UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI ECONOMIA "GIORGIO FUÀ"

Corso di Laurea Magistrale in Data Science per l'Economia e le Imprese

Analisi del rischio ipertensivo tramite tecniche di Machine Learning

Hypertensive risk analysis using Machine Learning techniques

Relatore: Prof. Domenico Potena

Tesi di Laurea di: Andrea Baldinini

Anno Accademico 2023 – 2024

"*Senza dati sei solo un'altra persona con un'opinione"*

W. Edwards Deming

INDICE

1. INTRODUZIONE

L'ipertensione arteriosa, comunemente nota come pressione alta, rappresenta una delle condizioni croniche più diffuse e significative a livello globale. Definita come una pressione del sangue persistentemente elevata nelle arterie, l'ipertensione è un fattore di rischio primario per molte malattie cardiovascolari, che sono la principale causa di morte nel mondo. Secondo l'Organizzazione Mondiale della Sanità (OMS), circa 1,13 miliardi di persone soffrono di ipertensione, una condizione che contribuisce a circa 7,5 milioni di decessi ogni anno, pari al 13% del totale delle morti globali.

1.1 CONTESTO E MOTIVAZIONI

L'ipertensione è spesso definita come una "killer silenzioso" poiché molte persone non manifestano sintomi evidenti fino a quando non si verificano complicazioni gravi come infarti, ictus e insufficienza renale.

Le cause dell'ipertensione sono multifattoriali e includono sia fattori genetici che ambientali. Tra i principali fattori di rischio modificabili troviamo l'alimentazione ricca di sodio, l'eccessivo consumo di alcol, il fumo, la mancanza di attività fisica e il sovrappeso. Inoltre, condizioni comorbide come il diabete e la dislipidemia possono aggravare il rischio di sviluppare ipertensione.

Nonostante i progressi nella comprensione e nel trattamento dell'ipertensione, la condizione rimane un problema di salute pubblica significativo a causa della sua alta prevalenza e delle sue gravi complicazioni. Una diagnosi precoce e una gestione efficace sono cruciali per prevenire le conseguenze a lungo termine dell'ipertensione e per migliorare la qualità della vita dei pazienti.

In questo contesto, le tecniche di machine learning emergono come strumenti potenti e innovativi per affrontare la complessità della previsione del rischio ipertensivo. Grazie alla loro capacità di analizzare

grandi quantità di dati clinici e demografici, gli algoritmi di machine learning possono identificare pattern complessi e relazioni non lineari che potrebbero sfuggire agli approcci statistici tradizionali.

L'applicazione di queste tecniche presenta diversi vantaggi significativi:

- Accuratezza e precisione: gli algoritmi di machine learning possono migliorare la precisione delle previsioni rispetto ai modelli tradizionali, permettendo di individuare con maggiore affidabilità gli individui a rischio.

- Personalizzazione della diagnosi: analizzando un'ampia gamma di variabili, si può supportare una medicina personalizzata, adattando le strategie di intervento alle specifiche caratteristiche cliniche e demografiche del paziente.

- Identificazione di nuovi fattori di rischio: gli algoritmi possono rivelare nuovi fattori di rischio e interazioni tra variabili che non erano precedentemente riconosciuti, arricchendo la comprensione della patogenesi dell'ipertensione.

1.2 CASO DI STUDIO E OBIETTIVO

Il presente studio si basa sull'analisi di pazienti cardiologici dell'INRCA di Ancona, Istituto di Ricovero e Cura a Carattere Scientifico. Ogni paziente possiede un rischio ipertensivo valutato attraverso uno Score in percentuale e una classe di rischio associata (basso, medio, alto). Attualmente, per motivi di letteratura clinica, lo Score viene calcolato utilizzando una combinazione di variabili: età, sesso, abitudine al fumo, livello di colesterolo e pressione sistolica.

Le prime tre, essendo descrittive, non presentano difficoltà significative nella raccolta dei dati. Tuttavia, la misurazione dei livelli di colesterolo e della pressione sistolica comporta costi elevati sia in termini di tempo che di risorse finanziarie per l'ospedale. Inoltre, queste misurazioni possono essere soggette a errori, poiché la rilevazione viene effettuata una tantum, rendendo i dati potenzialmente non rappresentativi dello stato reale del paziente.

L'obiettivo dello studio è sviluppare un modello predittivo che possa stimare lo Score ipertensivo e la corrispondente classe di rischio utilizzando esclusivamente variabili descrittive, le quali devono essere scelte tra quelle presenti nel dataset. Questo approccio mira a fornire ai

medici un utile strumento di supporto decisionale, consentendo loro di ottenere una stima del rischio ipertensivo di un paziente basata unicamente su informazioni facilmente accessibili, senza la necessità di esami costosi e potenzialmente inaccurati.

Per raggiungere questo obiettivo, sono state applicate tecniche di machine learning che hanno permesso di analizzare i dati disponibili e di sviluppare modelli predittivi robusti.

2. FONDAMENTI TEORICI

Nel corso di questo studio, sono state esplorate diverse tecniche di machine learning per trovare il giusto equilibrio tra la qualità dei dati a disposizione, l'efficienza computazionale nel training del modello e l'efficacia delle performance predittive. Dopo numerosi esperimenti e valutazioni, le tecniche che si sono dimostrate più adatte alle esigenze del nostro studio sono state il Decision Tree Regressor e il Decision Tree Classifier. Questi due algoritmi offrono un buon compromesso tra interpretabilità, velocità di addestramento e accuratezza delle previsioni.

2.1 DECISION TREE REGRESSOR

Il Decision Tree Regressor è un algoritmo di machine learning che prevede valori continui come output. L'algoritmo costruisce un albero decisionale in cui ogni nodo interno rappresenta una decisione basata su una singola variabile del dataset, e ogni foglia rappresenta un valore di output predetto.

Il processo di costruzione inizia con l'intero dataset di training e procede dividendo i dati in sottoinsiemi sulla base di condizioni che massimizzano la riduzione dell'impurità. La scelta della variabile e della soglia per ciascuna divisione è determinata dall'algoritmo al fine di migliorare la predizione dei valori continui.

L'obiettivo principale durante la costruzione dell'albero è minimizzare la varianza all'interno dei sottoinsiemi dei dati. Minori sono le varianze nei sottoinsiemi, migliore sarà la predizione del modello.

Uno dei principali vantaggi del Decision Tree Regressor è la sua interpretabilità: è facile da comprendere e visualizzare, rendendo evidente come le decisioni vengano prese. Tuttavia, un potenziale limite è la tendenza all'*overfitting*, soprattutto se l'albero è troppo profondo e complesso.

2.2 DECISION TREE CLASSIFIER

Il Decision Tree Classifier è un algoritmo di machine learning utilizzato per problemi in cui l'output è una classe discreta. Simile al Decision Tree Regressor, il classifier costruisce un albero decisionale in cui ogni nodo rappresenta una decisione basata su una variabile del dataset, e ogni foglia rappresenta una classe predetta.

La costruzione di un Decision Tree Classifier segue un processo iterativo in cui l'obiettivo principale è ridurre l'impurità dei nodi ad ogni divisione. L'impurità è una misura che quantifica la mescolanza delle classi all'interno di un nodo. L'algoritmo valuta diverse possibili divisioni e seleziona quella che riduce maggiormente l'impurità, migliorando così la purezza delle classi nei sottoinsiemi risultanti.

Due metriche comuni utilizzate per determinare le divisioni ottimali sono l'impurità di Gini e *l'information gain*:

- Impurità di Gini: Questa misura valuta quanto i dati in un nodo siano mescolati in termini di classe. Un valore di Gini minore indica una maggiore omogeneità delle classi all'interno del nodo. La minimizzazione dell'impurità di Gini aiuta a creare nodi con classi più pure.

- *Information Gain*: Basata sull'entropia, questa misura rappresenta la riduzione dell'incertezza nelle classi dei dati dopo una divisione. La massimizzazione dell'information gain riduce significativamente l'impurità, rendendo le classi nei nodi figli più omogenee.

Il Decision Tree Classifier è noto per la sua semplicità e facilità d'uso. È facilmente interpretabile e può gestire sia variabili numeriche che categoriche. Tuttavia, come il regressor, è suscettibile all'*overfitting* se non vengono applicate tecniche di *pruning* per limitare la profondità dell'albero. Il *pruning* aiuta a evitare che l'albero diventi troppo complesso e specifico ai dati di addestramento, migliorando così la generalizzazione su dati nuovi.

3. PREPARAZIONE DEI DATI

Il capitolo pone l'attenzione sull'analisi della struttura e della qualità del dataset. Viene qui esposto come sono state dapprima studiate le singole variabili, per poi essere trattate mediante specifiche tecniche laddove necessario ed infine selezionate per poter, in seguito, costruire i modelli predittivi.

3.1 DATASET E PRE-PROCESSING

Il dataset contiene informazioni dettagliate su pazienti cardiologici, con l'obiettivo di prevedere il rischio ipertensivo. Le variabili target utilizzate sono SCORE2 per la regressione e RISCHIO per la classificazione. Di seguito vengono riportate le principali caratteristiche del dataset e una descrizione delle variabili incluse.

Il dataset è composto da 7945 osservazioni e 39 variabili che includono dati demografici, abitudini di vita, e parametri clinici dei pazienti. Di seguito, una breve descrizione delle variabili più rilevanti:

- Sesso: 0 per femmina, 1 per maschio.
- Età.
- Altezza: in centimetri.
- Peso: in chilogrammi.
- BMI: indice di massa corporea.
- BSA: superficie corporea.
- Fumo: 0 per no, 1 per sì.

- n° sig /die: numero di sigarette fumate al giorno.
- Anni fumo: numero di anni in cui il paziente ha fumato.
- Pack Years (n°sig/die X Anni fumo) / 20: un'unità di misura per valutare l'esposizione al fumo.
- OSAS: presenza/assenza di sindrome delle apnee ostruttive del sonno.
- Scompenso cardiaco: presenza/assenza.
- Fibrillazione atriale: presenza/assenza.
- Cardiopatia ischemica: presenza/assenza.
- Ictus/TIA: presenza/assenza di ictus o attacco ischemico transitorio.
- Ateromasia carotidea/AOP: presenza/assenza di ateromasia carotidea o arteriopatia obliterante periferica.
- Diabete mellito: presenza/assenza.
- Dislipidemia: presenza/assenza.
- BPCO: presenza/assenza di broncopneumopatia cronica ostruttiva.

- Colesterolo totale: valore del colesterolo totale nel sangue.
- HDL: valore delle lipoproteine ad alta densità nel sangue.
- Trigliceridi: valore dei trigliceridi nel sangue.
- cLDL: valore delle lipoproteine a bassa densità nel sangue.
- nonHDL: valore delle lipoproteine non-HDL nel sangue.
- Terapia ipolip: indica se il paziente sia o meno in terapia ipolipemizzante.
- Creatinina: valore della creatinina nel sangue.
- eGFR: tasso di filtrazione glomerulare stimato.
- PAS: pressione arteriosa sistolica.
- PAD: pressione arteriosa diastolica.
- FC: frequenza cardiaca.
- CATsexsmo: categoria di rischio basata su sesso e fumo.
- NonHDLMmMOL: livelli di non-HDL in millimoli per litro.
- CATcol: categoria di rischio basata sul colesterolo.

- CatAge: categoria di rischio basata sull'età.

- CATPA: categoria di rischio basata sulla pressione arteriosa.

- **SCORE2**: score numerico che rappresenta il rischio ipertensivo in percentuale; variabile target della regressione.

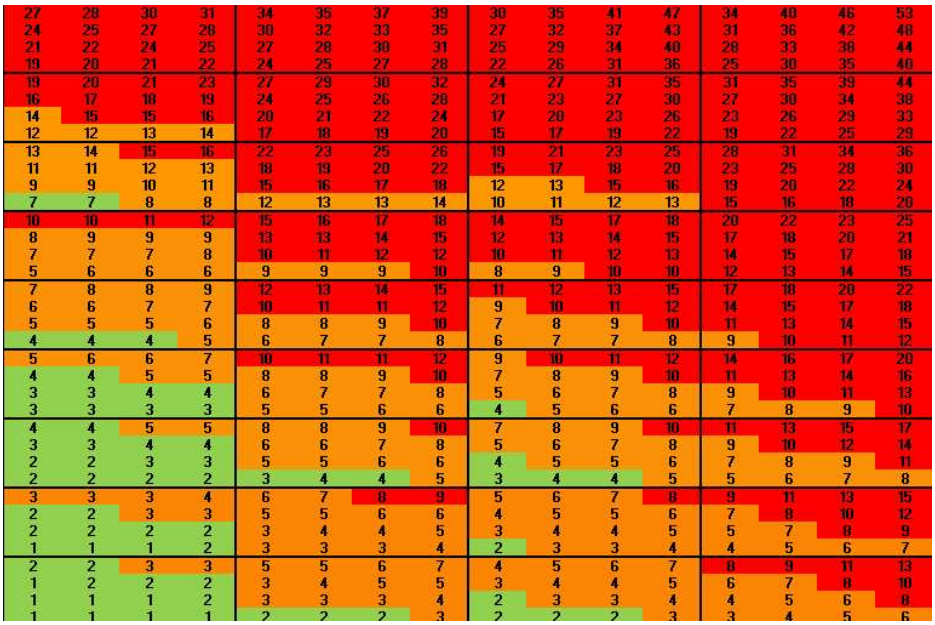

Fig. III.1 Tabella Score2

- **Rischio**: classe di rischio ipertensivo (0 per basso rischio, 1 per rischio medio, 2 per alto rischio); variabile target della classificazione.

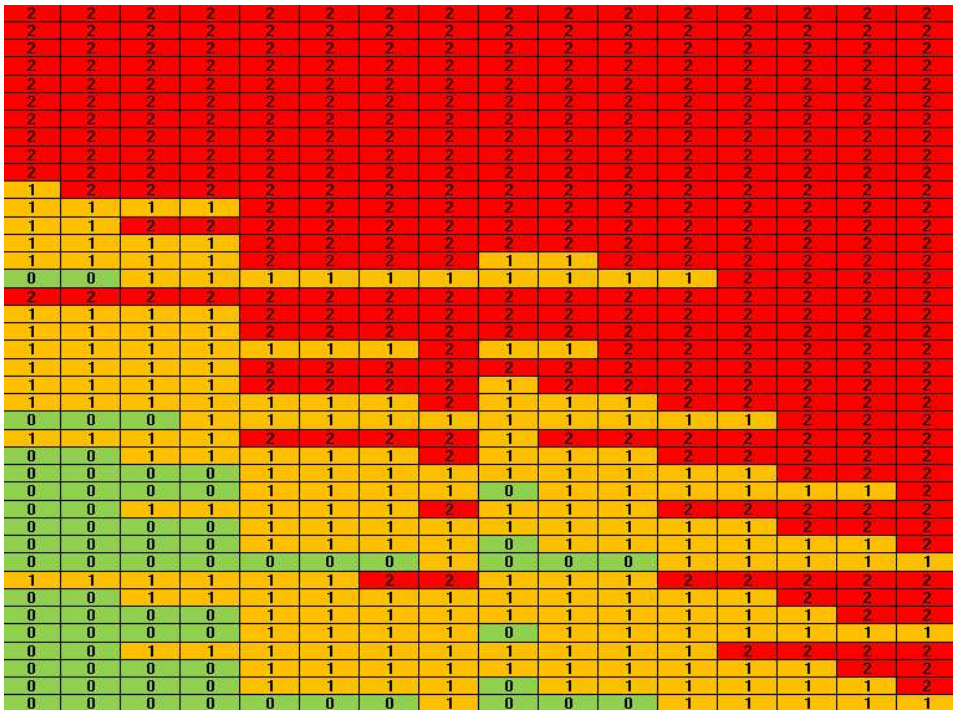

Fig. III.2 Tabella Rischio

Come si può osservare, sono numerose le variabili descrittive, la cui raccolta risulta semplice e priva di erronea determinazione. Al contempo, altre variabili sono il frutto di esami / osservazioni cliniche, passibili di *bias*.

