UNIVERSITÀ DEGLI STUDI DI BERGAMO

Dipartimento di
Ingegneria Gestionale,
dell'Informazione e della Produzione

Corso di laurea in
Ingegneria delle Tecnologie per la Salute
Classe n. L-9

Analisi del rischio di infarto miocardico basata su caratteristiche anamnestiche e sintomatologiche: correlazioni e regressione logistica

Candidato:
Giulio Ortoleva

Relatore:
Chiar.mo Prof. Ettore Lanzarone

Matricola n. 1070225

Anno Accademico
2024/2025

Indice

1 Introduzione

L'infarto miocardico rappresenta una delle principali cause di mortalità e morbilità a livello mondiale, con un impatto rilevante sia in termini clinici che socio-economici. La valutazione precoce e accurata del rischio nei pazienti che presentano sintomi sospetti costituisce quindi un aspetto fondamentale per una gestione efficace e tempestiva.

L'obiettivo di questa tesi è analizzare l'influenza delle caratteristiche sintomatologiche e anamnestiche sulla classificazione dei pazienti a elevato rischio di infarto miocardico. A tal fine, sono stati utilizzati metodi statistici classici per identificare associazioni significative tra le variabili e la classe di rischio finale, affiancati da tecniche di machine learning basate sulla regressione logistica.

Il confronto tra i due approcci consente non solo di verificare la coerenza dei risultati, ma anche di valutare l'efficacia dei modelli predittivi nel selezionare le feature più rilevanti.

2 Metodo

L'analisi è stata condotta in due fasi distinte: nella prima fase sono stati utilizzati test statistici per valutare la significatività di ciascuna feature, mentre nella seconda fase è stata applicata una tecnica di machine learning, in particolare la regressione logistica binaria, integrata con una procedura di selezione delle feature basata sul criterio di informazione di Akaike (AIC), il modello è stato impiegato per la predizione della classe di rischio rappresentata dalla feature "post-test".

Le variabili del dataset possono essere distinte in:
- **Quantitative**: età, peso (kg), indice di massa corporea (BMI), numero di fattori di rischio.
- **Qualitative**: sesso, sintomi (es. angina, dispnea, cardiopalmo, ecc.), fattori di rischio (es. diabete, ipertensione, fumo, ecc.), stime di rischio pre-test e post-test.

Per le variabili quantitative, è stata preliminarmente verificata la normalità della distribuzione mediante il test di Shapiro-Wilk. In caso di distribuzione normale, le differenze tra le due classi di rischio (alto vs basso) sono state valutate con il **t-test** per campioni indipendenti (Welch Two Sample t-test). Nel caso di distribuzione non normale, viene impiegato il test non parametrico di Wilcoxon, rivelato essere non necessario.

Per le variabili qualitative, sono stati applicati il test del Chi-quadro con correzione di Pearson. L'obiettivo era verificare l'indipendenza statistica tra le singole variabili e la classe di rischio post-test.

La soglia di significatività statistica è stata fissata a $\alpha = 0.05$. Le variabili con p-value inferiore a questa soglia sono state etichettate come significative.

Successivamente, è stato sviluppato un modello di regressione logistica binaria per stimare la probabilità di appartenenza alla classe di rischio alto. Sono state considerate due configurazioni:
- **Full model**: contenente tutte le variabili del dataset.
- **Back model**: ottenuto tramite **backward selection** basata su AIC, con l'obiettivo di selezionare il sottoinsieme di variabili più informativo e parsimonioso.

Infine, le performance dei modelli sono state valutate tramite l'area sotto la curva ROC (AUC), stimata con tecnica di **k-fold cross validation** (con $k = 9$).

