PREVISIONE DI UN INFARTO CON TECNICHE DI DATA MINING

23 Novembre 2022

Abstract

Le malattie cardiovascolari sono la prima causa di morte a livello globale. Ad oggi, nel mondo, 18 milioni di persone all'anno muoiono a causa di una di queste malattie, e si prevede che entro il 2030 questo numero salirà a 24 milioni, giungendo ad una media di oltre 66 mila morti al giorno [1].
Nello specifico, il nostro lavoro si concentra sull'infarto, una delle più comuni malattie cardiovascolari, ed in particolare sulla previsione del rischio d'infarto mediante strumenti di data mining. L'obiettivo è quello di individuare il miglior metodo di classificazione che permetta di stimare la probabilità che un individuo ha di avere un infarto utilizzando tecniche di classificazione come il KNN e la regressione logistica. Dai risultati delle analisi svolte si conclude che il miglior metodo risulta essere la regressione logistica; tuttavia, la ricerca deve proseguire per ovviare ad alcuni problemi incontrati durante le analisi. I risultati di questo lavoro possono essere utili sia nel campo diagnostico, sia dal punto di vista economico sanitario.

1 Introduzione

Secondo le stime dell'OMS, il 32% delle morti annuali è riconducibile ad una malattia cardiovascolare ed in particolare l'85% di queste è causata dall' infarto.
Analizziamo quindi nel dettaglio la malattia in questione e i sintomi più frequenti.

Si parla di infarto del miocardio, noto nel linguaggio comune semplicemente come infarto, in presenza di una necrosi di parte del tessuto cardiaco dovuta ad un'ostruzione di una delle arterie coronariche, il cui compito è quello di rifornire il cuore di sangue ossigenato. L'ostruzione, che può essere parziale o totale, è spesso dovuta a un accumulo di grasso, colesterolo o altre sostanze che formano una placca nelle arterie (aterosclerosi) che si rompe e determina una trombosi interrompendo il flusso di sangue e causando così la morte (necrosi) del tessuto.

La sintomatologia legata all'infarto del miocardio è variabile: non tutti i pazienti riferiscono gli stessi sintomi o li avvertono alla stessa intensità; in alcuni casi, l'infarto può essere asintomatico e in altri casi ancora il primo segnale di infarto è un arresto cardiaco improvviso. La manifestazione più tipica dell'infarto è una sensazione di peso o dolore al petto che dura più di dieci minuti. Il dolore può estendersi dal petto a uno o entrambe le braccia e può irradiarsi anche al collo, alla mascella e alla schiena. Inoltre, il dolore al petto può essere associato a nausea, bruciore di stomaco o dolore addominale, fiato corto, stanchezza, sudorazione fredda, stordimento o vertigini. Nella maggior parte

dei casi la comparsa dei sintomi dell'infarto è improvvisa, ma possono anche esservi segnali di avviso nel corso delle ore, dei giorni o delle settimane precedenti; ne sono un esempio un dolore al petto ricorrente o una sensazione di pressione (detta angina pectoris) che è scatenata dal movimento e che si risolve a riposo. L'angina è dovuta a una diminuzione temporanea di flusso sanguigno al cuore, una condizione che però non è così prolungata da condurre alla necrosi del tessuto.

Alcuni fattori, distinti in modificabili e non modificabili, possono esporre a un maggior rischio di aterosclerosi e di infarto. Sono fattori non modificabili l'aumentare dell'età, il sesso (in età giovanile e matura il rischio è maggiore negli uomini, ma dopo la menopausa femminile il rischio è lo stesso nei due sessi) e la familiarità (casi di infarto in famiglia espongono a un maggior rischio, soprattutto se occorsi dai 55 anni negli uomini e dai 65 nelle donne). Sono invece fattori di rischio modificabili il vizio del fumo, l'ipertensione arteriosa (che danneggia le arterie), alti livelli di colesterolo LDL (il cosiddetto colesterolo cattivo che restringe le arterie) o di trigliceridi, il diabete (l'eccesso di glucosio nel sangue danneggia le arterie e favorisce l'aterosclerosi), l'obesità (che è associata ad alti livelli di colesterolo e di trigliceridi, ipertensione e diabete), la sindrome metabolica (un quadro che include obesità, diabete e ipertensione), la sedentarietà (la mancanza di attività fisica contribuisce a innalzare i livelli di colesterolo ed espone al rischio di aumento corporeo), lo stress e l'uso di sostanze stupefacenti.

L'esecuzione di un elettrocardiogramma, un esame che registra l'attività elettrica del cuore, permette di confermare o escludere l'infarto poiché il muscolo cardiaco danneggiato presenta un'alterazione nella conduzione degli impulsi elettrici. [2]

Avendo appurato una conoscenza del dominio d'interesse è ora possibile applicare tool di data mining, che, in parole semplici, trasformano i dati grezzi in conoscenza pratica e oggigiorno, in ambito medico, stanno diventando un tema di rilevante importanza per costruire prognosi grazie ai dati disponibili.

La comunità scientifica, in tutto il mondo, ha testato svariati metodi di classificazione per prevedere un attacco cardiaco; oltre i classici strumenti di data mining quali il KNN e la regressione logistica, sono stati utilizzati tool di machine learning più complessi come le reti neurali, gli alberi decisionali e i modelli random forest (RF). Quest'ultimi sembrano essere il metodo di classificazione migliore, raggiungendo un'accuracy di oltre il 99% in dataset equivalenti a quello da noi analizzato. [3].
In generale però, la letteratura scientifica non ha espresso un parere unanime sul miglior metodo in assoluto, in quanto questo dipende non solo dalle variabili utilizzate, ma anche dalla grandezza del dataset, che porta ad utilizzare una tecnica piuttosto che un'altra, anche per monitorare la situazione da un punto di vista computazionale.

L'obiettivo di questo progetto è quello di individuare il miglior metodo di classificazione per prevedere se un paziente è a rischio infarto o meno, basandoci su fattori modificabili (colesterolo, pressione sanguigna, …) e non modificabili (età, sesso) dell'individuo in esame. Inizialmente sono stati considerati quattro diversi approcci: LDA, QDA, KNN e regressione logistica. Tuttavia, in seguito ad alcune analisi preliminari, i modelli LDA e QDA sono stati subito scartati; pertanto, ci

siamo concentrati su una comparazione tra KNN e regressione logistica. Dai risultati delle analisi svolte si conclude che il miglior metodo di classificazione risulta essere la regressione logistica.