3.1.1 PULIZIA E PREPARAZIONE DEL DATASET

In questo studio, è stato seguito un processo rigoroso per filtrare e preparare i dati, basato su criteri specifici derivati dalla letteratura medica e da considerazioni pratiche.

Il primo passo è stato applicare due filtri fondamentali. Sono stati inclusi solo pazienti con età superiore ai 40 anni. Questo criterio è motivato dalla letteratura medica che indica che il rischio ipertensivo e lo Score associato sono calcolabili in modo affidabile solo per questa fascia di età. Inoltre, sono stati considerati solo pazienti con livelli di trigliceridi inferiori a 400 mg/dL. Anche questo filtro è basato su evidenze mediche che suggeriscono limiti entro i quali il calcolo dello Score è valido.

Molti pazienti nel dataset originale mancavano di variabili essenziali per il calcolo dello Score, come il colesterolo totale. La mancanza di tali dati critici rende impossibile la stima accurata del rischio ipertensivo; pertanto, questi pazienti sono stati esclusi dall'analisi, lasciando solamente 3053 osservazioni fruibili.

Sono state rimosse tutte le variabili che presentavano più del 15% di valori mancanti. La scelta di questa soglia è stata motivata dalla

necessità di mantenere un bilancio tra la quantità di dati disponibili e la qualità delle informazioni. Variabili con troppi dati mancanti avrebbero potuto introdurre *bias* e compromettere la robustezza dei modelli.

Per garantire che l'analisi fosse basata su dati primari e per evitare potenziali circolarità nelle previsioni, sono state eliminate tutte le variabili calcolate, ossia quelle derivate da altre variabili del dataset (come il BMI e il BSA). Questo ha permesso di concentrare l'attenzione su misurazioni dirette e indipendenti.

Questi passaggi di pulizia e preparazione del dataset sono fondamentali per assicurare che i modelli di machine learning possano essere addestrati su dati di alta qualità, aumentando così l'affidabilità delle previsioni del rischio ipertensivo.

3.1.2 ANALISI DEI VALORI NULLI E TECNICHE DI GESTIONE

Nello studio sono state adottate diverse strategie per trattare i valori mancanti, creando tre differenti dataset. Di seguito, sono descritte in dettaglio le tecniche utilizzate e la *ratio* dietro ogni approccio.

Per gestire i valori nulli in modo più sofisticato, si è deciso di sfruttare la segmentazione dei dati mediante clustering. Questo approccio ha permesso di considerare la variabilità dei pazienti in base a caratteristiche simili.

Utilizzando il metodo dell'*Elbow* e l'analisi del coefficiente di *silhouette*, è stato possibile identificare il numero ottimale di cluster. *L'Elbow method* ci ha permesso di osservare il punto in cui l'incremento della somma dei quadrati delle distanze diminuisce significativamente, mentre il coefficiente di silhouette ci ha aiutato a valutare la qualità della separazione dei cluster.

Di seguito i risultati grafici:

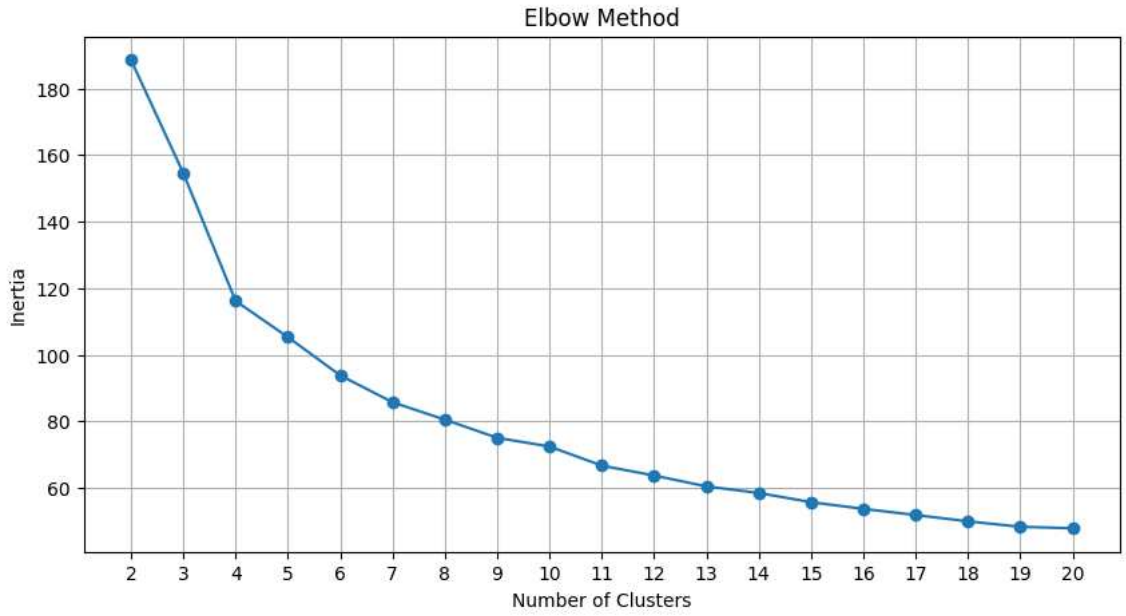

Fig. III.3 Risultati Elbow method

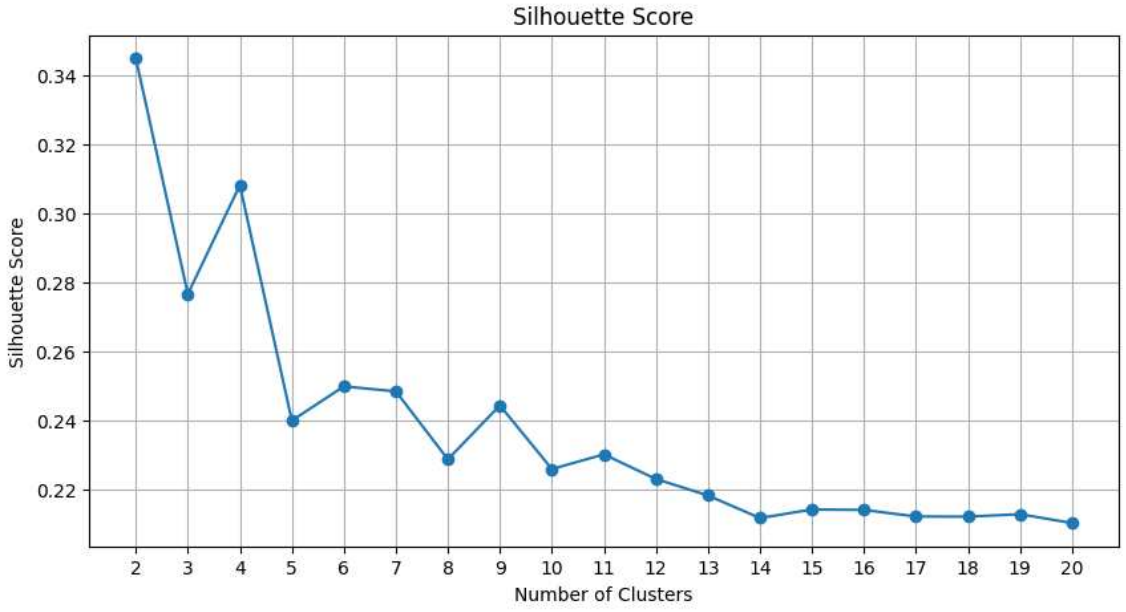

Fig. III.4 Risultati Silhouette score

Una volta determinato il numero ottimale – pari a 4 - sono stati riempiti i valori mancanti all'interno di ogni cluster. Per le variabili continue, è stata utilizzata la media dei valori presenti nel cluster, mentre per le variabili binarie, la moda.

Questo approccio ha il vantaggio di tenere conto delle similitudini tra i pazienti all'interno di ciascun cluster, garantendo un riempimento dei valori mancanti più appropriato e specifico rispetto a metodi globali.

Parallelamente, è stato creato un secondo dataset riempiendo i valori mancanti senza considerare i cluster.

Per le variabili continue, i valori nulli sono stati sostituiti con la media generale dell'intero dataset. Per le variabili binarie, la moda generale.

Questo metodo è più semplice e meno computazionalmente intensivo rispetto al riempimento per cluster, ma potrebbe non catturare la variabilità specifica dei pazienti come il metodo cluster-*based*.

Infine, è stato mantenuto un terzo dataset lasciando i valori nulli invariati.

Analizzando questo dataset, si è potuto osservare l'effetto dei valori mancanti sulle performance dei modelli, confrontando i risultati con quelli ottenuti dai datasets riempiti.

3.1.3 VISUALIZZAZIONE DEI DATI

A questo punto dell'analisi si è ritenuto necessario osservare graficamente le variabili. Si è esaminato come le variabili binarie si distribuiscono rispetto al target "Rischio". Sono stati scelti grafici a barre impilate per rappresentare visualmente queste distribuzioni, permettendo di osservare chiaramente le proporzioni di ciascun livello di rischio all'interno delle categorie binarie.

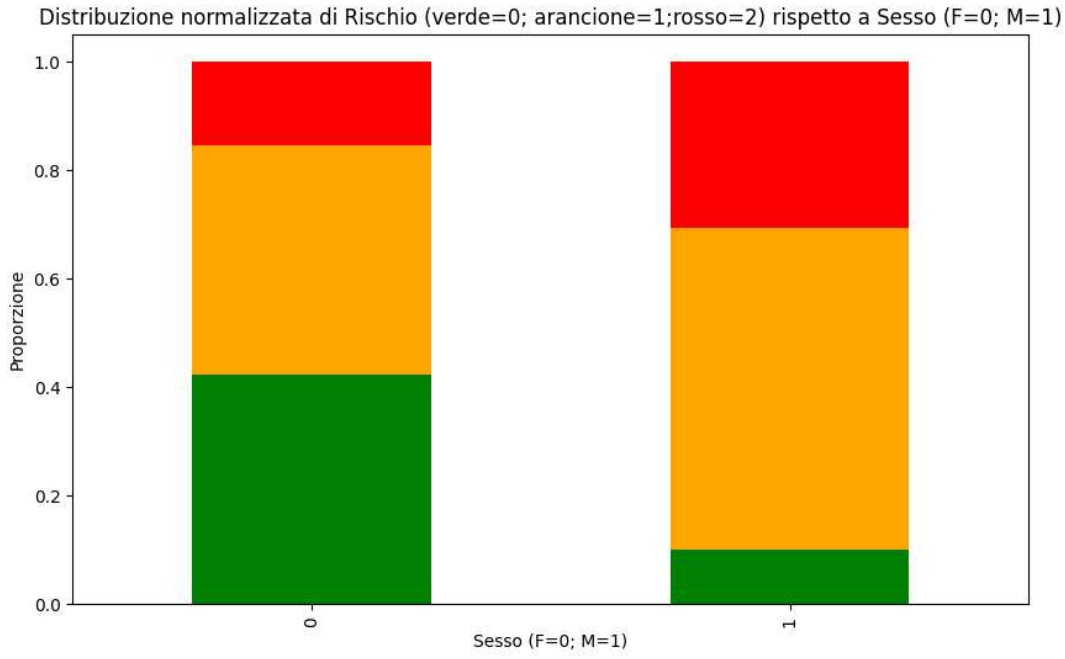

Fig. III.5 Distribuzione di Sesso rispetto a Rischio

L'osservazione riguardante la relazione tra il sesso e il rischio di ipertensione mostra che gli uomini sono più portati ad avere un rischio maggiore di ipertensione.

Uno dei fattori potrebbe essere la differenza nei livelli ormonali tra uomini e donne. Alcuni studi suggeriscono che gli ormoni sessuali come il testosterone possono incrementare la pressione sanguigna.

Risulta, infatti, inferiore al 10% il numero di pazienti maschili con rischio basso, rispetto al 40% femminile. Inoltre, si osserva una proporzione doppia di individui maschili a rischio alto, rispetto alla controparte.

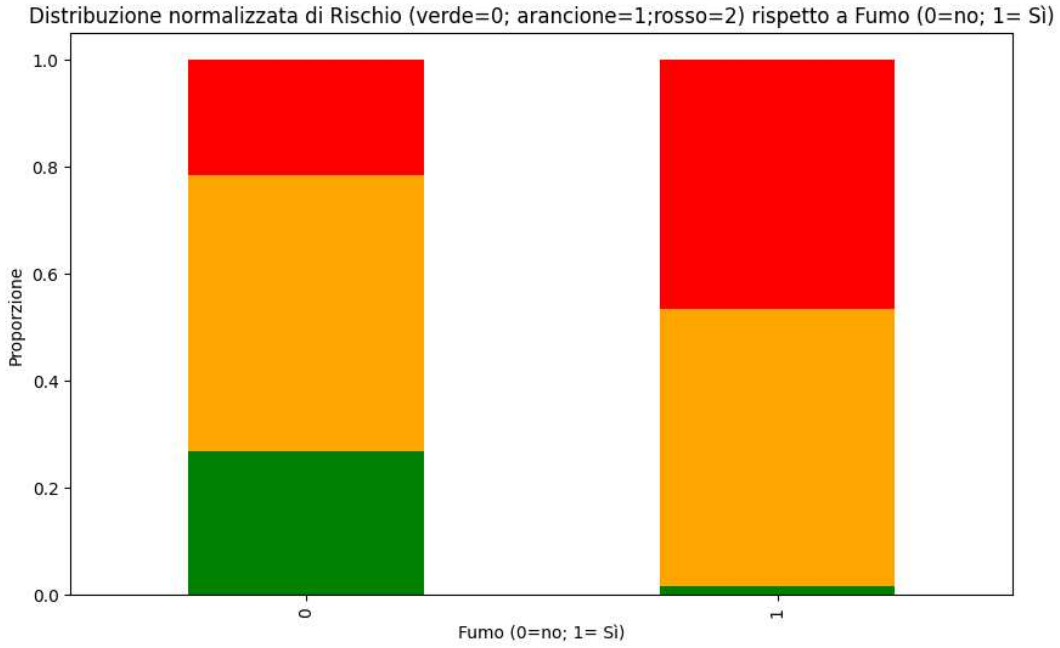

Fig. III.6 Distribuzione di Fumo rispetto a Rischio

Il grafico mostra una correlazione tra il fumo e il rischio di ipertensione confermando quanto sia importante considerare il tabagismo come un fattore di rischio determinante per lo sviluppo di questa patologia. Questa osservazione è in linea con la vasta letteratura clinica che ha

dimostrato ripetutamente il legame tra fumo e ipertensione.