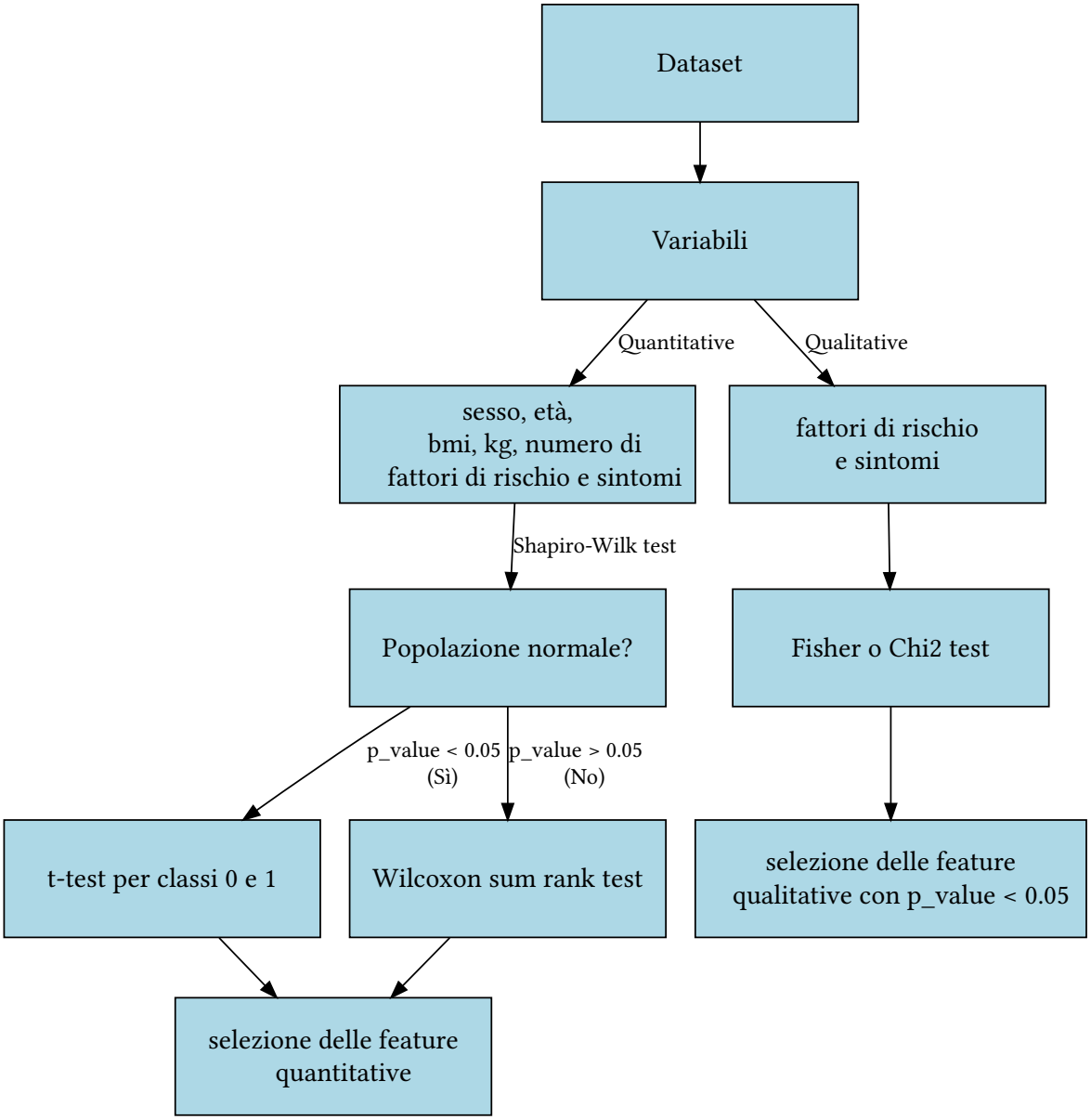

Figura 1: Schema dell'analisi (elaborazione dello studente)

3 Dataset

3.1 Informazioni generali

Il dataset utilizzato nello studio riguarda 2472 pazienti sottoposti a indagini scintigrafiche per sospetta ischemia e per accertamenti successivi. I dati sono stati acquisiti dalla struttura privata accreditata Gavazzeni. I dati sono inoltre stati anonimizzati.

3.2 Periodo

Il periodo di acquisizione va dal 3 gennaio 2019 al 3 gennaio 2024.

3.3 Altre informazioni

Il database è stato ricostruito a partire da dati grezzi in formato tabellare, i quali presentavano errori di digitazione e varianti non uniformi degli stessi valori categoriali.

3.4 Schema del dataset

Può essere suddiviso nelle seguenti tipologie di caratteristiche:

Sintomatologiche	**dtype**
angina	i64
astenia	i64
bruciore_retrosternale	i64
cardiopalmo	i64
costrizione_giugulare	i64
costrizione_mandibolare	i64
costrizione_retrosternale	i64
dispnea	i64
dolore_braccio_sinistro	i64
dolore_interscapolare	i64
epigastralgie	i64
lipotimia	i64
malessere	i64
no	i64
oppressione_epigastrica	i64
oppressione_retrosternale	i64
precordialgie	i64
scompenso_cardiaco	i64
sincope	i64
toracoalgie	i64
vertigini	i64

Demografiche	**dtype**
età	f64
kg	f64
sex	str: "Female" or "Male"

Tabella 1: Schema del dataset

Cliniche	**dtype**
post_test	str: "Basso" or "Alto"
pre_test	str: "Basso" or "Medio" or "Alto"

Fattori di rischio	**dtype**
diabete_insulino_dipendente	i64
diabete_non_insulino_dipendente	i64
dislipidemia	i64
familiarità	i64
fumo	i64
ipertensione	i64
obesità	i64
nessuno	i64

sintetiche	dtype
nfrc	i64: da 0 a 6
bmi	f64

*per le variabili a cui è assegnato dtype i64 i valori assunti sono 0 per l'assenza o 1 per la presenza del sintomo o del fattore di rischio.

