POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Sviluppo di modelli predittivi dello stato fisio-patologico nel paziente settico e analisi del bias socio-economico

Tesi di Laurea Triennale in
Ingegneria Biomedica

Autori:

Riccardo Sanna
Emma Torracca
Pietro Maria Zangrando

Codice persona: 10808761; 10806927; 10809427
Relatore: Prof. Riccardo Barbieri
Correlatore: Cristian Drudi

Anno Accademico: 2024-25

Abstract

Introduzione: La sepsi è una delle condizioni più pericolose, essendo la terza causa di morte a livello mondiale e la prima causa di morte negli ospedali, oltre che una delle più onerose nelle Unità di Terapia Intensiva (UTI). L'individuazione di tale stato è un punto cruciale per permettere l'azione tempestiva del personale medico, dato l'alto tasso di mortalità. Le decisioni cliniche possono essere facilitate dallo sviluppo tecnologico-informatico nel campo dell'intelligenza artificiale con metodi di machine learning.

Obiettivi: In questo studio è proposto lo sviluppo di modelli di predizione di mortalità per pazienti diagnosticati con sepsi o SIRS in Unità di Terapia Intensiva. Vengono inoltre effettuati dei test per verificare la possibile presenza di bias socio-economico nelle predizioni, per constatare se il modello predittivo differenzia determinate categorie rispetto ad altre.

Metodi: I dati analizzati vengono estratti dal database MIMIC-IV che è accessibile pubblicamente. Lo studio ha incluso 26.400 pazienti, l'estrazione dei quali è stata effettuata considerando i criteri Sepsis-3 e SIRS. I metodi di Machine Learning (ML) utilizzati per lo sviluppo dei modelli di predizione della mortalità in due istanti di tempo (a 90 giorni dall'ammissione in UTI e a fine ospedalizzazione) sono Logistic Regression (LR) e metodi di Gradient Boosting (GB) come XGBoost e LGBoost. Dalle predizioni effettuate tramite ML, sono stati ricavati dei risultati e delle analisi SHAP. In seguito è stata effettuata un'analisi per determinare il bias socio-economico attraverso la forzatura di variabili e la valutazione dei modelli.

Risultati: I risultati per d90d mostrano che l'accuratezza più elevata (0,82) è stata raggiunta dal modello LightGBM (LGB), che ha anche ottenuto un AUC-ROC di 0,84. Il modello XGBoost (XGB) ha ottenuto la sensibilità più elevata (0,70) e il miglior punteggio MCC (0,46). L'F1-score più alto per d90d è stato raggiunto da XGB (0,59), mentre la precisione più elevata è stata ottenuta dal modello LGB (0,65). Il valore di recall massimo per d90d è stato ottenuto da XGB (0,70). Per dinhosp, il modello LGB ha ottenuto la massima accuratezza (0,76) e MCC (0,48), mentre il modello XGB ha ottenuto la massima sensibilità (0,84). L'F1-score più alto per dinhosp è stato ottenuto da XGB (0,68), mentre la precisione più elevata è stata registrata dal modello LGB (0,70). Il valore di recall più

alto per dinhosp è stato ottenuto da XGB (0,84). In entrambi i set di dati, i modelli di regressione logistica hanno mostrato una sensibilità inferiore rispetto ai metodi ad albero, con F1-score e recall più bassi e una ridotta capacità di identificare correttamente la classe positiva.

Conclusioni: XGBoost è il modello che ha portato i risultati migliori in quanto a sensibilità e F1 score, da questo deriva un miglior modello diagnostico, il cui scopo è di diminuire il più possibile i casi di falsi negativi. Il modello è privo di bias socio-economico anche dopo la forzatura delle variabili, sebbene permangano minime differenze nella valutazione clinica dovute alla tipologia di assicurazione dei pazienti e possibili errori di misura da parte ospedaliera.

Parole chiave: Unità di terapia intensiva; Sepsi; Sindrome da Risposta Infiammatoria Sistemica (SIRS); Machine Learning; Bias socio-economico; Modelli di predizione di mortalità.

Indice

Abstract **i**

Indice **iii**

Introduzione **1**

1 Contesto clinico **3**
1.1 Unità di Terapia Intensiva . . . 3
1.2 Sindrome da Risposta Infiammatoria Sistemica (SIRS) . . . 3
1.3 Sepsi e Shock Settico . . . 4
1.3.1 Definizione di Sepsi e Shock Settico . . . 4
1.3.2 Epidemiologia e Fattori di rischio . . . 6
1.3.3 Impatto clinico ed economico . . . 7

2 Stato dell'Arte **9**
2.1 Modelli di predizione di mortalità in unità di terapia intensiva . . . 9
2.2 Obiettivi . . . 11