L'utilità di questo lavoro è duplice. Da un lato, la diagnosi precoce di un infarto, grazie al machine learning, può aiutare i medici ad essere consapevoli della situazione del paziente, facendo risparmiare a entrambi tempo e, in alcuni casi, salvare la vita del paziente. Inoltre, da un punto di vista economico, salvare un numero maggiore di persone significa avere meno pressione sugli ospedali, con un conseguente risparmio economico nel settore sanitario.

Il restante lavoro è organizzato come segue:

Sezione 2 - materiali: descrizione del dataset

Sezione 3 - metodi: fase di pre-processing, analisi esplorativa e applicazione dei metodi

Sezione 4 - risultati: comparazione dei risultati in termini di accuracy, sensibility, sensitivity, AUC e ROC curve per i metodi KNN e regressione logistica

Sezione 5 - discussioni: riassunto del problema, analisi, considerazioni sugli approcci utilizzati e possibili sviluppi futuri del lavoro.

Sezione 6 - bibliografia

2 Materiali

In questo lavoro è stato utilizzato il dataset "Heart Failure Prediction" [4], creato combinando le informazioni di cinque dataset relativi ad aree geografiche distinte e considerando 12 variabili comuni.

I cinque dataset utilizzati per la sua creazione sono:
- Cleveland: 303 osservazioni
- Ungheria: 294 osservazioni
- Svizzera: 123 osservazioni
- Long Beach VA: 200 osservazioni
- Set di dati Stalog (Cuore): 270 osservazioni

Il totale delle osservazioni è pari a 918, ottenute dopo l'eliminazione di 272 duplicati.
Al fine di prevedere un possibile infarto cardiaco viene considerato l'intero dataset, il quale contiene variabili utili al fine previsivo, delle quali riportiamo tutte le informazioni nella *Tabella 1*. La variabile "HeartDisease" nella nostra analisi viene identificata come variabile target.

Tabella 1: Variabili utilizzate nell'analisi

VARIABILE	DESCRIZIONE	MODALITÀ
HeartDisease	classe dell'output	**1**: infarto miocardico **0**: normale
Età	età del paziente	anni
Genere	sesso del paziente	**M**: maschio, **F**: femmina
ChestPainType	tipo di dolore toracico	**TA**: Angina tipica **ATA**: Angina atipica **NAP**: Dolore non anginoso **ASY**: Asintomatico
RestingBP	pressione sanguigna a riposo	mm Hg
Colesterolo	colesterolo sierico	mm/dl
FastingBS	glicemia a digiuno	**1:** se FastingBS > 120 mg/dl **0**: altrimenti
RestingECG	risultati dell'elettrocardiogramma a riposo	**Normal**: Normale, **ST**: presenta anomalie dell'onda ST-T (inversioni dell'onda T e/o innalzamento o depressione del tratto ST > 0,05 mV) (*Figura 1*) **LVH**: mostra una probabile o definita ipertrofia ventricolare sinistra secondo i criteri di Estes (*Figura 2*)
MaxHR	frequenza cardiaca massima raggiunta	valore numerico compreso tra 60 e 202
ExerciseAngina	dolore transitorio al torace o sensazione di pressione (che si manifesta quando il muscolo cardiaco non riceve una sufficiente quantità di ossigeno) indotto dall'esercizio fisico	**Y:** Sì **N:** No
Oldpeak	sottoslivellamento ST: alterazione dell'elettrocardiogramma di superficie.	Valore numerico misurato in depressione
ST_Slope	pendenza del segmento ST di picco da sforzo	**Up**: in salita **Flat**: in piano **Down**: in discesa

Figura 1: Elettrocardiogramma

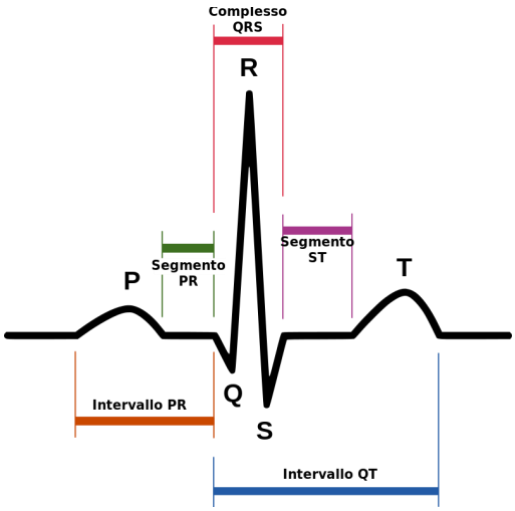
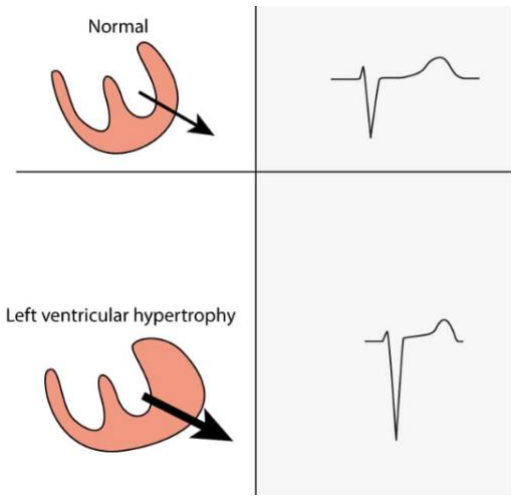


Figura 2: Ipertrofia ventricolare sinistra



Da un'analisi iniziale del dataset non si sono riscontrati valori mancanti. Tuttavia, da un'analisi più approfondita mediante le statistiche descrittive si è notata la presenza di valori atipici. Tale problema riguarda la variabile "Cholesterol", che presenta il 18.7% di valori pari a 0, incompatibili con il significato e i valori reali della variabile. Per ovviare a tale inconveniente abbiamo imputato agli zeri il valore della mediana, come verrà spiegato in modo più approfondito nella Sezione 3.
Un'altra variabile trattata è "RestingBP" che presentava un solo valore pari a 0, anch'esso incompatibile, e quindi in questo caso abbiamo eliminato l'osservazione.
L'analisi delle statistiche descrittive ci permette di delineare brevemente il campione (*Tab 2 e 3*). La percentuale di maschi è molto elevata (79%) (*Figura 3*), e l'età media è di 53.51 anni, con un massimo di 77 anni. Il colesterolo medio è pari a 243.2, quindi al di sopra della soglia critica indicata dal Ministero della Salute (*2004*) [5], il quale lo classifica come "Alto". Solo il 23% dei pazienti presenta una glicemia a digiuno superiore ai 120 mg/dl, mentre poco più della metà non ha avuto dolore al torace (*Figura 4*).
Infine, analizzando la variabile target "HeartDisease", si nota che il campione è ben bilanciato (*Figura 5*).