Ci sono diversi meccanismi attraverso i quali il fumo può influenzare la pressione sanguigna. Innanzitutto, le sostanze chimiche nocive presenti nel fumo del tabacco possono danneggiare le pareti dei vasi sanguigni, compromettendo la loro elasticità e aumentando la resistenza vascolare, il che porta a un aumento della pressione arteriosa.

Si osserva come la proporzione di pazienti fumatori ed a rischio basso sia irrisoria a confronto della quasi metà degli individui a rischio alto e l'altra metà intermedio.

3.1.4 ANALISI DELLE CORRELAZIONI

L'analisi delle correlazioni delle variabili presenti nel dataset non direttamente calcolanti lo Score2, ha permesso di studiare se e come vi fossero informazioni utili estraibili da esse, nonostante la loro natura esogena al calcolo dello Score. Il risultato è stato rappresentato in una *heatmap*.

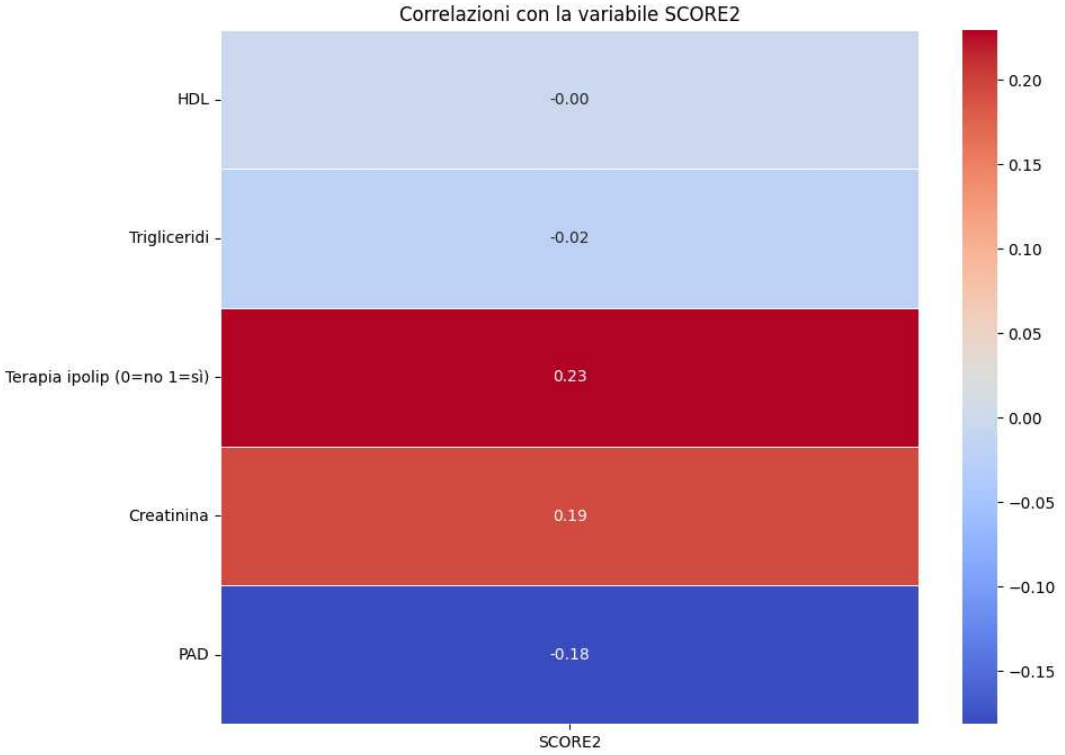

Fig. III.8 Correlazioni variabili e Score2

La scala cromatica, che va dal blu (correlazione negativa) al rosso (correlazione positiva), aiuta a visualizzare l'intensità e la direzione di queste correlazioni.

Osservando HDL (*High-Density Lipoprotein*) e Trigliceridi, notiamo una correlazione praticamente nulla con SCORE2 (-0.00 e -0.02), suggerendo che non abbiano un rapporto significativo con questa variabile target.

Passando alla Terapia ipolipemizzante, notiamo una correlazione positiva di 0.23 con SCORE2, la più alta tra tutte le variabili analizzate. Questo potrebbe suggerire un potenziale effetto diretto della terapia ipolipemizzante sulla variabile target.

La Creatinina, invece, presenta una correlazione di 0.19, indicativa di una leggera relazione positiva, ma comunque non molto forte.

Infine, il PAD (Pressione arteriosa diastolica) mostra una correlazione negativa di -0.18 con SCORE2, suggerendo una lieve relazione inversa.

In sintesi, nessuna delle variabili considerate presenta una correlazione particolarmente forte con la variabile target. Questo potrebbe indicare che, almeno in termini di correlazione lineare semplice, queste variabili non sono predittori sostanziali della variabile target. Tuttavia, è importante ricordare che la mancanza di correlazioni forti non implica necessariamente che queste variabili siano inutili per la previsione di SCORE2. Potrebbero comunque essere importanti in modelli più

complessi che considerano interazioni non lineari o combinazioni di variabili.

3.2 VALUTAZIONE VALORI NULLI E VARIABILI

Avendo introdotto quali siano state le tecniche di gestione di valori nulli, risulta ora necessario spiegare quale di esse sia stata scelta ed il motivo. Inoltre, attraverso l'analisi visuale dei dati e lo studio delle correlazioni, si è potuto studiare se e come ogni variabile incidesse, ponendo le basi per un approfondimento ed una scelta motivata delle variabili predittive da utilizzare.

3.2.1 SCELTA DELLA TECNICA DI GESTIONE DEI VALORI NULLI

Il confronto tra le varie tecniche di gestione è stato condotto mediante la medesima analisi approfondita per ottimizzare un modello di albero di regressione, con l'obiettivo di trovare i migliori parametri per migliorare la precisione del modello. Questo processo di ottimizzazione è stato effettuato sui cinque dataset differenti trattati nel paragrafo 3.1.2, ognuno pre-processato in modi diversi per gestire i valori mancanti e la standardizzazione.

Abbiamo iniziato definendo una serie di parametri chiave dell'albero di regressione da testare. Questi parametri includevano la profondità massima dell'albero, il numero minimo di campioni necessari per dividere un nodo ed il numero minimo di campioni richiesti per formare una foglia. La ricerca dei migliori parametri è stata eseguita tramite una tecnica che valuta sistematicamente tutte le combinazioni possibili dei parametri specificati, utilizzando una validazione incrociata a 10 *fold*. Questo metodo garantisce una valutazione robusta delle performance del modello, riducendo il rischio di *overfitting* e migliorando l'affidabilità dei risultati.

Dopo aver identificato il modello con i migliori parametri, abbiamo

valutato la sua performance calcolando l'errore quadratico medio. Questo ci ha permesso di stimare con precisione l'accuratezza del modello ottimizzato.

Questa procedura di ottimizzazione è stata applicata sui cinque dataset distinti:

- df_fill_std: dove i valori mancanti sono stati riempiti in modo semplice e successivamente i dati sono stati standardizzati.
- df_fill: con valori mancanti riempiti in modo semplice, ma non standardizzati.
- df_cl_std: dove i valori mancanti sono stati riempiti tramite tecniche di clustering e successivamente standardizzati.
- df_cl: con valori mancanti riempiti tramite clustering, ma non standardizzati.
- df: il dataset originale che presenta valori mancanti in alcune variabili e non è stato standardizzato.

Attraverso questa metodologia, siamo riusciti a valutare l'impatto delle diverse tecniche di *pre-processing* sui risultati del modello.

Di seguito i risultati ottenuti.

Dataset utilizzato	MSE da 10-fold Cross Validation
Df_fill_std	0.0000829
Df_fill	0.0000828
Df_cl_std	0.0000831
Df_cl	0.0000830
Df	0.0014898

Si osserva che la tecnica di riempimento dei valori nulli non comporta un significativo mutamento dell'errore. Al contrario, lasciare il dataset originale, rende peggiore la metrica in esame di 18 volte rispetto al migliore dei dataset precedenti.

Questa conclusione porterebbe a scegliere il dataset con i valori nulli riempiti mediante media delle variabili continue e moda delle binarie.

Trattando, però, dati medici, si è preferito approfondire ove fossero i valori nulli e se le variabili che li contenessero fossero o meno importanti per l'analisi.

3.2.1 ANALISI DELL'IMPORTANZA DELLE VARIABILI

L'analisi in merito all'importanza delle singole variabili per il modello di regressione con albero decisionale è stata svolta utilizzando il migliore dei dataset, ovvero df_fill.

L'importanza di ogni variabile viene calcolata durante la costruzione dell'albero, poiché il modello valuta continuamente quale caratteristica è la più migliore per fare predizioni precise. Questa importanza è misurata in base a quanto la caratteristica contribuisce a ridurre l'incertezza nelle previsioni del modello.

Questo valore numerico è calcolato in base a diversi fattori, come la frequenza con cui una caratteristica è utilizzata per dividere i nodi dell'albero e quanto questa divisione contribuisce a migliorare la purezza dei nodi figli.

Di seguito le singole variabili e la loro importanza, in ordine decrescente per quest'ultima.

- Età: 0.861478
- Sesso (F=0; M=1): 0.052247

- PAS: 0.051621
- Fumo (0=no; 1= Sì): 0.026326
- Colesterolo totale: 0.007227
- Creatinina: 0.000404
- HDL: 0.000303
- PAD: 0.000240
- Trigliceridi: 0.000154
- Terapia ipolip (0=no 1=sì): 0.000000

Di seguito la distribuzione cumulativa dell'importanza.

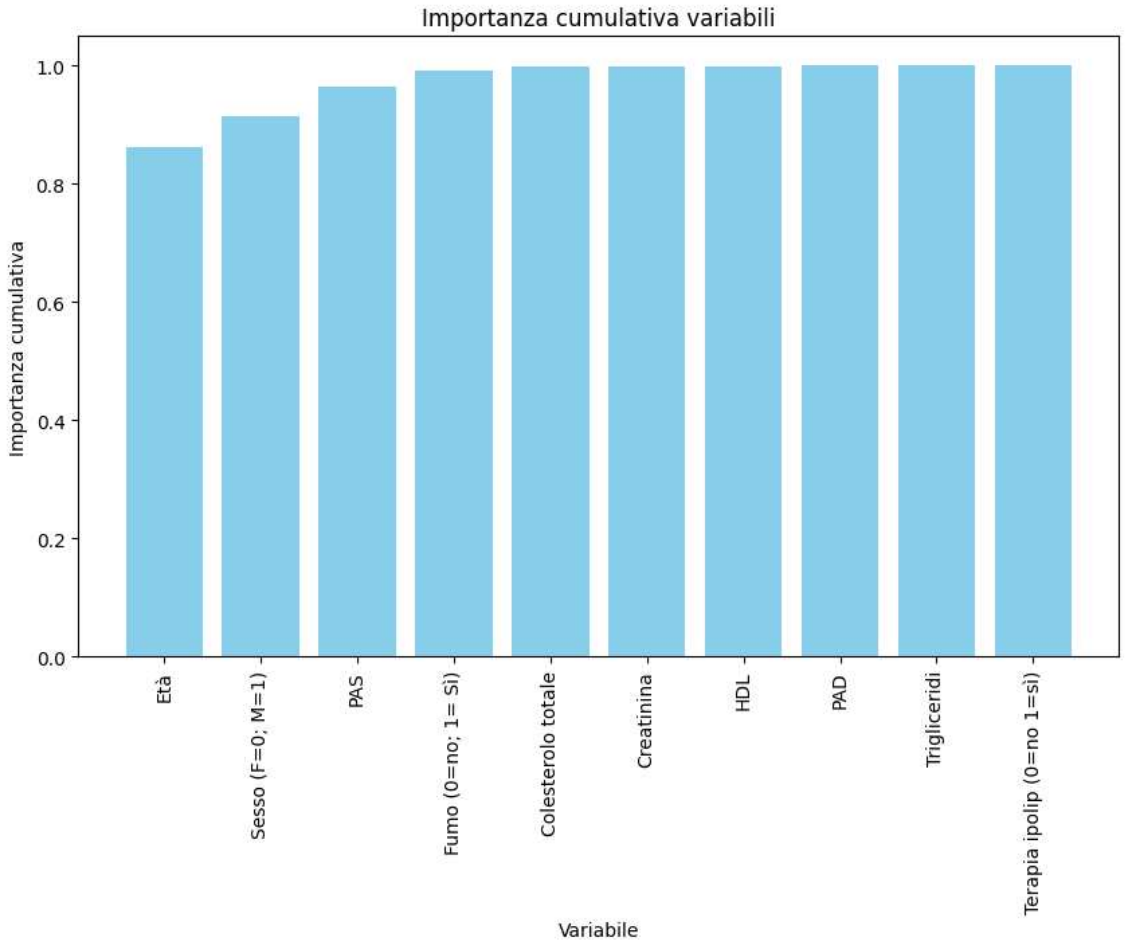

Fig. III.9 Importanza cumulativa variabili

Si osserva come le variabili che non costituiscono direttamente lo Score (Creatinina, HDL, PAD, Trigliceridi, Terapia ipolip) non incidano quasi per nulla nell'albero decisionale. Questa conclusione è coerente con l'analisi delle correlazioni trattata al paragrafo 3.1.4.

3.2.2 SCELTA DELLE VARIABILI DA UTILIZZARE

Dopo aver esaminato attentamente l'importanza delle variabili nel contesto della nostra analisi, si è deciso di eliminare le variabili meno rilevanti. Queste variabili, inoltre, erano le sole a presentare valori mancanti. Di conseguenza, rimuovendo queste variabili si è ottenuto un dataset privo di valori nulli, rendendo superflua l'implementazione di strategie di gestione dei dati mancanti.

Con il dataset ora privo di variabili non rilevanti, l'attenzione è stata posta sulle variabili rimaste, che coincidono esattamente con quelle utilizzate per calcolare lo Score. Questo fatto sottolinea che non vi sono altre variabili escluse dal calcolo che potrebbero essere rilevanti per la valutazione del rischio ipertensivo.

Le variabili rimaste includono il sesso, l'età e l'abitudine al fumo, tutte informazioni descrittive del paziente. Inoltre, abbiamo mantenuto la pressione arteriosa sistolica e il colesterolo totale, che sono variabili soggette a variazioni nel breve periodo e possono introdurre un certo grado di distorsione nelle previsioni.

Riprendendo in considerazione l'importanza delle variabili, si è optato per conservare solo le variabili descrittive. Questa decisione ha permesso

di mantenere un livello di spiegabilità dell'albero intorno al 94%, il che è considerato un compromesso accettabile. Utilizzare solo variabili oggettive che non sono suscettibili a variazioni nel tempo è preferibile, pur sapendo di sacrificare una piccola parte della spiegabilità per questo obiettivo.

4. ADDESTRAMENTO DEI MODELLI

Una volta selezionate le variabili rilevanti, si è proceduto alla costruzione effettiva dei modelli di regressione e classificazione. Per garantire che i modelli fossero addestrati e valutati correttamente, il dataset è stato diviso in due parti: un set di training e un set di test. Questa suddivisione è stata effettuata utilizzando una tecnica chiamata *holdout*, dove l'80% dei dati è stato assegnato al set di training e il restante 20% al set di test.