3.5 Descrizione delle variabili

- **post_test**: livello di rischio di infarto miocardico del paziente, dopo gli accertamenti.
- **pre_test**: stima da parte dei medici del livello di rischio che il paziente abbia un infarto miocardico, effettuata prima che vengano eseguiti esami diagnostici.
- **nfrc**: numero di fattori di rischio

3.6 Descrizione statistica

Il dataset analizzato comprende un totale di 2472 pazienti, di cui 717 (29,0%) classificati come a rischio Alto nel post-test e 1755 (71,0%) come a rischio Basso.

Gruppo	*Maschi (%)*	*Femmine (%)*
Alto	74.2	25.8
Basso	58.2	41.8
complessivo	39.8	60.2

Tabella 2: Distribuzione sesso per gruppo

Gruppo	*Età media*	*Peso medio (kg)*	*BMI medio*
Alto	69.5	79.2	27.3
Basso	65.7	75.2	26.4

Tabella 3: Confronto tra gruppi (Alto vs Non-Alto)

Variabile	*Media*	*Dev. Std*	*Mediana*	*Min*	*Max*
Età	66.8	10.8	68.1	18.3	91.7
Peso (kg)	76.3	15.5	75.0	40.0	165.0
BMI	26.7	4.6	26.0	15.7	62.9

