica avanzata). Tra i criteri di scoring considerati (AIC, BIC, BDeu), la struttura finale è stata selezionata tramite BDeu con *iss*=10, in quanto più stabile e interpretabile nel contesto di variabili discretizzate e campione moderato. Le indipendenze implicate dal DAG (d-separation) sono risultate complessivamente compatibili con i dati, confermando la plausibilità statistica della struttura selezionata.

Un risultato rilevante del modello è che la Markov blanket della variabile target coincide con l'insieme dei suoi genitori, $MB(hd) = \{exang, ca, thal\}$. Questo ha fornito una motivazione teorica e operativa per la riduzione del modello: alcune variabili laterali (ad esempio *trestbps*, *fbs*, *chol*, *restecg* e *thalach*) non aggiungono informazione diagnostica su *hd* una volta condizionato sulla Markov blanket, come supportato sia dal grafo sia da test empirici. Il modello ridotto conserva così i cammini informativi principali, migliorando leggibilità e facilità di utilizzo nelle query diagnostiche.

La stima parametrica è stata effettuata con approccio bayesiano (coerente con lo score BDeu) e ha prodotto CPT clinicamente interpretabili: l'età influenza l'anatomia coronarica ($age \to ca$); la catena sintomatologia/sforzo ($cp \to exang, oldpeak \to slope$) descrive la progressione verso alterazioni funzionali; l'esito *thal* dipende congiuntamente da *sex* e *slope*. Soprattutto, la CPT della target evidenzia che la probabilità di malattia cardiaca cresce marcatamente all'aumentare della severità anatomica (*ca*) e funzionale (*thal*) ed è ulteriormente amplificata dalla presenza di angina indotta da sforzo (*exang*).

Infine, il modello è stato utilizzato per inferenza esatta tramite Junction Tree con il pacchetto `gRain`. Le query diagnostiche hanno mostrato aggiornamenti posteriori coerenti: l'evidenza su variabili nella Markov blanket modifica direttamente e in modo sostanziale $P(hd = yes)$, mentre evidenze esterne (ad esempio *age*) influenzano la diagnosi solo indirettamente, propagandosi lungo i cammini attivi del DAG (ad esempio $age \to ca \to hd$). Gli scenari clinici costruiti su $\{exang, ca, thal\}$ hanno evidenziato

una netta stratificazione del rischio, dal quadro rassicurante con probabilità bassa di malattia fino a configurazioni ad alta severità con probabilità prossima a 1.

Nel complesso, la rete bayesiana sviluppata costituisce un modello probabilistico interpretabile, coerente con vincoli di dominio e utilizzabile per analisi inferenziali.