Un aspetto fondamentale di questa procedura è stato il mantenimento della proporzione tra le classi nella variabile target, che in questo caso rappresenta il Rischio. Per ottenere questo risultato, si è applicata la stratificazione durante il campionamento. Essa assicura che la distribuzione delle classi della variabile target sia simile sia nel set di training sia in quello di test. Questo è particolarmente importante quando si lavora con dati sbilanciati, dove alcune classi possono essere molto meno rappresentate rispetto ad altre.

Questo approccio è cruciale per evitare *bias* nel modello e assicurare che esso sia capace di generalizzare bene su dati non visti, fornendo previsioni accurate indipendentemente dalla classe di rischio.

4.1 MODELLI NON OTTIMIZZATI

Inizialmente, si è deciso di concentrarsi sull'addestramento di un modello di regressione e di un modello di classificazione utilizzando esclusivamente la ricerca dei parametri ottimali e la tecnica della cross-validation a X-fold. Questa scelta è stata fatta per semplificare l'approccio iniziale e garantire che i modelli fossero costruiti con i migliori parametri possibili, massimizzando così la loro accuratezza.

Tuttavia, è importante notare che in questo sotto-capitolo non vengono considerati alcuni aspetti critici che potrebbero influire significativamente sulle prestazioni dei modelli. In particolare, per quanto riguarda il modello di classificazione, non si tiene conto del bilanciamento delle classi. Esso è fondamentale quando si ha a che fare con dati squilibrati, dove una o più classi possono essere sottorappresentate rispetto ad altre. Ignorare questo aspetto può portare a un modello che ha buone prestazioni complessive ma che fallisce nel riconoscere correttamente le classi minoritarie.

Inoltre, un altro aspetto non considerato è la diversa importanza degli errori commessi in situazioni di rischio elevato rispetto a quelle di rischio basso. In molti contesti applicativi, un errore non ha sempre lo

stesso peso: ad esempio, in un sistema di diagnosi medica come questo, un errore di sottostima può avere conseguenze molto più gravi rispetto ad uno di sovrastima del rischio. Pertanto, ignorare la diversa gravità degli errori può portare a modelli che non sono ottimali dal punto di vista della sicurezza.

4.1.1 ALBERO DI REGRESSIONE

Il primo modello addestrato è stato l'albero decisionale di regressione.

Per migliorare le prestazioni del nostro modello, sono stati scelti attentamente i parametri con cui viene costruito l'albero decisionale. Questi parametri influenzano come l'albero si sviluppa e quanto bene può generalizzare su dati nuovi. La "griglia dei parametri" è una raccolta di possibili valori per ciascuno di questi parametri:

- **criterion**: è la funzione che misura la qualità di una suddivisione dell'albero.
- **max_depth**: la profondità massima dell'albero, ossia il numero massimo di livelli di suddivisione.
- **min_samples_split**: il numero minimo di campioni richiesti per poter dividere un nodo.
- **min_samples_leaf**: il numero minimo di campioni che un nodo foglia deve contenere.
- **min_weight_fraction_leaf**: la frazione minima di peso che un nodo foglia deve rappresentare.

- **max_leaf_nodes**: il numero massimo di nodi foglia nell'albero.
- **min_impurity_decrease**: il guadagno minimo di impurità richiesto per effettuare una suddivisione.
- **ccp_alpha**: un parametro di complessità utilizzato per la potatura dell'albero dopo che è stato completamente costruito.

Sono state, dunque, esplorate tutte le possibili combinazioni di parametri, utilizzando la cross-validation a 10-fold per valutare l'efficacia di ciascuna combinazione. La cross-validation divide i dati in parti, utilizza alcune parti per addestrare il modello e altre per testarlo, ripetendo il processo più volte per assicurarsi che i risultati siano affidabili.

Alla fine, è stata restituita la combinazione di parametri che ha ottenuto le migliori prestazioni in termini di errore quadratico medio e mediante questi parametri è stato addestrato il modello di regressione.

Graficamente l'albero risultante è il seguente.

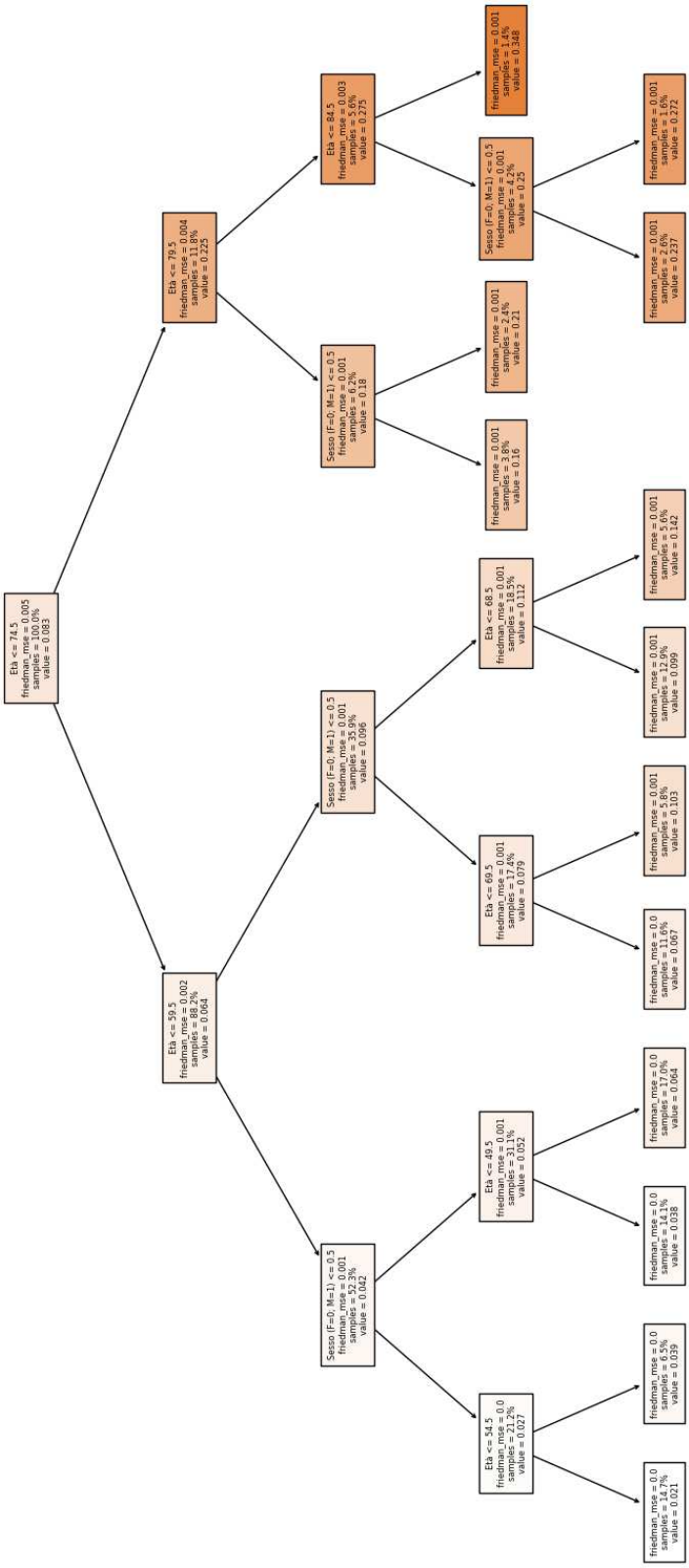

Fig. IV.1 Albero di regressione

È stata utilizzata una rappresentazione cromatica per visualizzare la distribuzione dello score predetto. Questa rappresentazione varia dai toni del bianco all'ocra, seguendo uno score crescente. Tale visualizzazione ha permesso di percepire in modo intuitivo come lo score si modifichi in base ai differenti valori predetti dal modello.

Inoltre, attraverso una corretta parametrizzazione, è stata impostata la profondità massima dell'albero, il numero minimo di osservazioni richieste per effettuare una suddivisione e i criteri di potatura. Questi parametri sono stati selezionati accuratamente per evitare la frammentazione dei dati, un fenomeno che avrebbe potuto causare *overfitting* al modello. In altre parole, è stato trovato un equilibrio ottimale tra la complessità del modello e la sua capacità di generalizzazione, garantendo che il modello fosse in grado di fare previsioni accurate non solo sui dati di addestramento, ma anche su nuovi dati.

Infatti, l'errore quadratico medio risultante dal confronto tra i valori predetti ed i reali del test set è risultato pari a 0.0004630, valore decisamente basso che verrà in seguito analizzato nel dettaglio.

4.1.2 ALBERO DI CLASSIFICAZIONE

Il secondo modello è quello di classificazione.

I parametri considerati nella nostra griglia sono i medesimi rispetto al paragrafo precedente, a parte ccp_alpha e la diversa ricerca di parametri per quanto riguarda il criterio.

Una volta definita la griglia dei parametri, è stata utilizzata la medesima tecnica della cross-validation a 10-fold per valutare le prestazioni del modello su diverse configurazioni di parametri, ma in questo caso stratificata per mantenere la corretta proporzione tra le classi tra i fold di train e quello di test.

Una volta ottenuti i parametri migliori in termini di massimizzazione dell'accuratezza, è stato addestrato l'albero di classificazione.

L'albero risultante è il seguente.

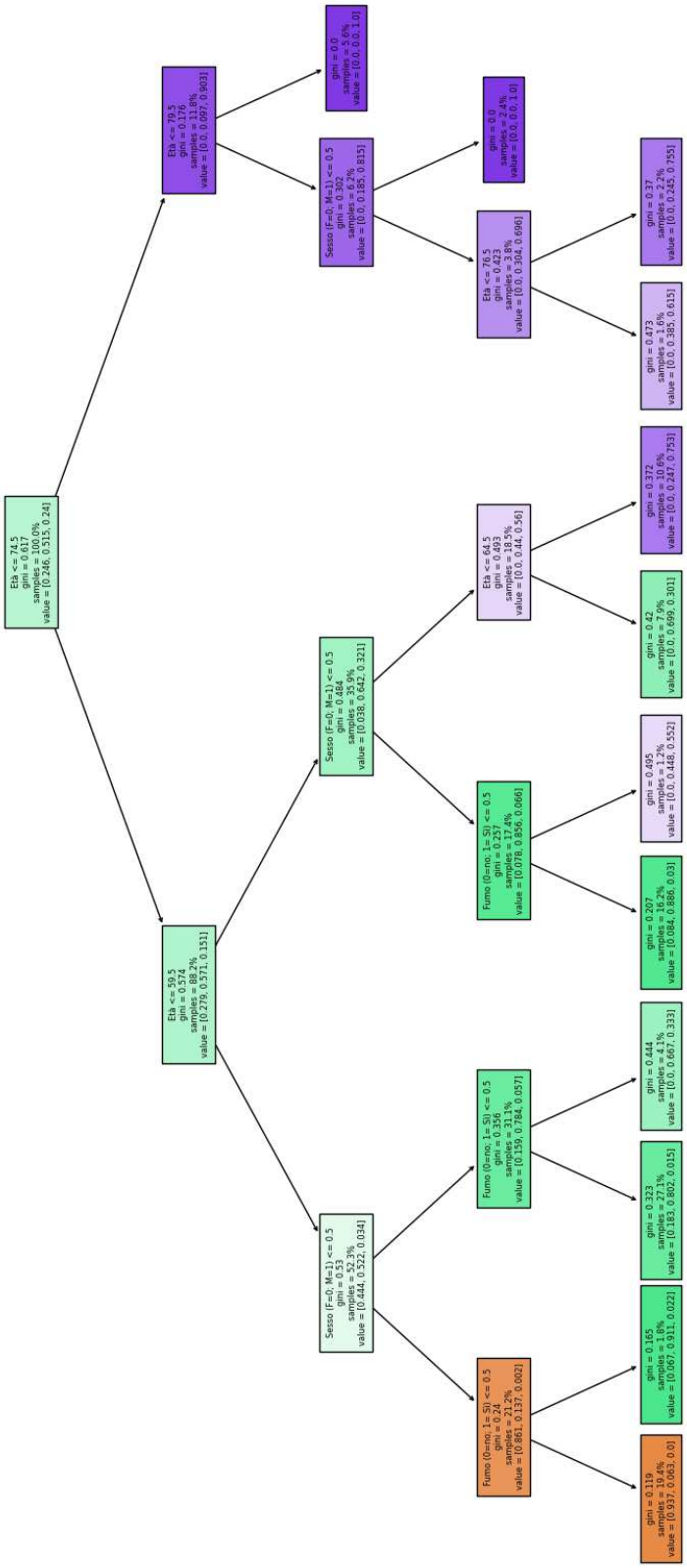

Fig. IV.2 Albero di classificazione

È stata utilizzata una rappresentazione cromatica per visualizzare la classe predominante in ogni nodo dell'albero (arancione per il rischio basso, verde per quello intermedio e viola per quello alto).

Anche in questo caso, questi parametri sono stati selezionati accuratamente per evitare la frammentazione dei dati, un fenomeno che avrebbe potuto causare *overfitting* al modello.

L'errore assoluto medio risultante dal confronto tra le classi predette e le reali del test set è risultato pari a 0.2258, valore decisamente basso che verrà in seguito analizzato nel dettaglio.

4.2 MODELLI OTTIMIZZATI

Una volta addestrati i modelli senza ottimizzazione, si è ritenuto utile ed interessante analizzare anche modelli ove venissero trattati lo sbilanciamento delle classi e la ponderazione differente degli errori in base al rischio.

Infatti, oltre alla ricerca dei parametri ottimali per l'albero decisionale, utilizzando metodi di ribilanciamento delle classi si può garantire che il modello dia la giusta importanza a tutte le classi durante l'addestramento, migliorando così la sua capacità di generalizzazione su dati non visti.

Inoltre, considerando anche la diversa importanza degli errori commessi in situazioni di rischio elevato rispetto a quelle di rischio basso, si garantisce che il modello tenga conto della diversa gravità degli errori durante l'addestramento e l'ottimizzazione.

L'integrazione di metodi di ribilanciamento delle classi e di ponderazione degli errori rappresenta un passo significativo verso la costruzione di un modello di classificazione più accurato ed affidabile, specialmente in casi di rischio in cui la precisione e la sicurezza sono cruciali.

4.2.1 ALBERO DI REGRESSIONE OTTIMIZZATO

L'obiettivo principale è trovare i migliori pesi, chiamati weight_1 e weight_2, da attribuire alle osservazioni in sede di ricerca del miglior criterio di split ad ogni passo di avanzamento dell'albero decisionale. In questo senso, i pazienti che presentano uno Score più elevato vengono presi maggiormente in considerazione e viceversa.

Il codice esegue una serie di iterazioni per testare diverse combinazioni di pesi e parametri del modello. Durante ogni iterazione, vengono valutati gli errori di previsione sullo Score reale attenzionando gli errori di sottostima del rischio e, se necessario, vengono aggiornati i pesi e i parametri del modello per migliorare le prestazioni complessive.