Tabella 4: Statistiche generali

3.7 grafici

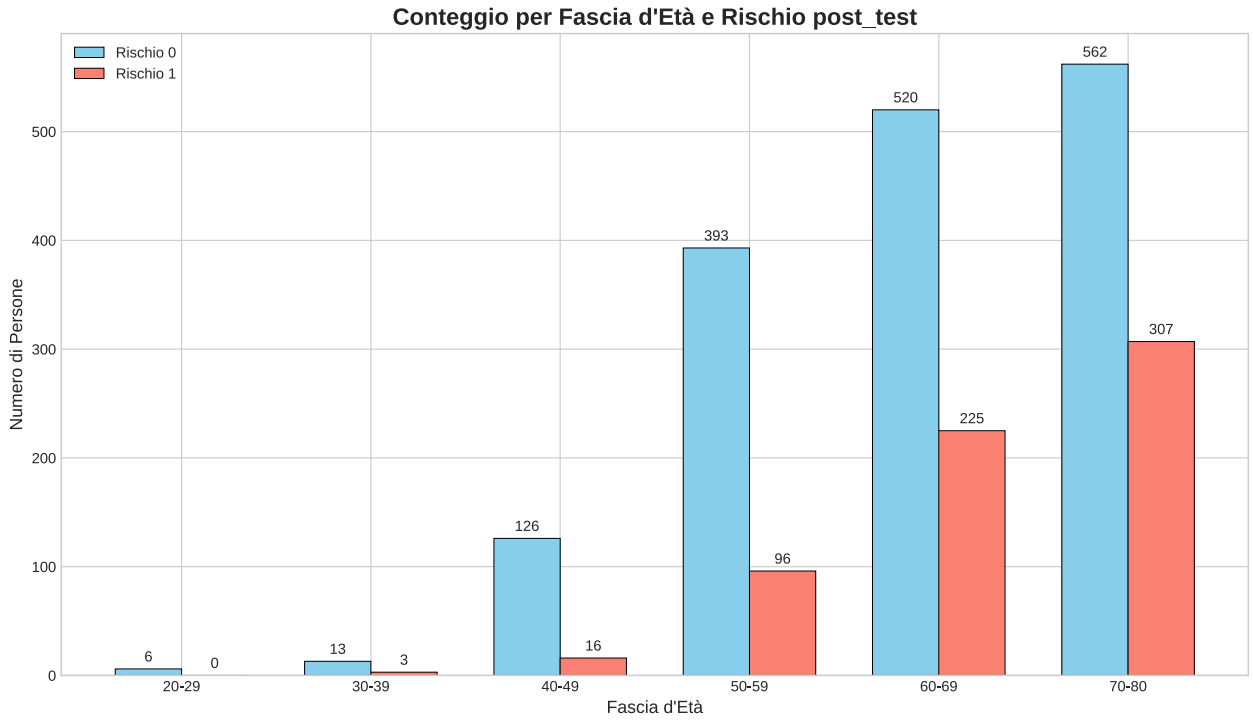

Figura 2: Distribuzione per fascia età, dove *Rischio 1* indica rischio Alto

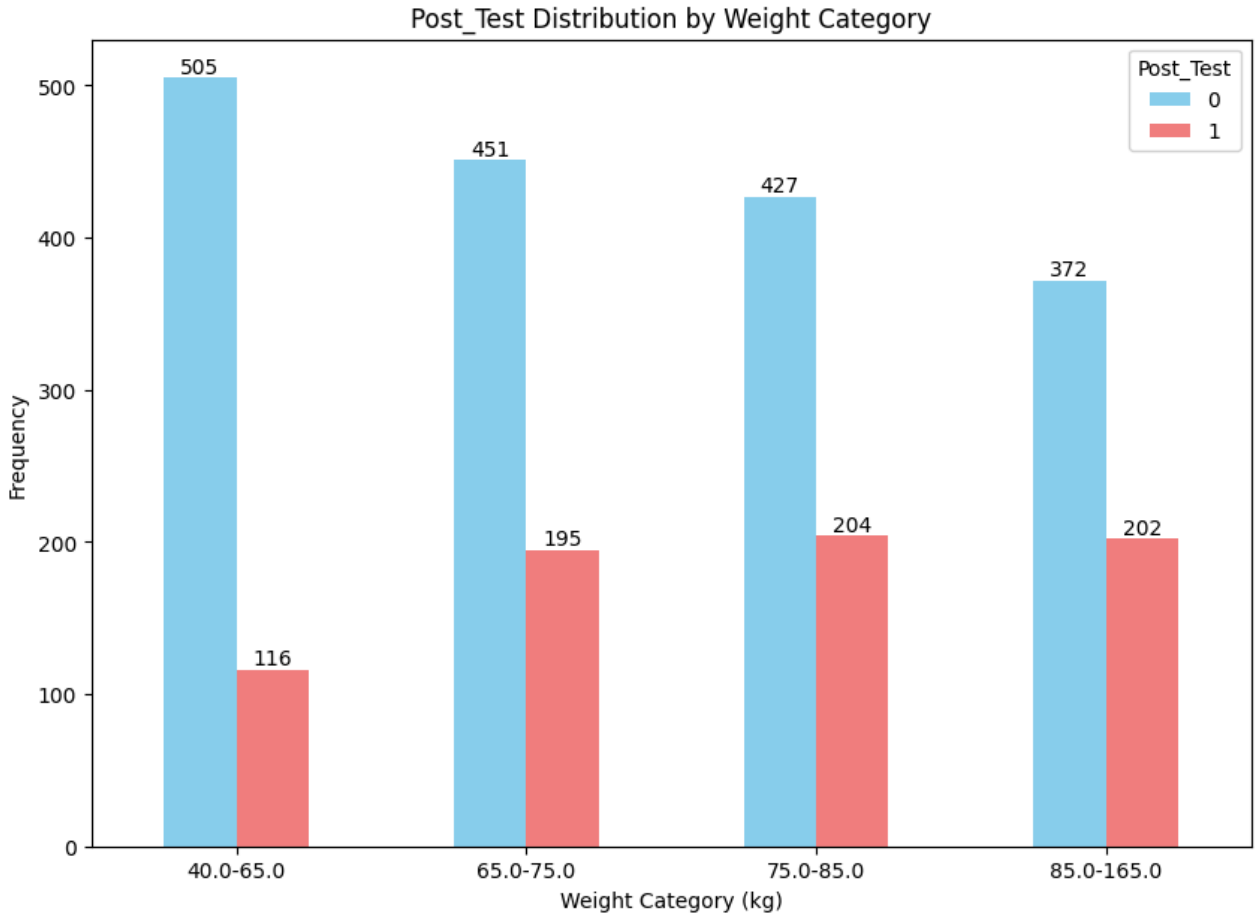

Figura 3: Distribuzione per categoria di peso, dove 1 indica rischio Alto

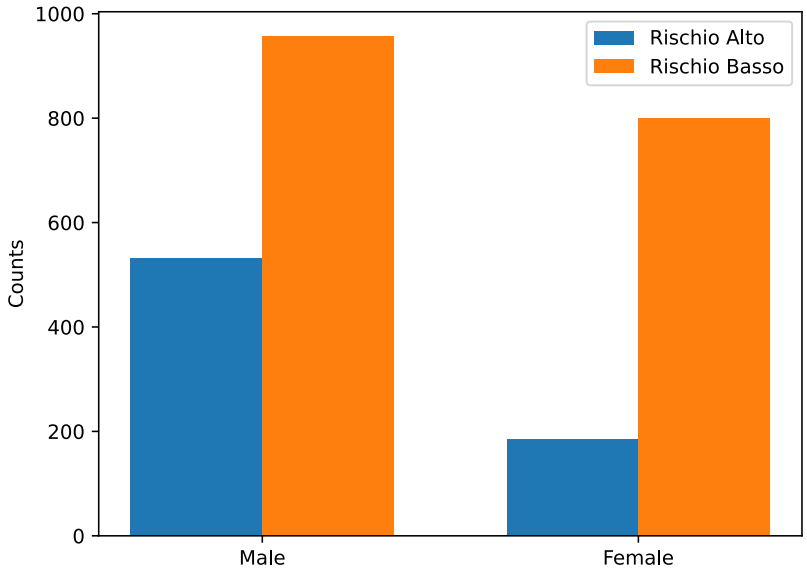

Figura 4: Frequenze per le categorie di rischio in base al sesso.

4 Risultati

4.1 Test di indipendenza: variabili categoriche

È stato applicato il test di indipendenza chi-quadro con correzione di Pearson alle variabili di natura categorica. L'analisi ha considerato i principali fattori di rischio, i sintomi, il sesso, il numero complessivo di fattori di rischio, nonché le condizioni di assenza di fattori di rischio e di assenza di sintomi.