Tabella 2: Statistiche descrittive delle variabili quantitative

VARIABILE	MINIMO	MEDIANA	MEDIA	MASSIMO
Age	28.0	54.0	53.51	77.0
RestingBP	80.0	130.0	132.50	200.0
Cholesterol	85.0	237.0	243.20	603.0
MaxHR	60.0	138.0	136.80	202.0
Oldpeak	-2.6	0.6	0.8870	6.2

Tabella 3: Statistiche descrittive delle variabili qualitative

VARIABILE	MODALITÀ	PERCENTUALE
Sex (*Figura 3*)	M	79%
	F	21%
ChestPainType (*Figura 4*)	ASY	54%
	ATA	18.8%
	NAP	22.1%
	TA	5%
FastingBS	0	76.7%
	1	23.3%
RestingECG	LVH	20.5%
	Normal	60.1%
	ST	19.4%
ExerciseAngina	N	59.6%
	Y	40.4%
ST_Slope	Down	6.9%
	Flat	50.1%
	Up	43%
HeartDisease (*Figura 5*)	0	44.7%
	1	55.3%

Figura 3: Variabile "Sex"

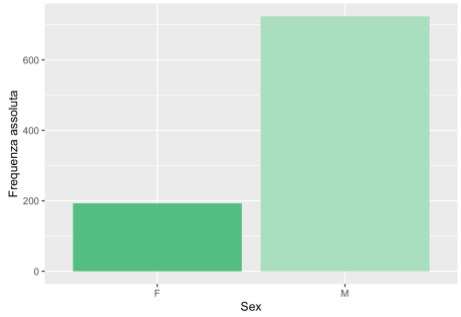


Figura 4: Variabile "ChestPainType"

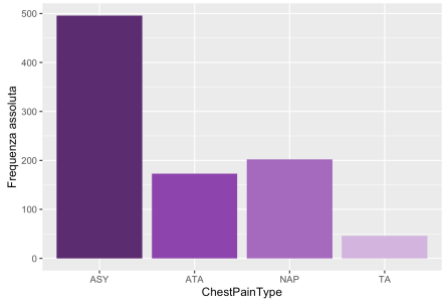
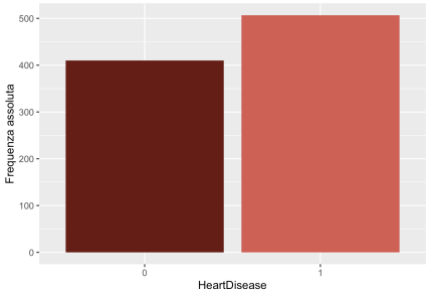


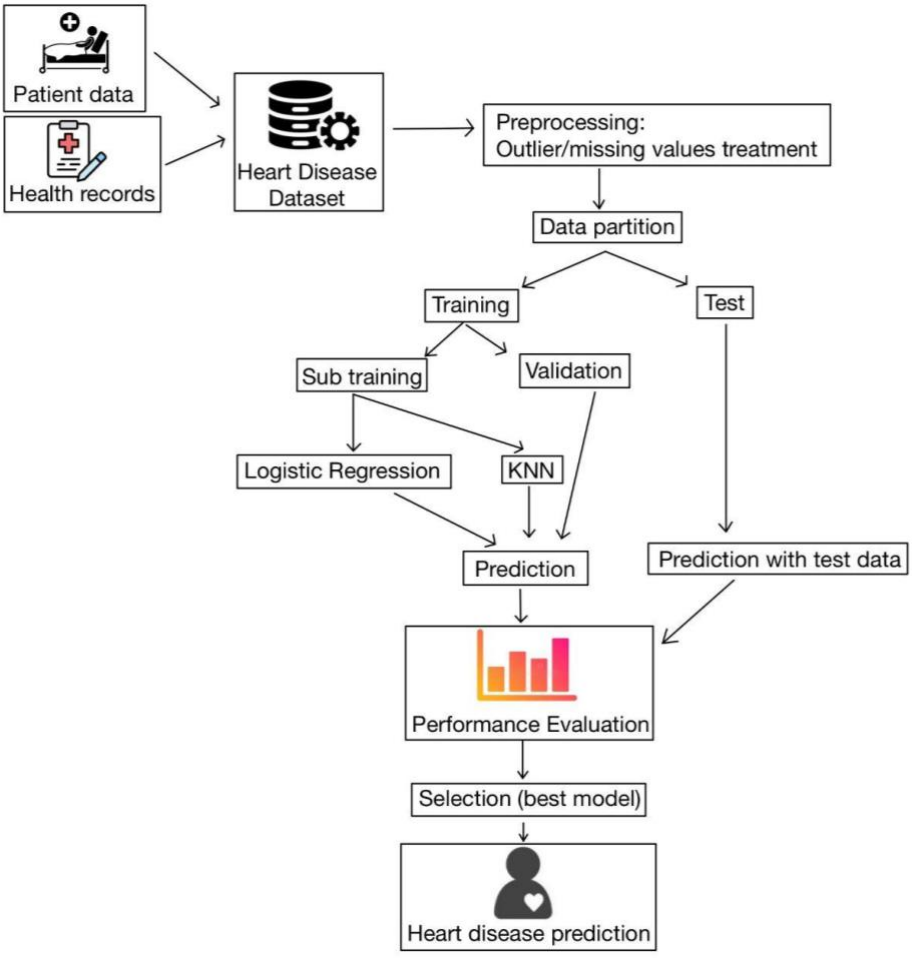
Figura 5: Variabile "HeartDisease"



3 Metodi

Riportiamo in questa sezione la descrizione dettagliata di tutte le operazioni che sono state eseguite sul dataset, i cui passaggi fondamentali sono riassunti nella *Figura 6*.