In primo luogo, si cerca di ottimizzare la sottostima relativa ai pazienti con Score superiore a 0.09. Alle iterazioni successive, qualora si dovesse ritrovare il medesimo errore, l'algoritmo analizza l'errore di sottostima dei pazienti con Score tra 0.04 e 0.09, estremi inclusi, aggiornando eventualmente i pesi migliori ed i parametri dell'albero.

Mediante i pesi ed i parametri ottimi, si è potuto addestrare il seguente albero di regressione ottimizzato.

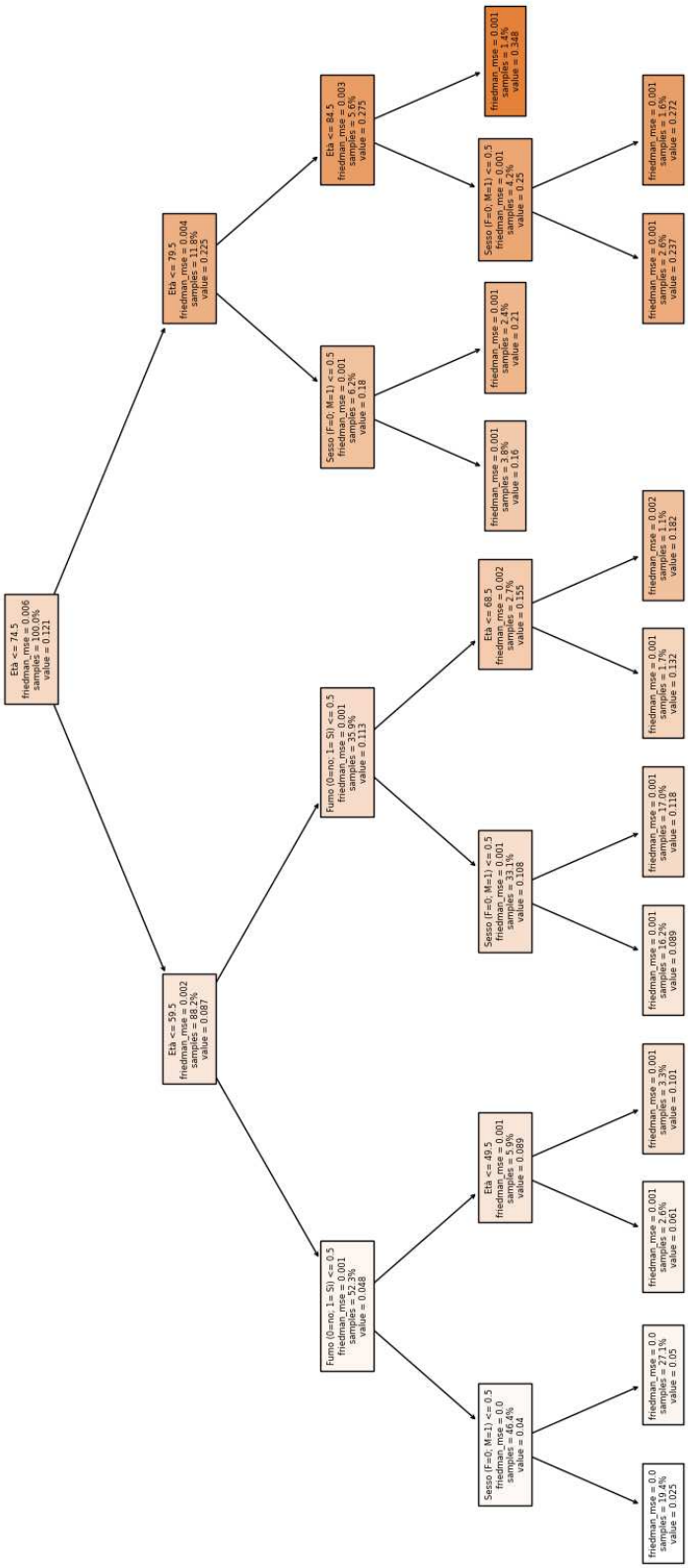

Fig. IV.3 Albero di regressione ottimizzato

Anche in questo caso è stata utilizzata una rappresentazione cromatica per visualizzare la distribuzione dello score predetto. Questa rappresentazione varia dai toni del bianco all'ocra, seguendo uno score crescente.

L'aver introdotto un'ottimizzazione per prevedere più accuratamente alcuni tipi di osservazioni ha certamente avvicinato il modello al rischio di overfitting. In tal senso, risultava ancora più importante una corretta parametrizzazione per poter bilanciare bene l'errore di addestramento e di generalizzazione.

Infatti, l'errore quadratico medio risultante dal confronto tra i valori predetti ed i reali del test set è risultato pari a 0.0005883, valore più alto rispetto all'albero di regressione non ottimizzato, ma comprensibilmente, dato che non sono stati curati alcuni tipi di errori, a discapito di altri presenti in minor numero, ma di importanza maggiore.

4.2.2 ALBERO DI CLASSIFICAZIONE OTTIMIZZATO

L'ottimizzazione dei pesi nel modello di classificazione mira a migliorare la precisione nel distinguere tra le diverse classi di rischio, con un'attenzione particolare a quelle più alte. Il processo inizia con l'assegnazione di pesi utili al ribilanciamento delle classi, in modo tale che si correggesse in parte lo sbilanciamento tra di esse.

Si definisce, in seguito, una serie di parametri del modello da esplorare durante la ricerca del miglior modello possibile, allo stesso modo dell'albero di classificazione non ottimizzato.

L'ottimizzazione in questo caso viene messa in atto mediante la ricerca dei migliori pesi da assegnare alle osservazioni durante l'addestramento, in modo tale da ponderare l'importanza di quelle osservazioni a rischio più elevato. I pesi infatti sono riferiti sia ai pazienti con rischio alto, che a quello intermedio, cercando di ottimizzare il primo citato e, in seguito, il secondo.

Il modello ottimale derivante dalla ricerca dei parametri migliori viene quindi utilizzato per effettuare predizioni cross-validate, e queste predizioni vengono confrontate con i valori reali di rischio per calcolare

gli errori di classificazione. In particolare, si conta il numero di errori di sottostima, il caso più delicato ed attenzionato durante una diagnosi medica, per le osservazioni a rischio alto. Se il numero di errori per questa classe è inferiore al minimo riscontrato in precedenza, vengono aggiornati i pesi e i parametri considerati ottimali. Se il numero di errori per la classe di rischio alto è uguale al minimo precedente, vengono valutati gli errori per la classe di rischio intermedio e si aggiornano i pesi e i parametri solo se anche questi errori sono inferiori al minimo riscontrato.

Al termine del processo, vengono determinati i pesi ottimali che minimizzano gli errori di classificazione per le classi di rischio medio e alto. Questo approccio permette di addestrare un modello di classificazione più accurato e bilanciato, migliorando la capacità del modello di distinguere tra diverse classi di rischio, con un'attenzione particolare alla riduzione degli errori nelle classi di rischio più critiche.

I pesi ed i parametri migliori vanno a costruire il seguente albero.

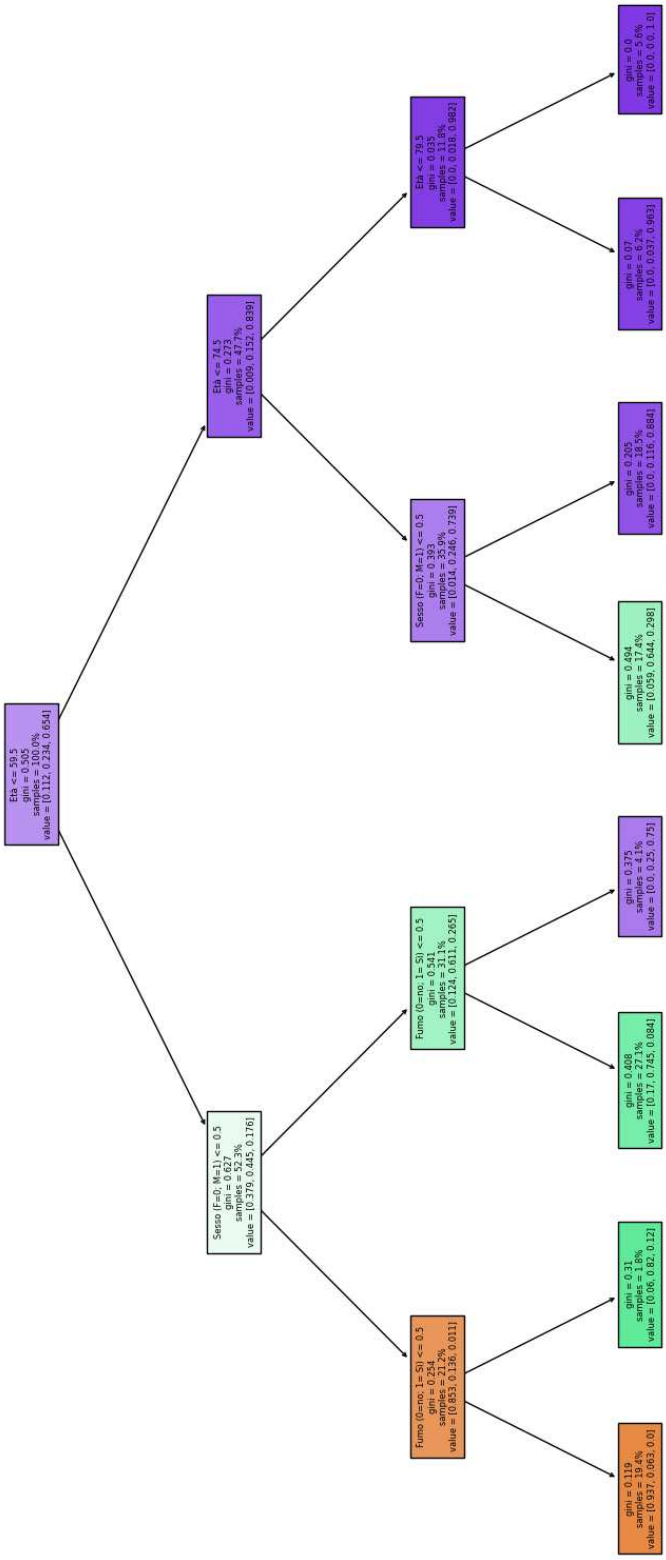

Fig. IV.4 Albero di classificazione ottimizzato

Anche in questo caso si può visualizzare la classe predominante in ogni nodo dell'albero (arancione per il rischio basso, verde per quello intermedio e viola per quello alto).

Questi parametri sono stati selezionati accuratamente per evitare la frammentazione dei dati, un fenomeno che avrebbe potuto causare *overfitting* al modello. Un rischio, questo, già accresciuto dalla presenza dell'ottimizzazione dei pesi.

L'errore assoluto medio risultante dal confronto tra le classi predette e le reali del test set è risultato pari a 0.2471, valore leggermente più alto del caso non ottimizzato, ma risulta essere un *trade-off* comunque accettabile.

4.3 VALUTAZIONE DEI MODELLI

L'analisi è proseguita con la valutazione, volta alla scelta, dei due modelli ottimizzati e no, dato che si è ritenuta importante, ma non decisiva l'osservazione rispettivamente di MSE di regressione e MAE di classificazione.

L'attenzione si è spostata sull'analisi visuale degli errori, in quanto in contesto clinico l'errore di sottostima comporta una valutazione del rischio pericolosa per la salute del paziente. Al contrario, quello di sovrastima richiede uno sforzo economico, strumentale e di tempo maggiore per l'ospedale, allocando risorse a pazienti che magari non necessiterebbero un'attenzione tale.

4.3.1 VISUALIZZAZIONE ISTOGRAMMA DEGLI ERRORI DI CLASSIFICAZIONE

Per iniziare, si è deciso di valutare l'accuratezza del modello di classificazione, concentrandoci su come questo modello performa in relazione alle diverse classi di rischio.

Per facilitare la comprensione e l'interpretazione dei risultati, si è scelto di rappresentare i dati attraverso una visualizzazione a istogrammi. Sull'asse delle ascisse sono riportate le tre classi reali di rischio, che sono ulteriormente suddivise in tre categorie di colore: rosso per la sottostima, verde per la classificazione corretta e blu per la sovrastima.

Di seguito i risultati del modello di classificazione non ottimizzato.

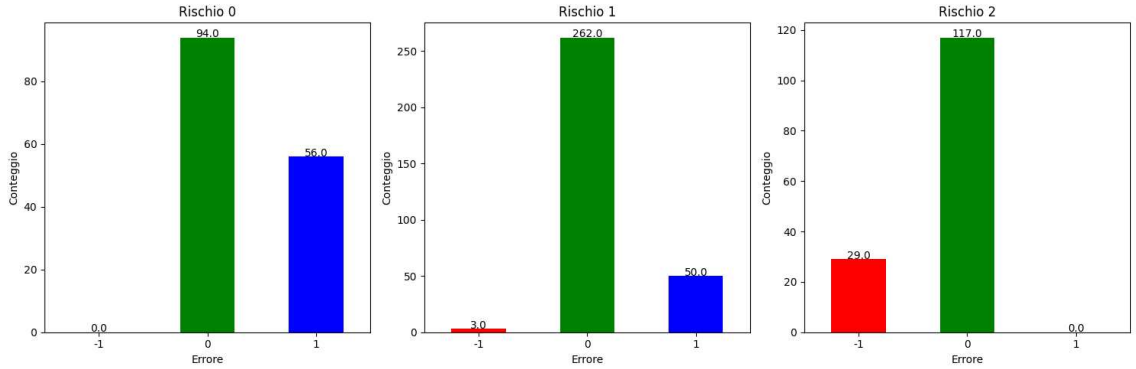

Fig. IV.5 Istogrammi errori di classificazione non ottimizzata

Rischio reale	Sottostime	Corretti	Sovrastime
Basso (0)	-	62.67%	37.33%
Intermedio (1)	0.95%	83.17%	15.87%
Alto (2)	19.86%	80.14%	-

Analizzando i risultati del modello, emerge una tendenza significativa nella predizione delle classi di rischio intermedio e alto, che risulta essere particolarmente accurata. Questo potrebbe essere attribuito alla maggiore rappresentanza di casi appartenenti a tali categorie nel dataset. Tuttavia, va notato che il modello tende a sovrastimare più del dovuto i pazienti a rischio basso, con oltre un terzo dei casi identificati erroneamente come appartenenti a questa categoria. Allo stesso modo, si registra una tendenza a sottostimare il rischio nelle classi più elevate, con circa un paziente su cinque mal classificato.

Questi risultati hanno implicazioni cruciali, specialmente in termini di gestione delle risorse ospedaliere. Una sovrastima così significativa, soprattutto tra le classi 0 e 1, potrebbe portare a esami e trattamenti superflui, creando inefficienze e aumentando i costi. Inoltre, le sottostime, che si attestano intorno al 20%, sono particolarmente

preoccupanti poiché indicano una sottovalutazione della gravità delle condizioni cliniche dei pazienti.

Questi dati sottolineano l'importanza di continuare a perfezionare il modello, non solo per migliorare la sua capacità di identificare accuratamente i diversi livelli di rischio, ma anche per garantire una gestione ottimale delle risorse sanitarie e una valutazione precisa della gravità delle condizioni dei pazienti.