3 Materiali e metodi **13**
3.1 Database MIMIC-IV . . . 13
3.2 Selezione Dati . . . 16
3.2.1 Identificazione dei Pazienti . . . 16
3.2.2 Estrazione delle Variabili Cliniche Rilevanti . . . 18
3.2.3 Calcolo dei Punteggi Diagnostici: SIRS e SOFA . . . 18
3.3 Criteri di Inclusione ed Esclusione . . . 19
3.3.1 Selezione Coorte . . . 20
3.4 Preprocessing . . . 21
3.4.1 Divisione in Train/Test . . . 21
3.4.2 K-Nearest Neighbors (KNN) . . . 21

3.4.3 Trattamento Outlier . . . 22
3.4.4 Mappatura delle Variabili Categoriali . . . 22
3.5 Algoritmi di Machine Learning . . . 23
3.5.1 Logistic Regression . . . 24
3.5.2 XGBoost . . . 25
3.5.3 LightGBM . . . 26
3.6 Ottimizzazione iperparametri . . . 27
3.6.1 Randomized SearchCV . . . 27
3.6.2 Bayesian Optimization con Optuna . . . 27
3.6.3 Grid Search Raffinato . . . 28
3.6.4 Analisi mediante SHAPLEY . . . 29
3.7 Analisi del Bias . . . 31
3.7.1 Analisi dei Missing Values . . . 31
3.7.2 Forzatura delle Variabili Demografiche . . . 31

4 Risultati **33**
4.1 Risultati a 90 giorni . . . 34
4.1.1 Logistic Regression . . . 34
4.1.2 XGBoost . . . 35
4.1.3 LGBM . . . 36
4.2 Risultati a fine ospedalizzazione . . . 37
4.2.1 Logistic Regression . . . 37
4.2.2 XGBoost . . . 38
4.2.3 LGBM . . . 39
4.3 Risultati analisi SHAPLEY . . . 40
4.3.1 Analisi XGB per mortalità a 90 giorni . . . 40
4.3.2 Analisi XGB per mortalità in ospedale . . . 42
4.4 Risultati al Bias-Socio Economico . . . 44
4.5 Sintesi dei risultati . . . 48

5 Discussione **51**
5.1 Modelli di predizione . . . 51
5.2 Bias socio-economico . . . 54
5.3 Analisi SHAP e confronto con i modelli di riferimento . . . 56
5.4 Innovazioni del Progetto . . . 58
5.4.1 Ottimizzazione degli Iperparametri . . . 58
5.4.2 Analisi del Bias Socio-economico . . . 58
5.4.3 Utilizzo del Database MIMIC-IV . . . 59

5.4.4 Ampliamento della Coorte 59

6 Conclusioni e Sviluppi Futuri **61**

Bibliografia **65**

A Appendice A **67**

A.1 Strumenti e Pacchetti Utilizzati nel Progetto 67

A.2 Confronto delle Coorti 68

A.3 Distribuzioni stimate 69

A.4 Griglie di iperparametri 70

Elenco delle figure **71**

Elenco delle tabelle **73**

Introduzione

La Sepsi è una disfunzione di organi, potenzialmente letale, causata da una risposta anomala dell'ospite a un'infezione e rappresenta una delle principali cause di morte nell'Unità di Terapia Intensiva con un'incidenza variabile tra il 14% e il 30% [1]. La Sepsi può degenerare in uno stato di Shock Settico, caratterizzato da ipotensione grave e ipoperfusione, che provocano un eccesso di lattato sierico e la conseguente condizione di respirazione anaerobica. Questi fattori portano ad un rischio di mortalità ancora più elevato della Sepsi (39%). La gravità dello stato del paziente viene misurata tramite il Sequential Organ Failure Assessment (punteggio SOFA), dato dalla sommatoria di un insieme di valori monitorati durante il ricovero del paziente. Se questi valori fuoriescono dall'intervallo di misura, incrementano il punteggio SOFA e se questo risulta essere maggiore di 2, si identifica lo stato del paziente come settico [1].

La condizione di Sepsi è spesso studiata in relazione a ricerche riguardanti metodi per la diagnosi e costruzione di modelli predittivi di mortalità. Questo è dovuto, oltre al tasso di mortalità elevato di questa condizione, agli alti costi di gestione del paziente settico, che risultano pari a 3.8 miliardi di dollari per gli Stati Uniti [2]. Il nostro lavoro intende portare al miglioramento dei modelli predittivi di mortalità già pubblicati, considerando anche l'impatto del bias socio-economico sul modello stesso, in modo da verificarne la robustezza. La scelta del metodo si basa sulla systematic review di Bao et al., che compara i metodi di machine learning per predire mortalità di pazienti settici; il lavoro di Wang et al. e Mollura et al. sono utilizzati come baseline per il confronto dei risultati del lavoro, possedendo i migliori risultati in ambito e un obiettivo simile a quello del seguente studio ([3–5]). Non esistono al momento pubblicazioni che affrontino l'argomento del bias socio-economico come fattore potenzialmente rilevante in un modello di predizione di mortalità nei pazienti settici del database MIMIC-IV: ciò rende questo progetto innovativo e un primo passo per la discussione di questo tema.