Figura 6. Flowchart



3.1 Analisi esplorativa

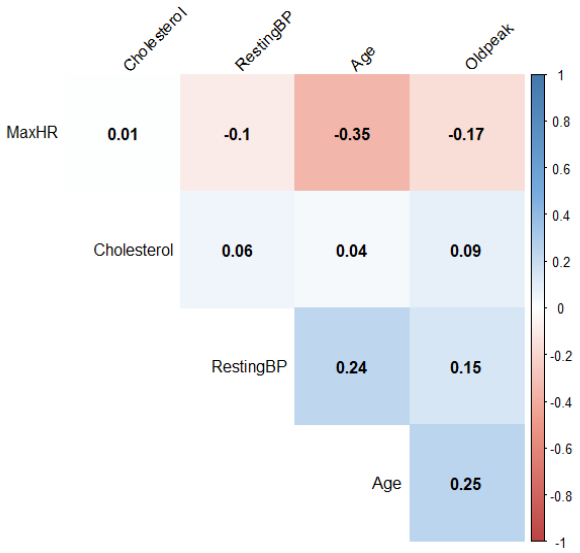
Il primo step dell'analisi è stato quello di dividere il dataset in *training* e *test set*. In particolare, abbiamo attribuito, in modo casuale, l'80% delle osservazioni del dataset al *training set* e il restante 20% al *test set*. È stata poi effettuata un'ulteriore divisione del *training set* in *sub training set* (65%) e *validation set* (35%). Questa fase è essenziale nel momento in cui si vanno ad applicare approcci di tipo supervisionato, il cui scopo è quello di fare generalizzazione e prevedere nuove istanze per la variabile target.

Sono state poi svolte le fasi di analisi esplorativa e di preparazione del dato, fondamentali nel processo di data mining, sul *training set*. In seguito sono state applicate le diverse operazioni di cleaning e sistemazione dei dati decise in questa fase anche a *validation set* e *test set.*

Per prima cosa sono stati trattati i valori mancanti, presenti solo per le variabili "Cholesterol" e "RestingBP". Per la variabile "Cholesterol" si è deciso di procedere con una strategia attiva: dal momento che i dati mancanti interessano un numero di unità considerevole (superiore alla soglia del 5% generalmente considerata), abbiamo sostituito ciascun missing con il valore mediano, calcolato a partire dai dati a disposizione. Nello specifico, abbiamo attribuito il valore mediano calcolato sul *sub training set* ai missing del *sub training set* stesso e del *validation set*, mentre il valore calcolato sul *training set* completo allo stesso e al *test set*. La scelta della mediana come valore plausibile con cui effettuare la sostituzione è dovuta al fatto che questa misura di posizione risulta essere meno influenzata dai valori anomali rispetto alla media. Abbiamo ritenuto inoltre che non fosse necessario imputare la mediana condizionata rispetto ad altre variabili, dato che nel dataset non sono presenti features altamente correlate con la variabile "Cholesterol". Per la variabile "RestingBP" abbiamo ritenuto invece più consona una strategia passiva: eliminare l'unica osservazione che presenta il dato mancante ed effettuare l'analisi esclusivamente sui dati a nostra disposizione.

Successivamente, tramite il calcolo della matrice di correlazione, abbiamo verificato eventuali effetti di multicollinearità tra le variabili quantitative. Dall'analisi della matrice *(Figura 7)* non si notano elevate correlazioni delle variabili e è stato quindi deciso di mantenerle tutte nell'analisi.

Figura 7: Matrice di correlazione tra le variabili quantitative oggetto di analisi

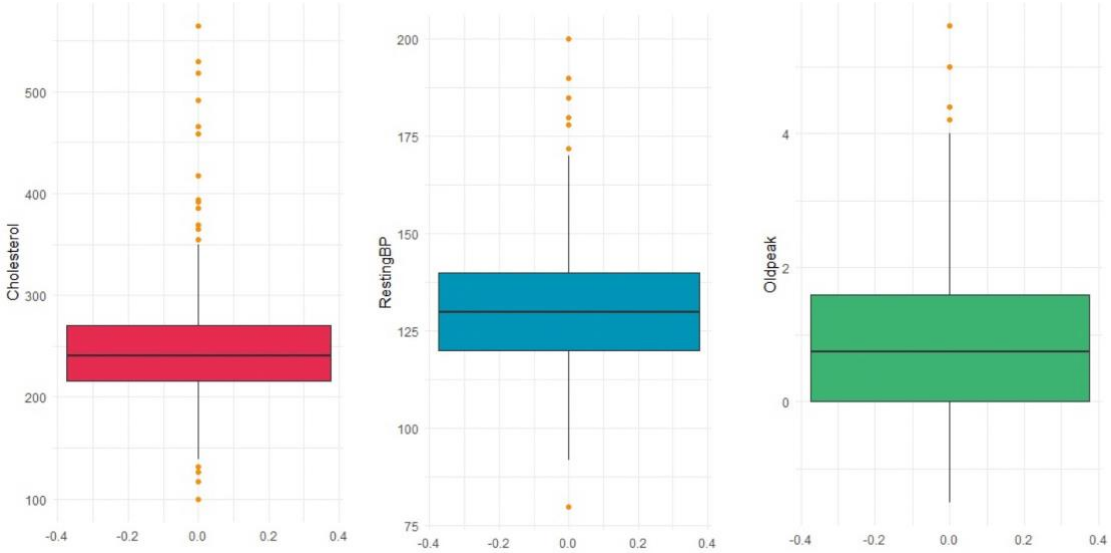


Attraverso i boxplot delle variabili sono stati analizzati la presenza di valori anomali e il range di variazione delle variabili, al fine di verificare la necessità di una eventuale normalizzazione o standardizzazione dei dati.

Dai boxplot è evidente che le variabili "Cholesterol", "RestingBP" e "Oldpeak" presentano valori anomali *(Figura 8)*; per ridurne l'influenza si è deciso di applicare una trasformazione logaritmica alle prime due, e di non trattare "Oldpeak" allo stesso modo dato che assume valori negativi e la presenza di outliers è ridotta, pertanto trascurabile.

Infine, abbiamo omogeneizzato le variabili effettuando una standardizzazione con mediana e MAD, essendo misure rispettivamente di posizione e di variabilità meno influenzate dalla presenza di outliers rispetto a media e deviazione standard.

Figura 8: Boxplot delle variabili "Cholesterol", "RestingBP" e "Oldpeak".



3.2 Verifica delle assunzioni

Nel secondo step sono state verificate le assunzioni per l'applicabilità dell'analisi discriminante, partendo dalla normalità delle variabili condizionate rispetto alla classe. Dai qqplot delle variabili *(Figure 9 e 10)*, che evidenziano code delle distribuzioni piuttosto pesanti, sembra che l'ipotesi non sia rispettata e per averne una conferma abbiamo utilizzato il test di normalità di Shapiro-Wilk. Il test è stato eseguito su un numero ristretto di variabili dal momento che l'analisi discriminante richiede che siano quantitative e continue.

I p-value del test portano all'accettazione dell'ipotesi nulla di normalità solo per due variabili (*Tabella 4*), "Age" e "MaxHR", e condizionatamente ad una sola classe. Abbiamo ritenuto quindi che considerare soltanto due variabili sulle dieci complessive del dataset fosse troppo riduttivo.