Il modello ottimizzato tenta proprio di essere più cauto nei confronti delle sottostime tra classe 2 reale e classe 1 predetta.

Di seguito i suoi risultati.

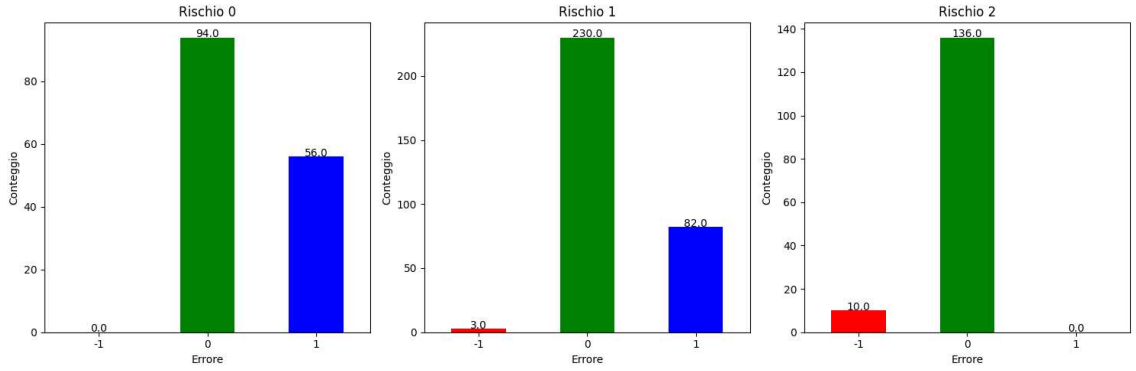

Fig. IV.6 Istogrammi errori di classificazione ottimizzata

Rischio reale	Sottostime	Corretti	Sovrastime
Basso (0)	-	62.67%	37.33%
Intermedio (1)	0.95%	73.02%	26.03%
Alto (2)	6.85%	93.15%	-

L'analisi dettagliata rivela interessanti variazioni nella classificazione dei pazienti tra le classi 1 e 2. Mentre la proporzione di corrette ed errori rimane sostanzialmente costante tra le classi 0 e 1, si osservano significative differenze quando si passa alla classe 2.

In particolare, il modello ottimizzato mostra una tendenza a sovrastimare il numero di pazienti a rischio intermedio, con un aumento di 32 casi rispetto alla versione precedente. Tuttavia, è importante notare che questa sovrastima è parzialmente compensata da un aumento di 19 casi di pazienti correttamente classificati a rischio alto.

Questi risultati suggeriscono che il modello ottimizzato sia in grado di meglio identificare e classificare i pazienti con un rischio alto, riducendo notevolmente il pericolo di sottostima. In altre parole, c'è una riduzione del 65.62% dei casi di sottostima rispetto al modello precedente.

Tuttavia, questa migliore capacità di identificare i casi a rischio alto comporta anche una maggiore richiesta di attenzione, con un aumento del 64% dei pazienti che richiedono ulteriori valutazioni rispetto ai pazienti sovrastimati nella versione precedente del modello.

Questi risultati sottolineano l'importanza di bilanciare attentamente la sensibilità e la specificità del modello, garantendo che sia in grado di identificare accuratamente i casi a rischio elevato senza generare un'eccessiva sovrastima, che potrebbe comportare l'allocazione inefficiente di risorse sanitarie.

4.3.2 VISUALIZZAZIONE BOXPLOT DEGLI ERRORI DI REGRESSIONE

Dopo un'attenta analisi dei modelli di classificazione e dei loro risultati generalmente soddisfacenti, abbiamo deciso di approfondire il rapporto tra questi modelli e i rispettivi modelli di regressione. Questo approccio ha permesso di esaminare le previsioni in modo più dettagliato e di identificare le aree in cui il modello di classificazione commette errori.

Un modo efficace per visualizzare questa relazione è tramite i boxplot degli errori di regressione, dove sull'asse delle ascisse rappresentiamo il Rischio reale, ulteriormente suddiviso per tipo di errore di classificazione e sull'asse delle ordinate l'errore di regressione.

Questo approccio ci consente di individuare rapidamente le discrepanze tra le previsioni del modello e la realtà clinica, evidenziando eventuali aree in cui il modello di regressione può migliorare.

Inoltre, questa visualizzazione ci permette di valutare la coerenza delle previsioni tra i modelli di classificazione e di regressione, consentendoci di identificare eventuali discrepanze nelle previsioni tra i due approcci. Ciò ci aiuta a comprendere meglio il comportamento dei modelli e a individuare eventuali aree di miglioramento per aumentare l'accuratezza

delle previsioni e migliorare la gestione dei rischi clinici.

Di seguito i risultati del modello di regressione non ottimizzato.

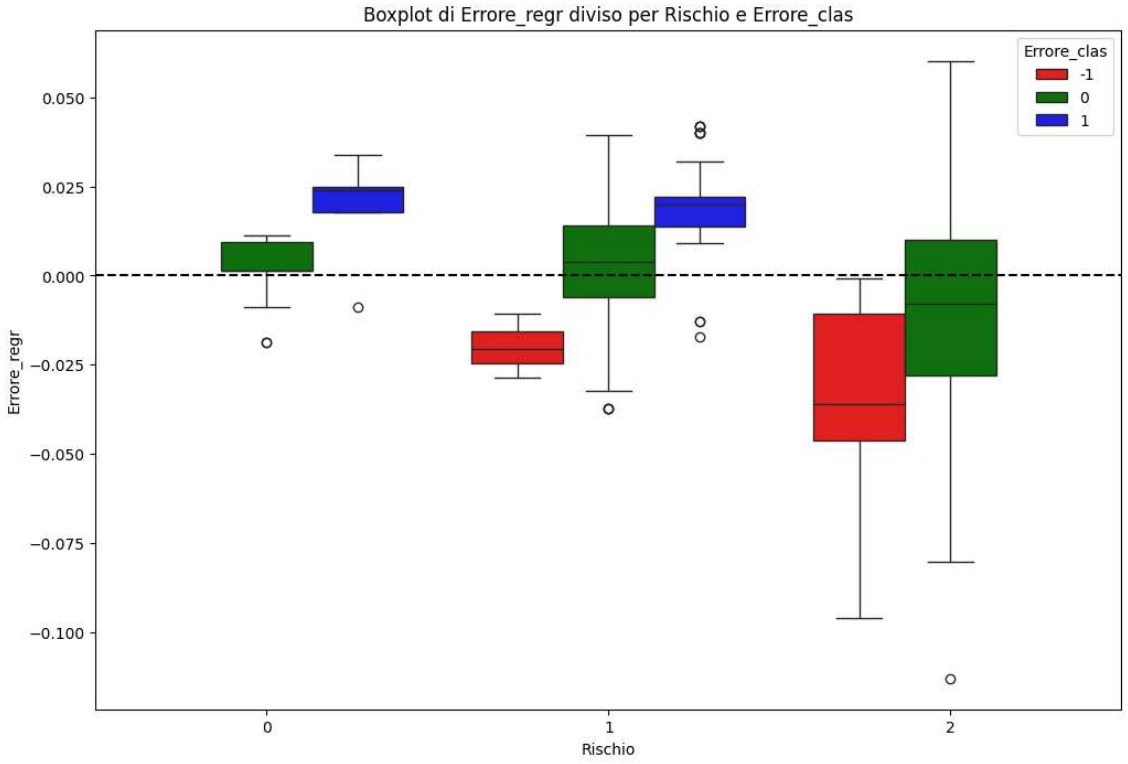

Fig. IV.7 Boxplot errori di regressione non ottimizzata

Studiando il grafico, si nota che per il rischio reale basso, l'errore di regressione è esiguo, trattandosi di una quasi perfetta predizione per i correttamente classificati ed un leggero errore di sovrastima nell'intorno del 2.5% per i pazienti classificati di classe intermedia.

In merito ai pazienti di rischio reale intermedio, la sottostima è esigua, intorno al -2.5%; i correttamente classificati riscontrano un errore tra il 4% ed il -4%, mostrando una leggera incertezza di previsione; infine, i

sovrastimati presentano un errore di regressione esiguo, intorno al 2.5%.

I pazienti di rischio reale alto, invece, mostrano maggiore difficoltà per il modello di regressione, in quanto i pazienti sottostimati risultano essere erroneamente predetti per valori che raggiungono persino il -10%, con una mediana intorno al -3%. Le osservazioni correttamente classificate di classe alta, nuovamente, riscontrano errori tra il -8% ed il 6%, indicando come il modello di classificazione sia maggiormente efficace.

Mediante il modello ottimizzato si è provato a ridurre l'incertezza legata prettamente alle sottostime di classe reale alta. Di seguito i risultati.

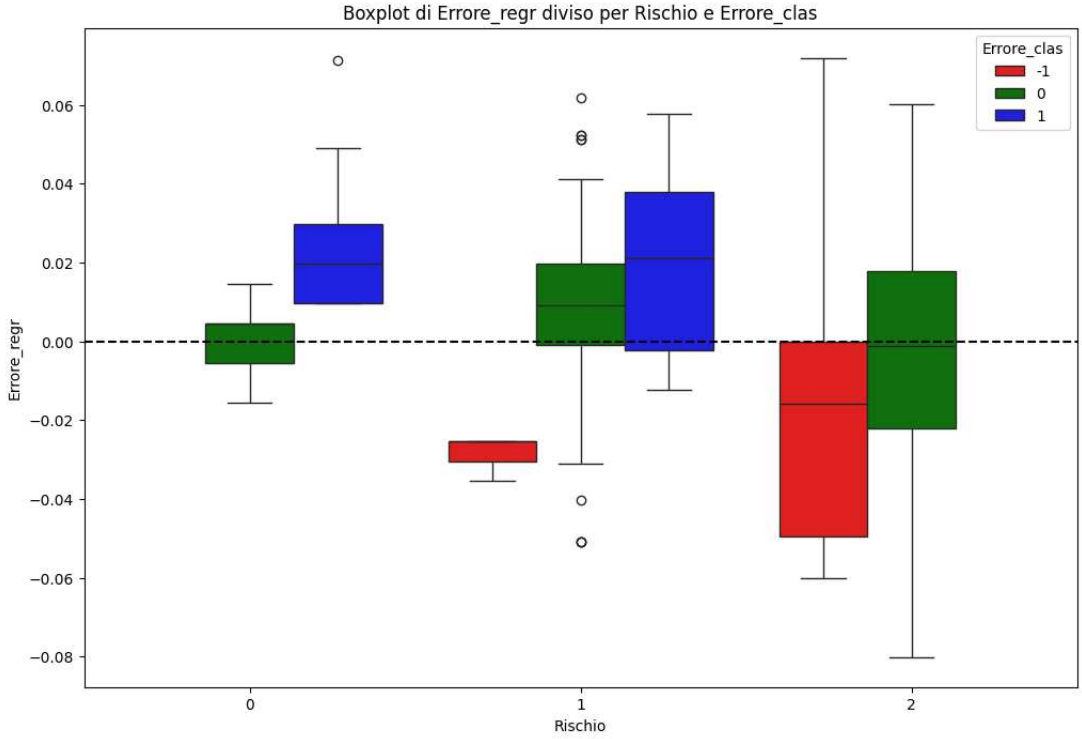

Fig. IV.8 Boxplot errori di regressione ottimizzata

Per i pazienti di rischio reale basso, si nota che mentre i correttamente classificati mostrano un errore di regressione simile al precedente, coloro che vengono sovrastimati raggiungono errori fino al 5%, nonostante la mediana scenda al 2%.

Per quanto concerne le osservazioni che presentano rischio reale intermedio, coloro che vengono sottostimati e correttamente classificati presentano un range di errore simile al precedente; i pazienti sovrastimati, invece, mostrano molta incertezza nella previsione mediante regressione, in quanto si osservano errori tra il -1% ed il 6%.

Infine, i pazienti di rischio reale alto mostrano molta incertezza per quanto riguarda le sottostime, mostrando addirittura valori sovrastimati per lo Score; l'errore per coloro che vengono correttamente classificati è simile al precedente modello.

In generale si osserva che il modello di regressione ottimizzato crei maggiore incertezza rispetto alla controparte non ottimizzata.

4.3.3 INTERVALLI DI CONFIDENZA

Nel corso dell'analisi, è emersa l'esigenza di integrare lo studio degli intervalli di confidenza per tenere conto dell'incertezza nelle previsioni, sia nei modelli di classificazione che in quelli di regressione.

Per questo scopo, sono stati registrati i valori reali di Rischio (per la classificazione) e Score (per la regressione) per ogni foglia di entrambi i modelli. Questo ha permesso di ottenere, con un livello di confidenza del 90%, l'intervallo di valori possibili per ogni foglia.

Per quanto riguarda le osservazioni del set di test, sono state valutate le probabilità a posteriori che venissero classificate nelle classi di rischio bassa, intermedia e alta, unitamente ai relativi intervalli di confidenza. Invece, per la regressione, sono state analizzate le probabilità a priori che l'osservazione venisse classificata in una determinata classe, in funzione dei valori dell'intervallo di confidenza di regressione predetto.

L'introduzione degli intervalli di confidenza ha portato alla necessità di definire due nuove classi di previsione dedicate all'incertezza. Infatti, sono stati identificati casi limite in cui si verificavano sovrapposizioni parziali degli intervalli per la classificazione o un'assegnazione non

chiara della classe risultante dal valore di previsione per la regressione. Per questi casi, sono state introdotte due nuove classi di output: "0-1" e "1-2", che rappresentano l'incertezza rispettivamente per i casi a cavallo tra il rischio basso e intermedio e tra quest'ultima e quella alta.

L'adozione di questa soluzione ha comportato, sia nei modelli ottimizzati che in quelli non ottimizzati, una riduzione dei casi di sovrastima della classe di rischio reale intermedia, a fronte però di una diminuzione del numero di osservazioni correttamente classificate nella classe di rischio alta.

Tuttavia, è importante sottolineare che una classificazione come incerta è stata comunque considerata corretta se non divergeva significativamente dalla classe di rischio reale. In questo senso, è preferibile avere una chiara consapevolezza dell'ambiguità di una classificazione piuttosto che una falsa certezza.

4.3.4 UNIONE DEI MODELLI

A questo punto dell'analisi, è stata esplorata la possibilità di creare un sistema di classificazione unico che integrasse i modelli di classificazione e regressione, favorendo una collaborazione tra di essi.

L'idea centrale era quella di sfruttare i punti di forza di entrambi i modelli per ottenere una maggiore robustezza e accuratezza nelle previsioni. Il funzionamento previsto era il seguente:

- Concordanza: se entrambi i modelli avessero concordato sulla previsione, questa sarebbe stata riportata come output definitivo. In questo caso, la somma delle "certezze" dei due modelli rafforzava la fiducia nella predizione.

- Incertezza: Se uno dei due modelli avesse presentato incertezza nella previsione, l'output finale avrebbe riportato questa incertezza. Questo approccio garantiva la trasparenza, evitando di riportare una falsa certezza quando la fiducia in una specifica classe non era completa.