4.1.1 Influenzano post-test

name	chi stat	p_value
pre_test	94.67035	0.00000
diabete nid	29.70884	0.00000
scompenso_cardiaco	25.88284	0.00000
ipertensione	23.01494	0.00001
nfrc	37.08662	0.00022
nessuno (fattori di rischio)	12.04392	0.00242
angina	10.69416	0.00476
cardiopalmo	9.57210	0.00835
astenia	8.45871	0.01456
dislipidemia	7.70195	0.02126
fumo	6.60912	0.03672
no (sintomo)	6.17208	0.04568

Tabella 5: Variabili con p-value < 0.05 in ordine crescente

4.1.2 Non influenzano post-test

name	chi stat	p_value
diabete_insulino_dipendente	5.27744	0.07145
familiarità	4.97367	0.08317
dolore_interscapolare	4.68025	0.09632
bruciore_retrosternale	3.94677	0.13899
costrizione_giugulare	3.91206	0.14142
sincope	2.67343	0.26271
costrizione_retrosternale	2.65234	0.26549
dolore_braccio_sinistro	2.59096	0.27377
oppressione_retrosternale	2.15820	0.33990
lipotimia	1.83229	0.40006
obesità	1.61882	0.44512
precordialgie	1.16822	0.55760
oppressione_epigastrica	1.07422	0.58443
malessere	0.59524	0.74258
costrizione_mandibolare	0.52872	0.76770
epigastralgie	0.39621	0.82028
toracoalgie	0.29029	0.86490
vertigini	0.29029	0.86490
dispnea	0.02708	0.98655

Tabella 6: Variabili con p-value > 0.05 in ordine crescente

4.2 Test d'indipendenza: variabili quantitative

4.2.1 Test di normalità: età, bmi, kg

Risultati del test di Shapiro-Wilk per la normalità della distribuzione delle categorie di post_test.

- Variabile: **età**

classe di rischio	lt005	distribuzione	p value
Basso	true	Normale	1.814e-13
Alto	true	Normale	6.01e-08

- Variabile: **Kg**

classe di rischio	lt005	distribuzione	p value
Basso	true	Normale	2.2e-16
Alto	true	Normale	1.464e-09

- Variabile: **bmi**

classe di rischio	lt005	distribuzione	p value
Basso	true	Normale	2.2e-16
Alto	true	Normale	6.018e-11

Le distribuzioni risultano essere normali. Viene eseguito il Welch Two Sample t-test:

Nome Variabile	Media: Alto	Media: Basso	p-value	Differenza Significativa
ETÀ	69.52	65.72	$< 2.2e-16$	Sì
KG	79.17	75.15	$1.419e-09$	Sì
BMI	27.30	26.43	$8.755e-06$	Sì

nota: Alto e Basso sono le varianti di post_test

4.3 Risultati complessivi

Nota: Il livello di soglia per il p-value è impostato a 0.05

name	p-value	Significatività	Full Model coeff.	Back Selec. coeff.
obesità	4.451e-1	No	NA (escluso da R)	
diabete_insulino_dipendente	7.145e-2	No	0.390	
familiarità	8.317e-2	No	0.303	
costrizione_giugulare	1.414e-1	No	1.333	1.309
costrizione_retrosternale	2.654e-1	No	0.648	0.608
dispnea	9.865e-1	No	0.076	
oppressione_epigastrica	5.844e-1	No	0.123	
oppressione_retrosternale	3.399e-1	No	0.440	0.425
precordialgie	5.576e-1	No	0.260	0.259
sincope	2.627e-1	No	0.755	0.783
vertigini	8.648e-1	No	0.073	
sex	2.2e-16	Sì	0.889	0.848
pre_test	1.339e-19	Sì	0.411	0.479
nfrc	2.162e-4	Sì	−0.272	
diabete_non_insulino_dipendente	3.538e-7	Sì	0.433	
dislipidemia	2.125e-2	Sì	0.318	
fumo	3.671e-2	Sì	0.264	
ipertensione	1.005e-5	Sì	0.411	
angina	4.762e-3	Sì	1.299	1.244
cardiopalmo	8.345e-3	Sì	−0.146	
età	2.2e-16	Sì	0.036	0.039
kg	1.419e-09	Sì	0.007	0.011
bmi	8.755e-06	Sì	0.016	

Tipo di test t-test chi quadro

- **FM coeff**: i coefficienti del modello completo
- **back coeff**: coefficienti del modello con le variabili selezionate tramite AIC backward selection
- **Significatività**: Per chi-quadro indica se c'è dipendenza con post_test, mentre per il t-test indica se le distribuzioni della feature per le categorie di post_test sono significativamente diverse.

4.4 Performance regressione logistica

Viene valutata la AUC con cross validation (9 k-folds) per il modello allenato con tutte le feature (Full Model), e il modello in seguito alla tecnica di selezione delle feature, AIC backward selection (Back Model).