Considerato quanto detto, abbiamo deciso di non procedere con la classificazione tramite LDA e QDA, le quali avrebbero portato sicuramente a valori di accuracy non elevati, ma di utilizzare approcci alternativi, quali regressione logistica e KNN.

Figura 9: qqplot delle variabili condizionatamente alla classe HeartDisease.

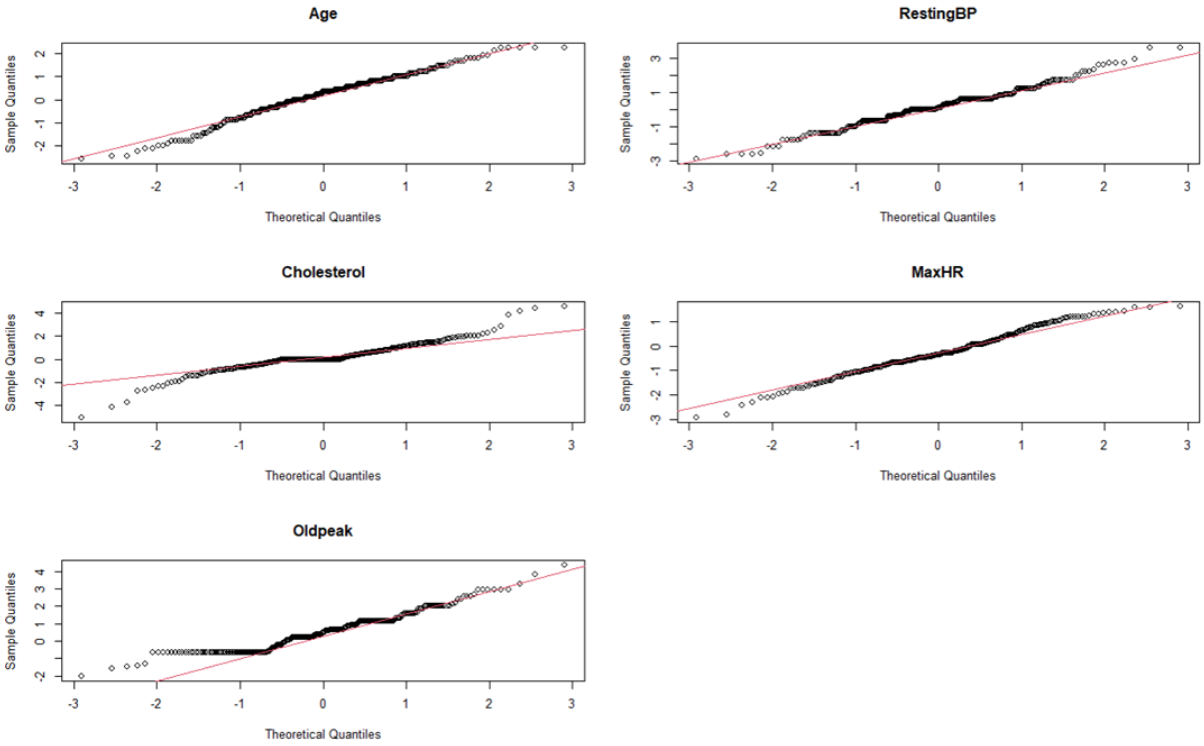


Figura 10: qqplot delle variabili condizionatamente alla classe no Heart Disease.

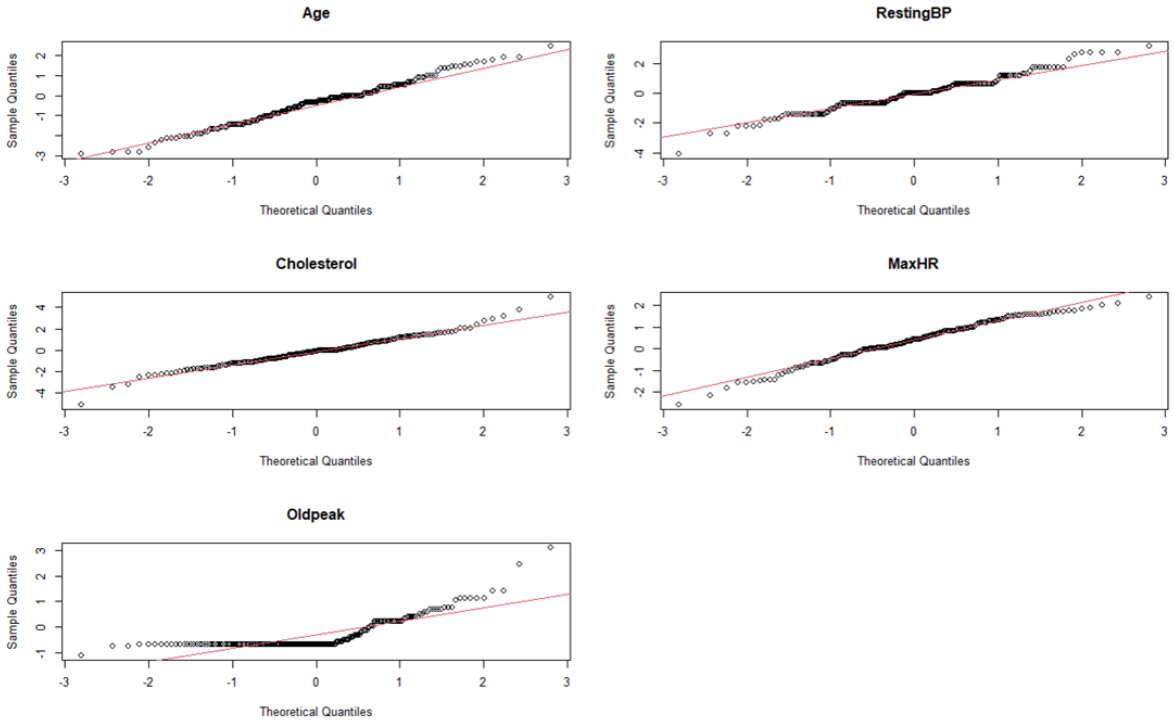


Tabella 4: P-values del test di normalità di Shapiro-Wilk condizionatamente alle classi di "HeartDisease".

VARIABILE	Heart Disease	Normal
Age	0.00108	0.27898
RestingBP	0.01146	0.00217
Cholesterol	0.00000	0.00587
MaxHR	0.05342	0.02167
Oldpeak	0.00000	0.00000

3.3 Classificazione

Metodo 1: Regressione logistica

Data la presenza di una variabile target qualitativa dicotomica, come primo modello abbiamo testato la regressione logistica.