- Disaccordo: Se i due modelli avessero prodotto previsioni differenti, l'output sarebbe diventato l’incertezza tra le due classi

predette discordanti. In questo scenario, l'incertezza derivava dalla discrepanza tra le valutazioni dei due modelli, indicando la necessità di un'ulteriore analisi o di informazioni supplementari per una classificazione definitiva.

L'obiettivo principale di questo sistema collaborativo era quello di ridurre al minimo gli errori di classificazione. Introducendo l'incertezza come output in caso di dubbio, si privilegiava la trasparenza rispetto a una falsa certezza che avrebbe potuto portare a decisioni errate.

4.3.5 RISULTATI DEI MODELLI

Al termine della fase di costruzione dei modelli, si è palesata la necessità di scegliere quale fosse il migliore. Di seguito le matrici di confusione delle quattro combinazioni possibili tra regressione e classificazione ottimizzate e no.

Classe Reale *Classi predette*

	0	0-1	1	1-2	2
0	64	47	37	2	0
1	1	72	138	63	41
2	0	2	11	30	103

Fig. IV.9 Matrice di confusione modelli uniti e non ottimizzati

In figura si osservano i risultati del modello unito ove sia regressione che classificazione non sono state ottimizzate. Il totale dei correttamente classificati (in verde), è pari all'84.62% dei pazienti. In dettaglio, i 64 individui correttamente a rischio basso (pari al 10.47% del totale) sono coloro ai quali giustamente non risulta essere utile alcun

approfondimento medico.

Per quanto concerne le sottostime, si osserva la presenza di 2 pazienti (0.33% del totale) che vengono predetti a rischio basso/intermedio, mentre sarebbero da classificare a rischio alto. Questi casi sono stati approfonditi e come si osserva dal grafico seguente, il motivo consta nell'età (41 e 43 anni).

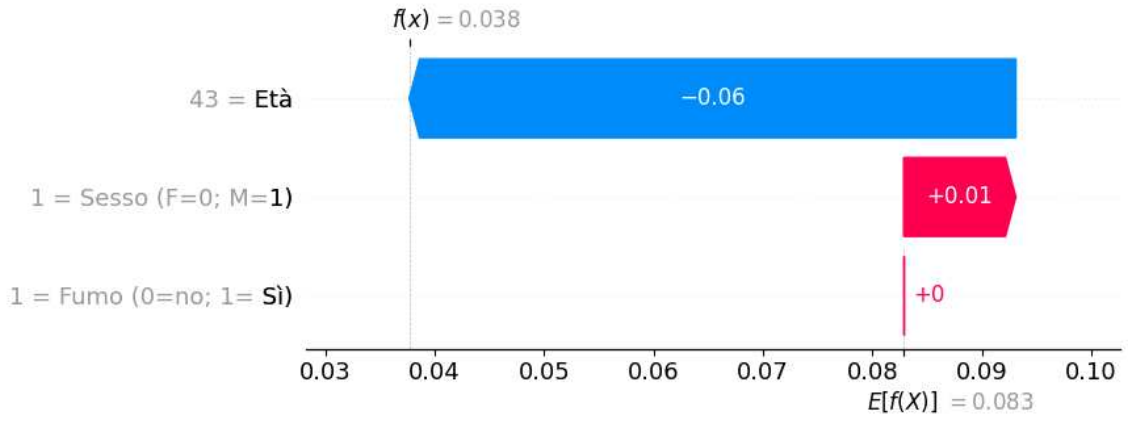

Fig. IV.10 Analisi Shap Values di un paziente sottostimato

Entrambi i pazienti risultano essere maschi e fumatori, entrambi fattori noti per aumentare il rischio, come confermato durante la fase di esplorazione iniziale dei dati. Tuttavia, l'età di questo paziente è tra le più basse nel dataset, il che contribuisce a una riduzione del rischio predetto del 6%. Per illustrare meglio questa situazione, consideriamo il valore reale dello Score di questo individuo, che è pari a 0.10. In contrasto, il modello prevede un valore di 0.04, dunque si colloca

logicamente tra la classe 0 e la classe 1, piuttosto che nella classe reale 2.

Questo modello risulta, dunque, sottostimare il 2.29% dei pazienti, la maggior parte relativi al rischio reale alto.

Per quanto concerne le sovrastime, invece, riscontriamo un 13.09% del totale di pazienti, ben divisi tra sovrastima di rischio basso verso l'intermedio e quest'ultimo predetto a rischio alto.

Si osservano due casi di sovrastima più importante, poiché di rischio reale basso, ma predetto medio/alto. Di seguito l'analisi di uno dei due (poiché con caratteristiche simili).

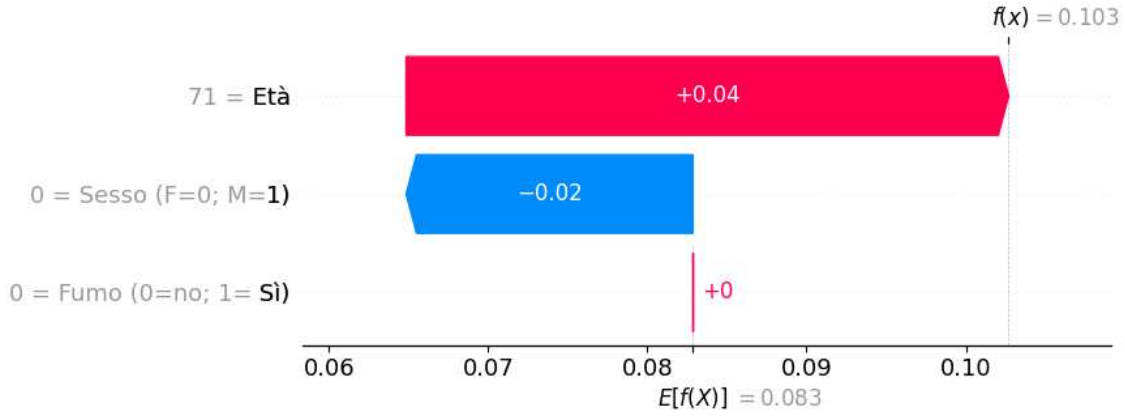

Fig. IV.11 Analisi Shap Values di un paziente sovrastimato

Entrambe le pazienti sono donne e non fumatrici, dunque si tenderà ad abbassare, giustamente, la previsione dello Score. Però il fattore età incide accrescendo di 4 punti la previsione, portandola a 0.10, mentre lo

Score reale risulta 0.07.

Procedendo al secondo candidato, di seguito viene riportata la matrice di confusione del modello ove la regressione non viene ottimizzata, al contrario della classificazione.

Classe Reale	*Classi predette*				
	0	0-1	1	1-2	2
0	64	47	37	2	0
1	1	66	136	40	72
2	0	0	10	8	128

Fig. IV.12 Matrice di confusione regressione non ottimizzata e classificazione ottimizzata

Il totale dei correttamente classificati è pari all'80.03% dei pazienti, in calo rispetto al modello precedente. Rimangono gli stessi coloro ben classificati di rischio basso.

Per quanto concerne le sottostime, si osserva l'assenza dei due pazienti suddetti con grave sottostima. Il modello di classificazione ottimizzata, come da suo scopo, è riuscito a classificare meglio i pazienti ad alto

rischio.

Questo modello risulta, dunque, sottostimare l'1.80% dei pazienti, in calo rispetto al precedente, la maggior parte relativi al rischio reale alto.

Per quanto concerne le sovrastime, invece, riscontriamo un aumento fino al 18.17% del totale di pazienti, con il picco riguardante coloro di rischio intermedio, predetti a rischio alto. Purtroppo, l'ottimizzazione tende appunto a cautelare la previsione, accrescendo il numero di pazienti predetti a rischio alto.

Si osserva nuovamente la presenza dei due casi di sovrastima più importante, poiché, come dimostrato, relativi al modello di regressione non ottimizzata.

In seguito, viene proposta la matrice di confusione del terzo modello candidato, ove la regressione è stata ottimizzata, al contrario della classificazione.

Classe Reale	*Classi predette*				
	0	0-1	1	1-2	2
0	94	0	56	0	0
1	3	0	237	34	41
2	0	0	13	30	103

Fig. IV.13 Matrice di confusione regressione ottimizzata e classificazione non ottimizzata

Il totale dei correttamente classificati è pari all'81.51% dei pazienti, in aumento rispetto al modello precedente, ma in calo rispetto al primo. Sale al 15.38% il numero di pazienti correttamente classificati a rischio basso, dunque per cui risultano superflui ulteriori esami. L'aumento è riscontrabile anche per i pazienti correttamente classificati a rischio intermedio, a discapito del numero di incerti tra classe 0 ed 1, pari a 0. Evidentemente, la regressione ottimizzata tende a non prevedere l'esistenza di pazienti in questa fascia.

Per quanto concerne le sottostime, si osserva nuovamente l'assenza dei due pazienti suddetti con grave sottostima.

Questo modello risulta, dunque, sottostimare il 2.62% dei pazienti, il

dato più alto finora registrato, incrementando entrambi i casi di sottostima presenti.

Per quanto concerne le sovrastime, invece, riscontriamo un 15.88% del totale di pazienti, abbastanza bilanciato con una leggera predominanza di sovrastime da rischio basso ad intermedio.

Si osserva, invece, l'assenza dei due casi di sovrastima più importante.

Infine, viene proposta la matrice di confusione dell'ultimo candidato, ove entrambi i modelli sono stati ottimizzati.

Classe Reale	*Classi predette*				
	0	0-1	1	1-2	2
0	94	0	56	2	0
1	3	0	227	13	72
2	0	0	5	13	128

Fig. IV.13 Matrice di confusione modelli uniti ed ottimizzati

Il totale dei correttamente classificati è pari all'77.74% dei pazienti, il dato più basso registrato. Troviamo nuovamente al 15.38% il numero di

pazienti correttamente classificati a rischio basso, constatando ulteriormente l'impatto della regressione ottimizzata in questo campo.

Per quanto concerne le sottostime, si osserva nuovamente l'assenza dei due pazienti suddetti con grave sottostima.

Questo modello risulta, dunque, sottostimare l'1.31% dei pazienti, il dato più basso finora registrato.

Per quanto concerne le sovrastime, invece, riscontriamo un 20.95% del totale di pazienti, piuttosto bilanciato con una predominanza di sovrastime da rischio intermedio ad alto, a causa della cautela predittiva risultante dalla classificazione ottimizzata.

Si osserva, nuovamente, l'assenza dei due casi di sovrastima più importante.

5. CONCLUSIONE

In quest'ultimo capitolo, vengono descritte la *ratio* con la quale è stato scelto il modello migliore tra quelli proposti e le proposte di implementazione di tale modello e di una visualizzazione *ad hoc* per l'analisi visuale delle previsioni. L'idea è quella di fornire al clinico uno strumento automatico di classificazione del rischio ipertensivo ed uno semi-automatico, come supporto alle decisioni. Infine, qualche considerazione e proposta di approfondimento relativa allo studio.

5.1 SCELTA DEL MODELLO

Nel paragrafo 4.3.5 sono stati proposti vari modelli, ciascuno con i suoi punti di forza e debolezza, a volte simili e altre volte completamente opposti, a causa delle diverse combinazioni di tecniche di regressione e classificazione utilizzate. Questo ha reso complesso il processo di valutazione dei modelli, poiché è emerso il problema di come bilanciare i vari aspetti da considerare. Minimizzare l'errore generale oppure il numero totale di sottostime o sovrastime non era un criterio sufficiente da solo, poiché eccellere in uno di questi obiettivi spesso comportava risultati peggiori in un altro.

Di fronte a questa sfida, è stato fondamentale il confronto con i medici, che possiedono una profonda conoscenza del dominio clinico. Questi esperti hanno suggerito di porre particolare attenzione alle sottostime quando la classe predetta è 0 ma il rischio reale è intermedio, poiché in tali casi i pazienti potrebbero essere erroneamente mandati a casa senza ulteriori indagini. Hanno, inoltre, sottolineato l'importanza di ridurre le sovrastime tra la classe predetta 1 e quella reale 0. In questi casi, ai pazienti classificati come a rischio intermedio viene prescritta una terapia ipolipemizzante con statine, che abbassa il livello di colesterolo.

Sebbene tale trattamento non rappresenti un rischio per la salute di chi ha livelli di colesterolo bassi, comporta comunque costi clinici significativi legati ai medicinali e agli strumenti di somministrazione. Infine, i medici hanno indicato come cruciale la riduzione degli errori di sovrastima tra la classe predetta 2 e quella reale 1, poiché la terapia per i pazienti ad alto rischio è ancora più aggressiva e gli esami necessari per confermare il rischio sono più invasivi e costosi.

Seguendo queste linee guida, in ordine di priorità, si è notato che tutti i modelli tendevano a sottostimare in maniera simile i rischi bassi e intermedi. Di conseguenza, nessuno dei modelli è stato escluso immediatamente sulla base di questo criterio. Successivamente, analizzando il numero di sovrastime tra la classe 0 e 1, sono stati scelti i modelli costituiti dalla regressione non ottimizzata, poiché la versione ottimizzata produceva il 51,35% in più di sovrastime di questo tipo. Infine, considerando l'ultimo criterio di scelta, si è deciso di scartare il modello con classificazione ottimizzata, che generava il 75,60% in più di sovrastime tra la classe predetta 2 e quella reale 1.

Alla fine, la scelta è ricaduta sul primo modello proposto, caratterizzato da una regressione e una classificazione entrambe non ottimizzate. Questo modello è stato ritenuto il migliore in base alle linee guida fornite

dai medici e ai risultati delle analisi effettuate, rappresentando un compromesso accettabile tra i diversi obiettivi di minimizzazione degli errori.

5.2 ANALISI DEI RISULTATI

Risulta necessario, dopo la scelta del modello, analizzare più minuziosamente i vari casi che si presentano in base ai risultati.

Si riporta di seguito, nuovamente, la matrice di confusione, per praticità di studio.

Classe Reale	*Classi predette*				
	0	0-1	1	1-2	2
0	64	47	37	2	0
1	1	72	138	63	41
2	0	2	11	30	103
Tot.	65	121	186	95	144

Fig. V.1 Matrice di confusione modelli uniti e non ottimizzati

Ponendosi nei panni dei medici, si osserva che 65 pazienti vengono predetti di classe 0 e, perciò, non viene condotto alcun approfondimento clinico, garantendo all'ospedale un risparmio in tempo, strumenti e

personale del 10.47% rispetto al totale dei pazienti analizzati allo stato attuale. Inoltre, di questi il 98.46% risulta corretto.

Passando alla previsione 0-1, il 98.35% dei classificati risulta correttamente in quell'intervallo di rischio. Per questa categoria di pazienti, i medici tendono ad approfondire l'anamnesi in sede di colloquio con l'individuo ed, eventualmente, approfondiscono con l'esame del colesterolo. In questo modo, probabilmente, anche l'1.65% errato potrebbe essere correttamente classificato in seconda sede.

Per quanto concerne i pazienti predetti di classe 1, il 74.19% viene correttamente classificato, mentre il 19.89% viene sovrastimato ed il restante 5.92% sottostimato. I medici, come citato nel paragrafo precedente, operano mediante somministrazione di terapia ipolipemizzante con statina, per ridurre il colesterolo. Inoltre, conducono esami ulteriori per comprendere meglio la situazione del paziente. Questo *modus operandi* risulterebbe parzialmente efficace verso coloro ai quali viene sottostimato il rischio, mentre risulterebbe superfluo e, dunque, costoso per l'ospedale per i pazienti sovrastimati.