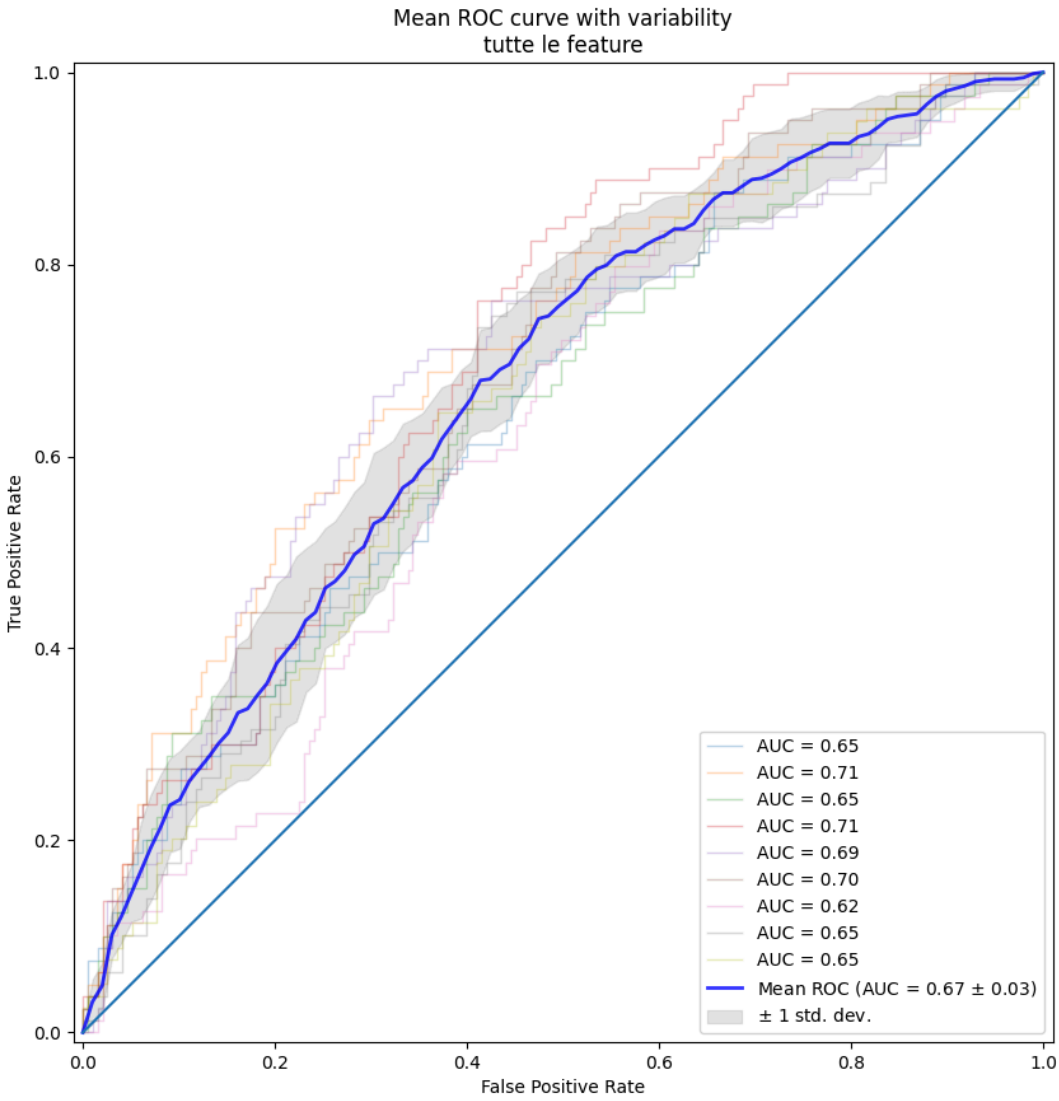

Figura 5: Modello con tutte le features

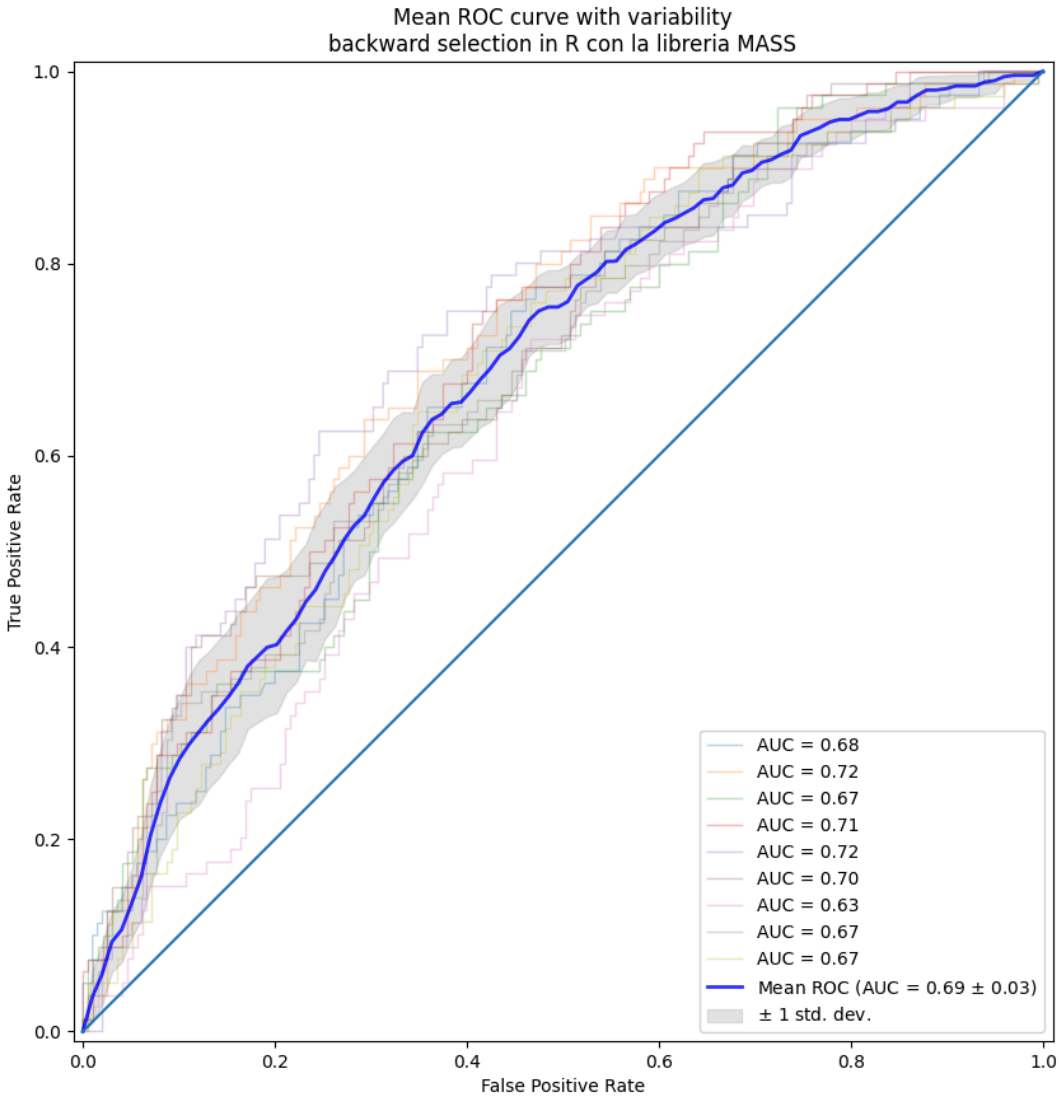

Figura 6: Modello con AIC backward selection

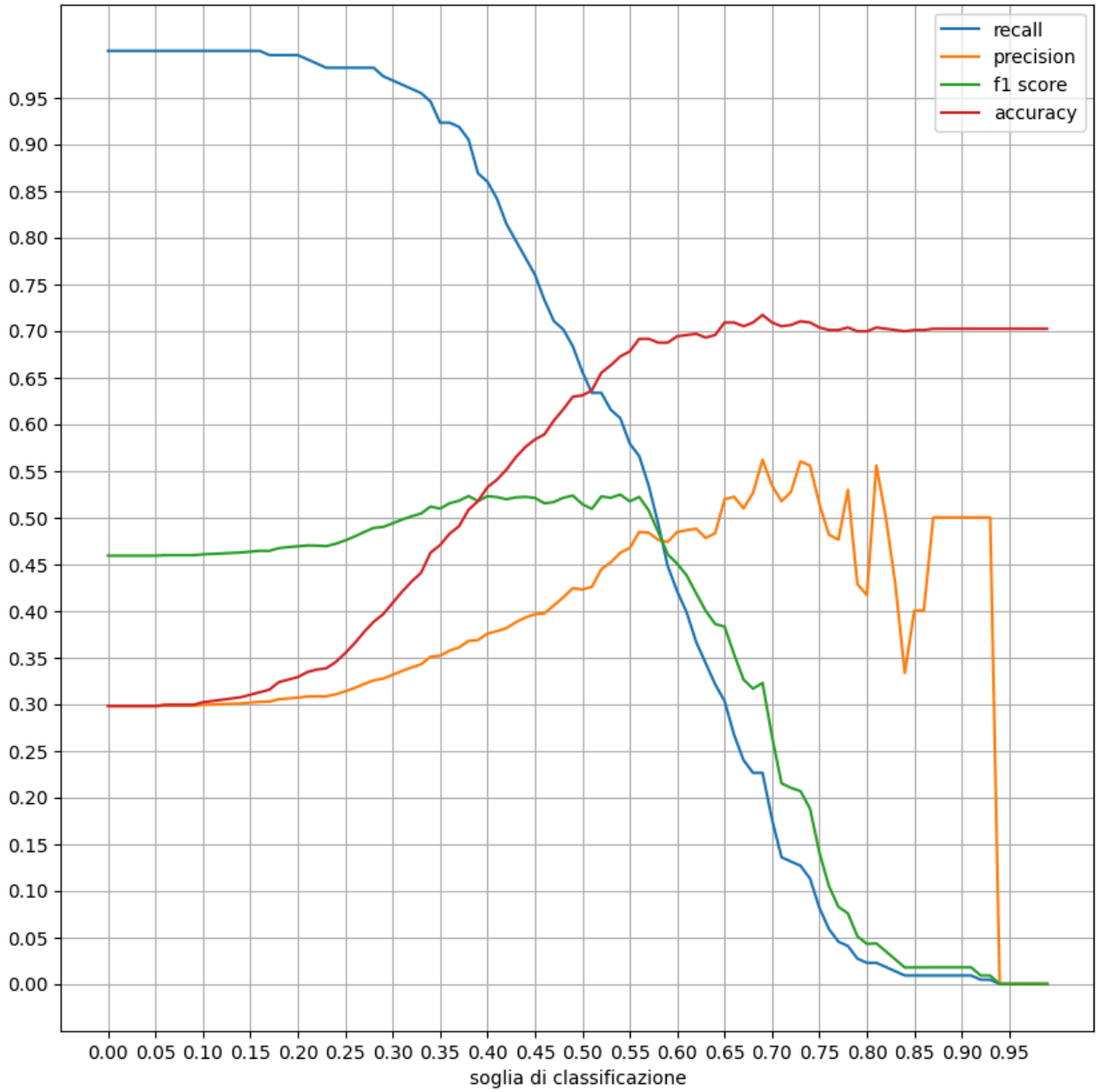

Figura 7: Metriche di performance del modello al variare della soglia di classificazione per modello con feature selezionate AIC

5 Discussioni e conclusioni

L'analisi condotta ha permesso di individuare un insieme di variabili anamnestiche e sintomatologiche significativamente associate al rischio di infarto miocardico. L'impiego combinato di test statistici classici e regressione logistica ha consentito di verificare la solidità dei risultati e di distinguere tra fattori con maggiore capacità predittiva e variabili a contributo marginale.

L'applicazione della regressione logistica, supportata dalla procedura di selezione delle feature basata sull'AIC, ha evidenziato un miglioramento delle prestazioni predittive: l'AUC media è passata da 0.67 (modello completo) a 0.69 (modello semplificato). Sebbene l'incremento sia contenuto, il risultato conferma l'utilità di strategie di selezione mirata per ridurre la complessità del modello e, al contempo, migliorarne la capacità discriminante.