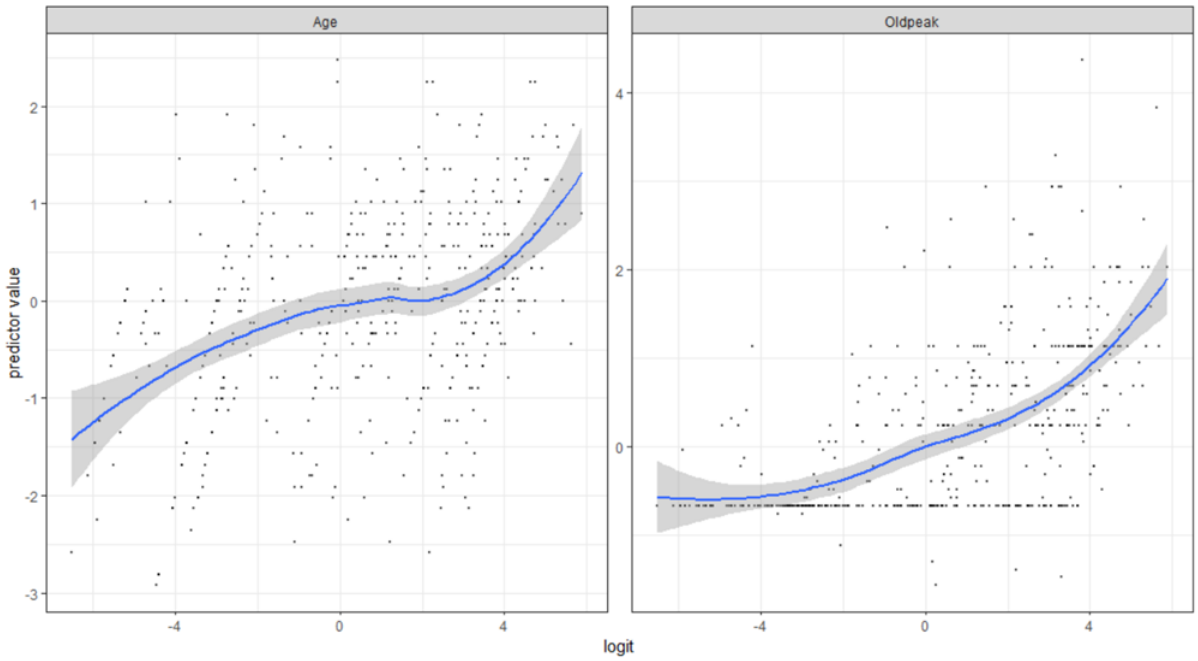
In primo luogo, abbiamo stimato il modello sul *sub training set* inserendo tutte le variabili e abbiamo applicato una selezione stepwise basata sul Akaike Information Criterion (AIC) al fine di identificare le variabili più informative a fini previsivi.

Si è proceduto eliminando i punti influenti che avrebbero potuto influenzare la stima del nostro modello e successivamente è stata ripetuta l'operazione di stima e di selezione delle variabili.

Il modello ottenuto in quest'ultima fase, contenente le variabili "Age", "Sex", "ChestPainType", "FastingBS", "OldPeak" e "ST_Slope risulta il modello migliore, sempre secondo il criterio di parsimoniosità basato su AIC.

Nel passo successivo si è svolta la fase di verifica della validità delle assunzioni preliminari: è necessario, infatti, verificare che vi sia una relazione lineare tra la logit (rapporto tra la probabilità di possedere l'attributo in esame e la probabilità di non possederlo) e le variabili esplicative. Anche se graficamente questa relazione sembra non essere perfettamente rispettata dalle due variabili "Age" e "Oldpeak" *(Figura 11)*, abbiamo deciso di procedere comunque con l'analisi, considerando anche il fatto che nel modello sono presenti altre quattro covariate informative (per le quali, essendo di tipo qualitativo e essendo trattate dal modello come dummy, non siamo in grado di testare la linearità). Questa scelta è stata effettuata nella consapevolezza del possibile errore ma anche nell'impossibilità di disporre, al momento, di metodi migliori.

Figura 11: Grafico contenente la logit in ascissa e i predittori in ordinata.



È stata poi valutata sul *validation set* la capacità previsiva del modello selezionato precedentemente. Sono stati ottenuti dei valori di accuracy e di AUC (Area Under the ROC Curve) sufficientemente elevati per procedere alla fase successiva di ristima del modello su *training set* e *validation set* e di previsione su *test set*.

Metodo 2. KNN (K-nearest neighbors)

Un approccio alternativo alla regressione logistica è l'algoritmo KNN. Basandosi sul calcolo delle distanze tra le osservazioni, è necessario escludere dal modello tutte le variabili qualitative sconnesse, ovvero "Sex", "FastingBS", "RestingECG" e "ExcerciseAngina".

Essendo il KNN un metodo di classificazione non parametrico, non è richiesta la stima di un modello nel sub *training set*, ma si procede valutando la capacità di classificazione sul *validation set*. Sono stati testati tutti i valori di k compresi tra 1 e 100 ed abbiamo ottenuto K = 55 come parametro di tuning ottimale, cioè il valore che massimizza l'accuracy del modello. Lo stesso K è stato poi utilizzato per prevedere la variabile target e valutare le performance dell'algoritmo anche nel *test set*.

4 Risultati

In questa sezione vengono discussi i principali risultati ottenuti dall'applicazione del KNN e della regressione logistica per stabilire quale dei due metodi ha migliori capacità previsive per la variabile target "Heart Disease".

Il modello di regressione logistica selezionato nel sub *training set* contiene le seguenti variabili: "Age", "Sex", "ChestPainType", "FastingBS", "Oldpeak" e "ST_Slope". Come già detto, le variabili esplicative quantitative considerate dal modello, ovvero Age e Oldpeak, non hanno una perfetta relazione lineare con la logit, ma riteniamo ragionevole proseguire con l'applicazione della regressione logistica, consapevoli dell'errore che si sta commettendo.

Lo studio prosegue con la valutazione della capacità di classificazione del modello proposto, che viene quindi applicato sul *validation set*. L'applicazione restituisce buoni risultati, con un'accuracy pari a 0.8249.