Proseguendo, i pazienti il cui rischio predetto ricade nell'intervallo 1-2, risultano corretti per il 97.89%, mentre a coloro che vengono sovrastimati viene somministrata la suddetta terapia e vengono svolti

esami ancora più approfonditi. Questo comporta una spesa ancora maggiore per l'ospedale, ma il numero esiguo di questi casi non è motivo di preoccupazione eccessiva.

Infine, i pazienti che vengono classificati di rischio alto, risultano corretti per il 71.53%, mentre il restante 28.47% viene sovrastimato. I medici ritengono che questa sovrastima sia importante, ma non come la precedente, poiché ai pazienti a rischio alto viene somministrata una terapia ancora più incisiva sul livello di colesterolo e gli esami vengono effettuati con maggiore celerità ed in maggior numero. In questo modo, comunque, un paziente di rischio intermedio avrebbe dei benefici, a discapito di spese moderatamente maggiori dell'ospedale.

5.3 PROPOSTA DI IMPLEMENTAZIONE

Una volta ottenuti tutti i risultati necessari, è stato fondamentale capire quali informazioni fornire ai medici e come presentarle in modo efficace. Il *trade-off* principale riguardava l'equilibrio tra l'interpretabilità dei risultati predittivi e la loro correttezza statistica.

Considerando il punto di vista dei professionisti del settore, era evidente che fosse necessario uno strumento che combinasse intuitività e accuratezza, permettendo al medico di interpretare i dati in modo autonomo. Si è quindi deciso di fornire ai clinici due diversi tipi di risultati:

- Modello Automatico di Previsione: Come descritto precedentemente, questo modello utilizza una combinazione di modelli di classificazione e regressione non ottimizzati per fornire una classe di rischio. Questo approccio genera una previsione unica basata su modelli che collaborano tra loro.

- Visualizzazione Separata dei Modelli: Questo strumento presenta in un unico grafico i risultati dei due modelli (classificazione e regressione) in modo separato. Questa visualizzazione permette di

osservare chiaramente come ciascun modello contribuisce alla previsione complessiva.

Di seguito si riporta un esempio dell'output generato dalla visualizzazione, volta al supporto alle decisioni.

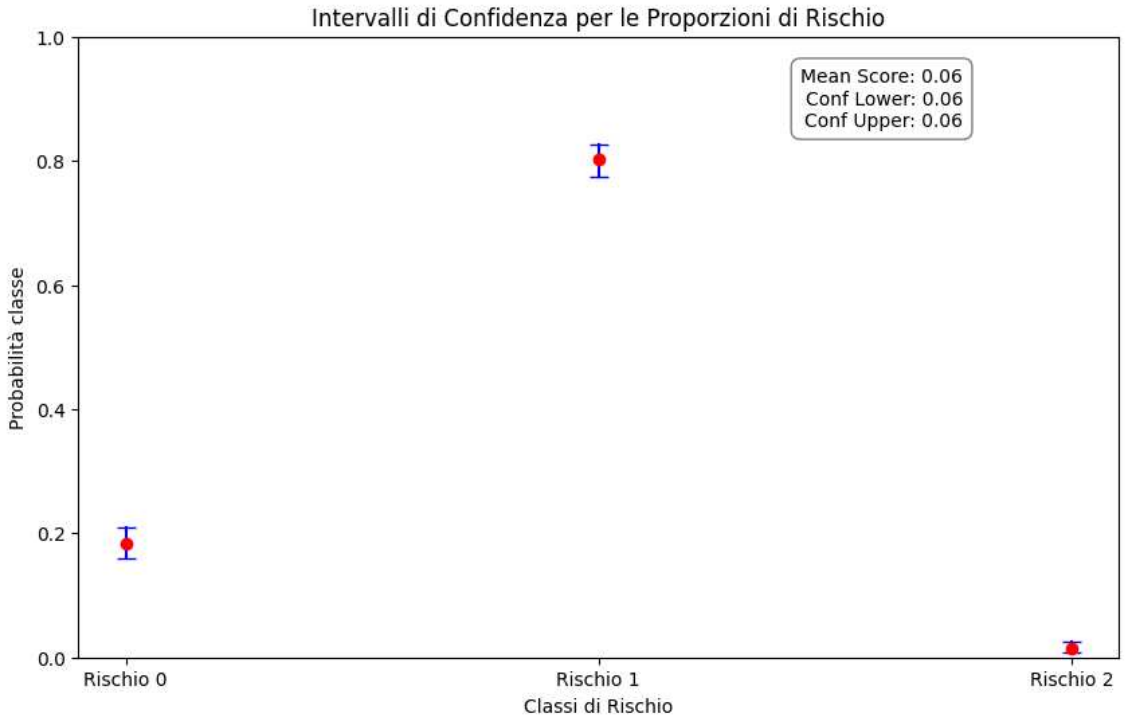

Fig. V.2 Visualizzazione proposta per supporto alle decisioni

Questo è l'output di una osservazione del test set estratta casualmente, il cui Score reale risulta 0.05 e la classe di rischio reale è pari a 1. In questo caso il modello di classificazione mostra che all'80% si prevede che il rischio sia effettivamente quello intermedio, al 20% sia quello basso e quasi nulla la probabilità che sia a rischio alto. Ogni classe ha il suo intervallo di confidenza, così da mostrare graficamente i casi in cui non

sia lampante la classe predetta proposta dal modello. In alto a destra, nel box, i risultati del modello di regressione mostrano un intervallo molto piccolo, intorno al valore predetto di 0.06, di un solo punto percentuale maggiore rispetto allo Score reale.

I due modelli separati, a livello di performance predittive, operano in maniera peggiore rispetto al sistema automatico di collaborazione e previsione del rischio presentato al paragrafo precedente. Tuttavia, questo tipo di visualizzazione aiuta il medico a comprendere se vi sia incertezza o meno nella previsione, potendo aggiungere al risultato il suo giudizio professionale, ponderando i risultati con eventuali altre osservazioni effettuate in sede di visita del paziente.

Si ritiene che entrambi gli strumenti, semi-automatico visuale ed automatico predittivo, possano influenzare la decisione del medico in maniera non invasiva e senza l'ambizione di sostituire il giudizio del clinico.

5.4 CONSIDERAZIONI FINALI

Nel corso di questa tesi, è stato esplorato e sviluppato un sistema di classificazione del rischio ipertensivo basato su modelli di regressione e classificazione. La ricerca si è focalizzata sulla creazione di uno strumento in grado di supportare i medici dell'INRCA in sede di anamnesi del paziente.

Dopo un'attenta analisi delle diverse combinazioni di modelli di regressione e classificazione, abbiamo identificato il modello migliore in base alle linee guida fornite dai medici. Questo modello, caratterizzato da una regressione e una classificazione non ottimizzate, è stato scelto per la sua capacità di bilanciare efficacemente tra la minimizzazione degli errori di sottostima e sovrastima.

I risultati ottenuti hanno mostrato una buona performance del modello nel classificare correttamente i pazienti, con un tasso di classificazione corretta del 84.62%, di sottostima del 2.29% e sovrastima del 13.09%.

Inoltre, è stata creata una visualizzazione *ad hoc* che permette ai medici di vedere separatamente i risultati dei modelli di regressione e classificazione. Questa aiuta a identificare incertezze nelle previsioni e a

integrare il giudizio professionale del medico con i dati forniti dal modello. Questo approccio non solo migliora la trasparenza del processo decisionale, ma permette anche ai clinici di fare aggiustamenti basati su osservazioni cliniche non catturate dal modello.

Il sistema proposto offre un supporto decisionale significativo senza sostituire il giudizio clinico. Il modello automatico di classificazione del rischio fornisce una base solida su cui i medici possono fare affidamento, mentre il sistema di visualizzazione semi-automatico offre flessibilità e un controllo maggiore sui risultati. Questo duplice approccio può portare a decisioni cliniche più informate e potenzialmente a migliori esiti per i pazienti.

Per future ricerche, si suggerisce di curare maggiormente la raccolta dei dati, così da poter fornire ai modelli un maggior numero di variabili e osservazioni, rendendo sempre più robusti i risultati. L'ideale sarebbe poter attingere a dati nazionali, così da poter introdurre col tempo uno strumento di *screening* ipertensivo della popolazione.

In conclusione, questo lavoro rappresenta un passo avanti nel campo della previsione del rischio ipertensivo, offrendo strumenti utili e pratici per supportare il lavoro dei clinici e migliorare la qualità delle cure fornite ai pazienti.

6. APPENDICE TECNICA

```
tree_regr = DecisionTreeRegressor(
    ccp_alpha=0.0,
    criterion='friedman_mse',
    max_depth=4,
    max_leaf_nodes=15,
    min_impurity_decrease=0.01,
    min_samples_leaf=10,
    min_samples_split=9,
    min_weight_fraction_leaf=0.01,
    random_state=42
)
model_regr = tree_regr.fit(X_train, y_regr_train)
```

Fig. VI.1 Modello di regressione

```
tree_clas = DecisionTreeClassifier(
    class_weight=None,
    criterion='gini',
    max_depth=4,
    max_features=None,
    max_leaf_nodes=15,
    min_impurity_decrease=0.0,
    min_samples_leaf=8,
    min_samples_split=7,
    min_weight_fraction_leaf=0.0,
    random_state=42
)
model_clas = tree_clas.fit(X_train, y_clas_train)
```

Fig. VI.2 Modello di classificazione

```python
weight_1 = 1
weight_2 = 1
param_grid = {
    'criterion': ['squared_error', 'friedman_mse'],
    'max_depth': list(range(3, 5)),
    'min_samples_split': list(range(7, 12)),
    'min_samples_leaf': list(range(8, 14)),
    'min_weight_fraction_leaf': [0.0, 0.01],
    'max_leaf_nodes': list(range(10, 31, 5)),
    'min_impurity_decrease': [0.0, 0.01, 0.05],
    'ccp_alpha': [0.0, 0.01]
}
best_params = {}
best_weight1 = 1
best_weight2 = 1
best_model = None
count_err_1 = float('inf')
count_err_2 = float('inf')
X_tree = X_train
df_tree = train
for weight_1 in range(1,3):
    for weight_2 in range(1,5):
        print("Peso_1:", weight_1)
        print("Peso_2:", weight_2)
        grid_search_cv = GridSearchCV(estimator=DecisionTreeRegressor(), param_grid=param_grid, cv=5, scoring='neg_mean_squared_error', verbose=1, n_jobs=-1)
        grid_search_cv.fit(X_tree, y_regr_train, sample_weight=[weight_2 if score2 > 0.09 else (weight_1 if 0.03 < score2 < 0.10 else 1) for score2 in y_regr_train])
        cv_model = grid_search_cv.best_estimator_
        y_pred_cv = cross_val_predict(cv_model, X_tree, y_regr_train, cv=10)
        df_pred = df_tree.copy()
        df_pred["Predizioni"] = y_pred_cv
        df_pred["Errore"] = df_pred['Predizioni'] - df_pred['SCORE2']
        numero_di_righe_2 = ((df_pred['SCORE2'] > 0.09) & (df_pred['Errore'] < 0)).sum()
        if numero_di_righe_2 < count_err_2:
            count_err_2 = numero_di_righe_2
            print("curr_err2:", count_err_2)
            best_params = grid_search_cv.best_params_
            best_weight2 = weight_2
            best_model = cv_model
        elif numero_di_righe_2 == count_err_2:
            numero_di_righe_1 = ((df_pred['SCORE2'] > 0.03) & (df_pred['SCORE2'] < 0.10) & (df_pred['Errore'] < 0)).sum()
            print("Err_1_pre:", count_err_1)
            if numero_di_righe_1 < count_err_1:
                count_err_1 = numero_di_righe_1
                print("curr_err1:", count_err_1)
                print("curr_err2:", count_err_2)
                best_params = grid_search_cv.best_params_
                best_weight1 = weight_1
                best_weight2 = weight_2
                best_model = cv_model
```

Fig. VI.3 Codice per ottimizzazione modello di regressione

```python
weight_0 = 1.0
class_weights = {0: 1, 1: 1, 2: 2}
param_grid = {
    'criterion': ['gini', 'entropy'],
    'max_depth': list(range(3, 5)),
    'min_samples_split': list(range(7, 13)),
    'min_samples_leaf': list(range(8, 12)),
    'min_weight_fraction_leaf': [0.0, 0.01],
    'max_leaf_nodes': list(range(10, 31, 5)),
    'min_impurity_decrease': [0.0, 0.01]
}
best_params = {}
best_weight_1 = 1
best_weight_2 = 1
best_model = None
count_err_1 = float('inf')
count_err_2 = float('inf')
X_tree = X_train
df_tree = train
for weight_1 in range(1,3):
    for weight_2 in range(1,5):
        sample_weights = [weight_2 if rischio == 2 else (weight_1 if rischio == 1 else weight_0) for rischio in y_clas_train]
        print("Peso 1:", weight_1)
        print("Peso 2:", weight_2)
        grid_search_cv = GridSearchCV(estimator=DecisionTreeClassifier(random_state=42, class_weight=class_weights), param_grid=param_grid,
                                      cv=StratifiedKFold(n_splits=10, shuffle=True, random_state=42), scoring='accuracy', verbose=1, n_jobs=-1)
        grid_search_cv.fit(X_tree, y_clas_train, sample_weight=sample_weights)
        cv_model = grid_search_cv.best_estimator_
        y_pred_cv = cross_val_predict(cv_model, X_tree, y_clas_train, cv=StratifiedKFold(n_splits=10, shuffle=True, random_state=42))
        df_pred = df_tree.copy()
        df_pred["Predizioni"] = y_pred_cv
        df_pred["Errore"] = df_pred['Predizioni'] - df_pred['Rischio (verde=0; arancione=1;rosso=2)']
        numero_di_righe_rischio2 = ((df_pred['Errore'] == -1) & (df_pred['Rischio (verde=0; arancione=1;rosso=2)'] == 2)).sum()
        if numero_di_righe_rischio2 < count_err_2:
            count_err_2 = numero_di_righe_rischio2
            print("current_err_2:", count_err_2)
            best_params = grid_search_cv.best_params_
            best_weight_2 = weight_2
            best_model = cv_model
        elif numero_di_righe_rischio2 == count_err_2:
            numero_di_righe_rischio1 = ((df_pred['Errore'] == -1) & (df_pred['Rischio (verde=0; arancione=1;rosso=2)'] == 1)).sum()
            print("err_1_pre:", count_err_1)
            if numero_di_righe_rischio1 < count_err_1:
                count_err_1 = numero_di_righe_rischio1
                print("current_err_2:", count_err_2)
                print("current_err_1:", count_err_1)
                best_params = grid_search_cv.best_params_
                best_weight_1 = weight_1
                best_weight_2 = weight_2
                best_model = cv_model
```

Fig. VI.4 Codice per ottimizzazione modello di classificazione

7. BIBLIOGRAFIA E SITOGRAFIA

- https://www.who.int/news-room/fact-sheets/detail/hypertension
- https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html
- https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html