I fattori che si sono confermati significativi in entrambe le analisi sono stati: età, peso, sesso, rischio stimato pre-test e presenza di angina. In particolare:

- **Età**: si associa positivamente al rischio, in accordo con le *linee guida ESC[1]* , che identificano l'età come determinante centrale nella valutazione del rischio cardiovascolare e utilizzano algoritmi dedicati (SCORE2 e SCORE2-OP) per stimarne l'impatto sulla probabilità di eventi, inclusi infarto miocardico e ictus.
- **Peso** (obesità): anch'esso associato a un incremento del rischio, in linea con quanto riportato nelle *linee guida ESC[1]* , che classificano l'adiposità tra i principali fattori di rischio modificabili per le malattie cardiovascolari aterosclerotiche.
- **Sesso**: il sesso maschile presenta una probabilità sensibilmente maggiore di essere classificato a rischio elevato, coerentemente con le evidenze ESC, che distinguono esplicitamente le stime di rischio per uomini e donne e riconoscono differenze sostanziali nella distribuzione e gestione del rischio cardiovascolare.
- **Rischio pre-test**: risulta fortemente predittivo, a conferma della validità clinica delle valutazioni iniziali effettuate dai medici.
- **Angina**: Seppur risultato significativo per post-test, bisogna considerare che il campione dei pazienti con angina è ridotto (24 casi) e questo potrebbe aver causato instabilità, ma bisogna anche tener conto che è un sintomo specifico per l'ischemia e quindi il rischio di infarto miocardico.

nome feature	min	medio	max	tipo
età	0.732	2.672	3.668	quantitativa
kg	0.44	0.839	1.815	quantitativa
angina	0	/	1.245	bool
pre_test	0	0.479	0.958	3 livelli
sex	0	/	0.848	bool

Tabella 7: Impatto delle feature sul rischio ottenuto moltiplicando i coefficienti per i valori minimi, medi e massimi rispettivamente.

Bibliografia

[1] F. L. J. Visseren *et al.*, «2021 ESC Guidelines on cardiovascular disease prevention in clinical practice: Developed by the Task Force for cardiovascular disease prevention in clinical practice with representatives of the European Society of Cardiology and 12 medical societies With the special contribution of the European Association of Preventive Cardiology (EAPC)», *European Heart Journal*, vol. 42, fasc. 34, pp. 3227–3337, 2021, doi: 10.1093/eurheartj/ehab484.