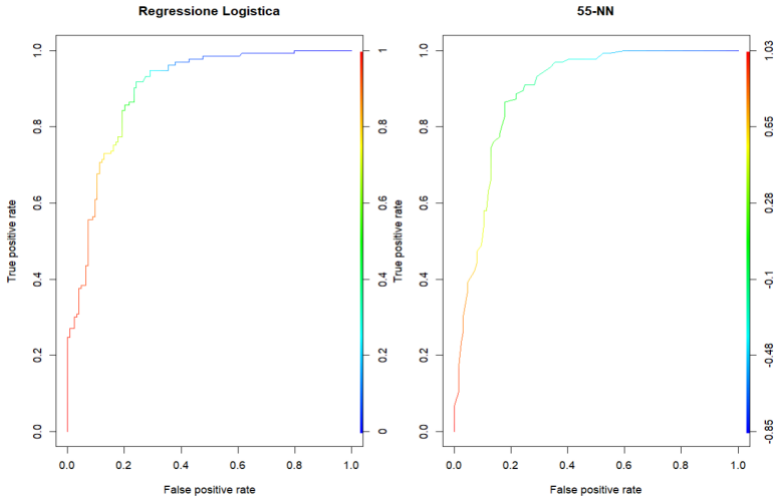
Parallelamente viene testato anche il modello KNN. Un primo studio ci permette di definire K=55 il valore che massimizza l'accuracy, che è pari a 0.8444.

Ad un primo confronto, si osserva che l'Area Under The Roc Curve (AUC) relativa alla regressione logistica applicata al *validation set* è pari a 0.8936, mentre per l'algoritmo KNN è pari a 0.8907.

In seguito, vengono visualizzate le curve Receiver Operating Characteristic (ROC). Queste vengono costruite valutando per diversi valori soglia della regola decisionale, la proporzione di veri positivi *(sensibilità)* e la proporzione di falsi positivi *(1 - specificità).* L'area sottostante alla curva ROC, ovvero l'AUC, è una misura di accuratezza. Tanto maggiore è l'area sotto la curva (cioè tanto più la curva si avvicina al vertice del grafico) tanto maggiore è il potere discriminante del test.

Guardando al grafico (*Figura 8*), è evidente che regressione logistica e KNN sono entrambi dei buoni modelli e sembrano essere ugualmente performanti.

Figura 12: Grafici rappresentanti le curve ROC calcolate sul validation set.



Il modello di regressione logistica finale, stimato sul *training set* completo, è il seguente:

$$logit(\pi(x)) = \ln\left(\frac{P(HeartDisease = 1|X)}{P(HeartDisease = 0|X)}\right) =$$
$$= -0.6889 + 0.2794 * Age + 1.7940 * SexM - 2.2391 * ChestPainTypeATA - 2.0674$$
$$* ChestPainTypeNAP - 1.3968 * ChestPainTypeTA + 1.1430 * FastingBS1 + 0.3681$$
$$* Oldpeak + 1.4307 * ST\_SlopeFlat - 1.1437 * ST\_SlopeUp$$

dove

$$SexM = \begin{cases} 1 \; se \; Sex = M \\ 0 \; se \; Sex = F \end{cases};$$

$$ChestPainTypeATA = \begin{cases} 1 \; se \; ChestPainType = ATA \\ 0 \; altrimenti \end{cases};$$

$$ChestPainTypeNAP = \begin{cases} 1 \; se \; ChestPainType = NAP \\ 0 \; altrimenti \end{cases};$$

$$ChestPainTypeTA = \begin{cases} 1 \; se \; ChestPainType = TA \\ 0 \; altrimenti \end{cases};$$

$$FastingBS1 = \begin{cases} 1 \; se \; FastingBS1 = 1 \\ 0 \; altrimenti \end{cases};$$

$$ST\_SlopeFlat = \begin{cases} 1 \; se \; ST\_Slope = 1 \\ 0 \; altrimenti \end{cases}$$

$$ST\_SlopeUP = \begin{cases} 1 \; se \; ST\_Slope = Up \\ 0 \; altrimenti \end{cases}$$

La significatività dei parametri del modello è riportata nella *Tabella 5*, dove *** indica un p-value inferiore a 0.001, ** compreso tra 0.001 e 0.01 e * compreso tra 0.01 e 0.05. Si nota che tutti i parametri sono significativi e quindi la totalità delle variabili selezionate sono utili al fine di prevedere nuove istanze della variabile target.

Tabella 5: significatività dei parametri

VARIABILE	P-value	Significatività
Intercetta	0.191378	
Age	0.019238	*
SexM	$1.33 * 10^{-9}$	***
ChestPainTypeATA	$5.22 * 10^{-10}$	***
ChestPainTypeNAP	$6.79 * 10^{-14}$	***
ChestPainTypeTA	0.005665	**
FastingBS1	0.000102	***
Oldpeak	0.001005	**
ST_SlopeFlat	0.002311	**
ST_SlopeUp	0.016753	*

La stima del modello ci permette di dare un'interpretazione alle stime dei parametri. Considerato un generico parametro beta, se la variabile a cui si riferisce è numerica, il parametro restituisce la variazione della logit per un incremento unitario della variabile considerata; mentre se la variabile è categorica, restituisce la variazione della logit per una categoria relativamente all'altra. In particolare, in ambito medico, quando il coefficiente della variabile è positivo, la variabile stessa corrisponderà ad un fattore di rischio. Al contrario, se negativo, corrisponderà ad un fattore di protezione.
Guardando ai segni delle nostre stime, possiamo concludere che i pazienti più anziani presentano una

probabilità maggiore di avere un infarto, a parità di tutte le altre variabili. Lo stesso ragionamento vale anche per la variabile "Oldpeak": i pazienti con maggiori variazioni dell'elettrocardiogramma avranno maggiore probabilità di avere un infarto, a parità di tutto il resto. Inoltre, notiamo che i pazienti di sesso maschile sono più a rischio rispetto a quelli di sesso femminile, così come i pazienti con un maggiore livello di glucosio nel sangue. Per la variabile "ChestPainType" la probabilità che un paziente soffra di infarto è maggiore se riporta un dolore anginoso tipico, atipico o un diverso tipo di dolore rispetto alla categoria di riferimento, che è asintomatico. Infine, un paziente che presenta una pendenza in piano del segmento ST è maggiormente a rischio rispetto alle altre categorie, mentre lo è meno se la pendenza è in salita.

Lo studio di entrambi i metodi di classificazione prosegue con l'applicazione sul *test set*.

La regressione logistica presenta un'accuracy pari a 0.8913 e AUC a 0.9286, mentre il KNN un'accuracy pari a 0.7935 e AUC a 0.8907. Per le considerazioni finali, si riassumono di seguito (*Tabelle 6 e 7*) i risultati principali ottenuti durante l'analisi.

Tabella 6: Risultati ottenuti sul validation set.

	Accuracy	**Sensitivity**	**Specificity**	**AUC**
Regressione	0.8249	0.8188	0.8319	0.8935545
KNN	0.8444	0.8394	0.8500	0.8907349

Tabella 7: Risultati ottenuti sul test set.

	Accuracy	**Sensitivity**	**Specificity**	**AUC**
Regressione	0.8913	0.8824	0.9024	0.9286308
KNN	0.7935	0.8000	0.7857	0.8907349

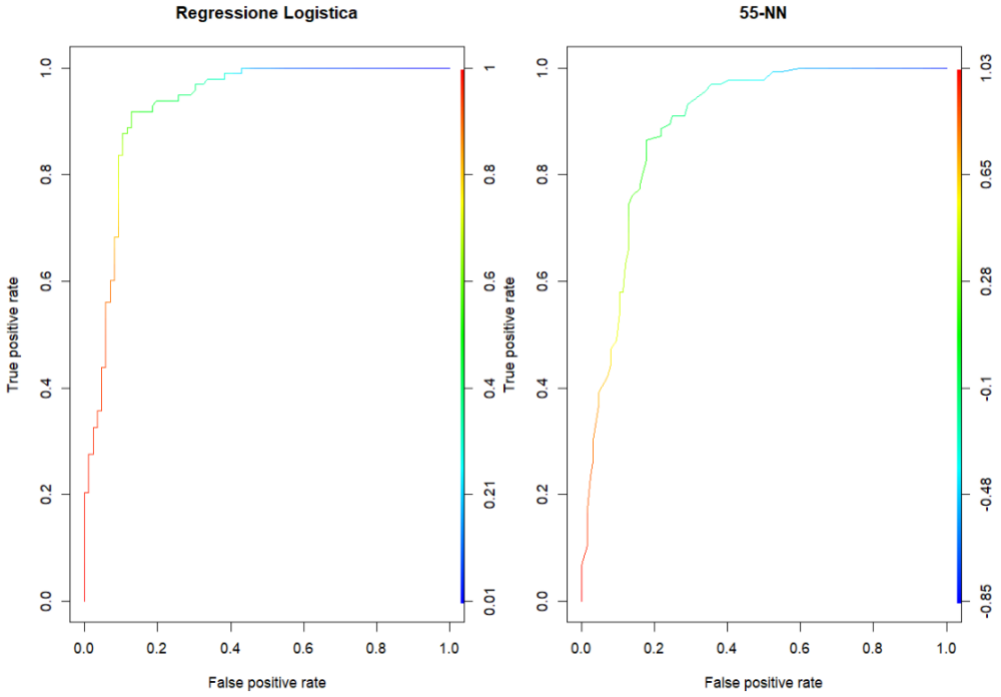
Dai risultati generali si può stabilire che entrambi i metodi hanno, in generale, delle prestazioni abbastanza soddisfacenti.

Le performance sul *validation set* sono migliori per il KNN, mentre il modello che restituisce esiti migliori sul *test set* è la regressione logistica.

Le curve ROC calcolate sul *test set* (*Figura 13*), confermano che l'AUC presenta valori maggiori per la regressione logistica, come già detto in precedenza.

Tenendo conto di tutti i risultati ottenuti e le considerazioni fatte in precedenza, il modello da noi scelto come migliore è la regressione logistica.

Figura 13: Grafici rappresentanti le curve ROC calcolate sul test set



5 Discussione

La ricerca presenta l'analisi comparativa per le prestazioni dei metodi della regressione logistica e del KNN per la classificazione dei dati.

Lo scopo dello studio era quello di creare un modello che ci permettesse l'assegnazione, ad ogni nuova osservazione presa in considerazione, della classe di appartenenza. Nello specifico, l'obiettivo era quello di prevedere, sulla base delle variabili fornite dal dataset, se il paziente fosse a rischio infarto.

Le prestazioni sono state confrontate in base a metriche di valutazione quali accuracy, sensitivity, specificity e AUC.

Visti i risultati ottenuti nel *test set*, siamo portati a preferire la regressione logistica. Questa presenta un'accuratezza migliore rispetto al KNN e dimostra di avere capacità migliori nell'assegnare correttamente la classe di appartenenza corretta ad una nuova osservazione. La nostra scelta è inoltre dovuta al fatto che per applicare l'algoritmo KNN siamo state costrette ad eliminare variabili informative per il problema in esame, di cui invece possiamo tenere conto nella regressione logistica. Come sottolineato durante lo svolgimento dello studio, le variabili non sembrano regredire perfettamente su un modello lineare, ma si ritiene comunque ragionevole l'approssimazione di questo tipo.

Inoltre, durante le operazioni di approfondimento e studio del fenomeno in analisi, abbiamo riscontrato un ampio utilizzo di questa tecnica in ambito di ricerca medica a fini predittivi.

In un eventuale sviluppo di questa analisi, una maggiore attenzione dovrebbe essere posta alla variabile "Cholesterol". Nel dataset preso in oggetto sono presenti molti valori pari a zero, ai quali è stata imputata la mediana non condizionata. Si potrebbe quindi integrare l'analisi con altri dataset, grazie alle cui variabili sarebbe possibile stimare con precisione "Cholesterol" e ottenere così risultati migliori.

Un'ulteriore considerazione sul lavoro è che il 79% delle osservazioni riguarda pazienti maschi. Dalla letteratura medica è noto che i sintomi dell'infarto possono manifestarsi in modo diverso in base al sesso [2]. Per questo, non è scontato che il modello abbia la stessa performance su un dataset bilanciato per quanto riguarda la variabile "Sex".

Futuri lavori si potrebbero orientare alla ricerca di un modello che meglio possa classificare le osservazioni. Problemi simili sono ampiamente studiati in ambito scientifico e, confrontando vari lavori svolti [3], il modello più performante sembra essere il Random Forest, che restituisce un'accuracy spesso superiore al 99%.

Infine, si riterrebbe utile, integrare il dataset con nuove variabili, ritenute predittive e significative per il caso medico analizzato. Il fatto che il paziente sia fumatore e la presenza di casi di infarto in famiglia, infatti, stando alla letteratura scientifica, sono variabili correlate con il verificarsi di un infarto.

6 Bibliografia

[1] Novartis Italia, 2021: Malattie cardiovascolari: entro il 2030 attesi 24 milioni di decessi nell'anno in tutto il mondo, Milano.

[2] Humanitas, 2020 : Cuore: che cos'è un infarto e come si interviene, Milano.

[3] National Library of Medicine, 2022 : Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques, USA.

[4] Kaggle, 2021: Heart Failure Prediction Dataset, Madrid.

[5] Ministero della salute, 2004: LINEE GUIDA PER LA PREVENZIONE DELL'ATEROSCLEROSI, Roma.