1 | Contesto clinico

1.1. Unità di Terapia Intensiva

La Medicina di Terapia Intensiva è quella branca della medicina specializzata nella cura di pazienti in stato critico o grave. Questi pazienti infatti hanno un alto rischio di sviluppare condizioni che minacciano la loro vita o stanno già vivendo tali condizioni [6]. L'unità di terapia intensiva è un blocco ospedaliero che ospita pazienti in condizioni patologiche gravi che necessitano di pronta assistenza e cure mediche adeguate. In questa unità è necessario il continuo monitoraggio dei parametri vitali per permettere ai medici anestesisti-rianimatori di seguire l'andamento clinico e, se possibile, predire gli eventi critici permettendo così un intervento tempestivo. La strumentazione messa a supporto dei pazienti monitora valori come la pressione arteriosa e altre variabili per verificare in che condizioni di salute essi si trovino, momento per momento. Una delle maggiori criticità per i pazienti ricoverati sono le Infezioni Correlate all'Assistenza, che in UTI facilmente degenerano in setticemia o sepsi, e quindi sono molto diffuse. Il medico in UTI difficilmente riesce a sfruttare a pieno le numerose variabili strumentali registrate e ciò potrebbe portarlo a decisioni non ottimali durante il trattamento per quello specifico paziente. Negli ultimi anni quindi molti ricercatori hanno proposto l'applicazione di tecniche di machine learning per migliorare l'assistenza in Terapia Intensiva sfruttando nella sua interezza la grande quantità di dati generati: tecniche che possono migliorare la qualità dell'assistenza e gli esiti clinici dei pazienti, ottimizzando le procedure eseguite all'interno dell'unità di terapia intensiva.

1.2. Sindrome da Risposta Infiammatoria Sistemica (SIRS)

La Sindrome da Risposta Infiammatoria Sistemica (SIRS) è uno stato infiammatorio dell'organismo in risposta a un fattore di stress nocivo, risposta che, quando eccessiva, comporta un'emergenza medica potenzialmente mortale. Il fattore nocivo può essere di diversa natura come trauma, ustione, pancreatite o altro, mentre se associato ad un'infezione può dare origine a una condizione di Sepsi ([1]). È dunque al fine di una

migliore descrizione e comprensione della Sepsi che risulta necessario definire in maniera approfondita cosa sia la SIRS. Non vi sono criteri ben definiti per il riconoscimento della SIRS, come in realtà anche della Sepsi, ma si prendono come indicatori per la diagnosi: la temperatura corporea, la frequenza cardiaca, la frequenza respiratoria e la conta dei globuli bianchi. È necessario sottolineare, però, come questi criteri non indichino necessariamente una risposta sregolata e potenzialmente letale, essendo talvolta presenti in pazienti non settici e che non presentano alcuna infezione ([1]).
I criteri di diagnosi della SIRS sono i seguenti: la temperatura $> 38°C$ o $< 36°C$, frequenza cardiaca > 90/min, frequenza respiratoria > 20/min o $PaCO_2 < 32$ mm Hg (4.3 kPa) e la conta dei globuli bianchi $> 12\,000/mm^3$ o $< 4\,000/mm^3$ o $> 10\%$ bande immature.

1.3. Sepsi e Shock Settico

1.3.1. Definizione di Sepsi e Shock Settico

La Sepsi è una delle principali cause di morte nell'Unità di Terapia Intensiva (UTI) nel mondo ed è stata definita nel 2016 dal 'Third International Consensus Definition for Sepsis and Septic Shock (Sepsi-3)', istituito dall'European Society of Intensive Care Medicine, come una disfunzione di organi pericolosa per la vita, causata da una risposta disregolata dell'ospite all'infezione: l'insufficienza acuta di più organi (polmoni, reni, fegato, ecc.) si manifesta con ipotensione, febbre alta e stato confusionale. ([1, 2]). La Sepsi può degenerare in una condizione di Shock Settico, ossia uno stato caratterizzato da ipotensione grave e ipoperfusione cellulare, con conseguente aumento di lattato sierico, prodotto della condizione di respirazione anaerobica. Si può definire lo Shock Settico come una condizione clinica in cui, nonostante l'adeguata somministrazione di fluidi, è necessaria la ulteriore somministrazione di vasopressori così da mantenere la pressione arteriosa media ≥ 65 mmHg e il livello sierico di lattato rimane > 18 mg/dL [2 mmol/L]. La condizione di Sepsi viene misurata per mezzo del Sequential Organ Failure Assessment (punteggio SOFA) a figura 1.2 [7] in base a diversi parametri, quali la conta delle piastrine, il rapporto tra pressione e frazione d'ossigeno inspirata, bilirubina sierica, pressione e vasopressori e creatinina. Per convenzione si determina uno stato di Sepsi nel caso in cui il punteggio SOFA risulti maggiore di 2 ([7]). Nella pratica clinica il paziente con sospetto di infezione viene dapprima valutato tramite il qSOFA che è uno strumento di screening rapido per identificare pazienti a rischio di sepsi tramite l'analisi di soli 3 valori: frequenza respiratoria, GCS (Glasgow Coma Scale) e pressione arteriosa sistolica come da diagramma in figura 1.1.

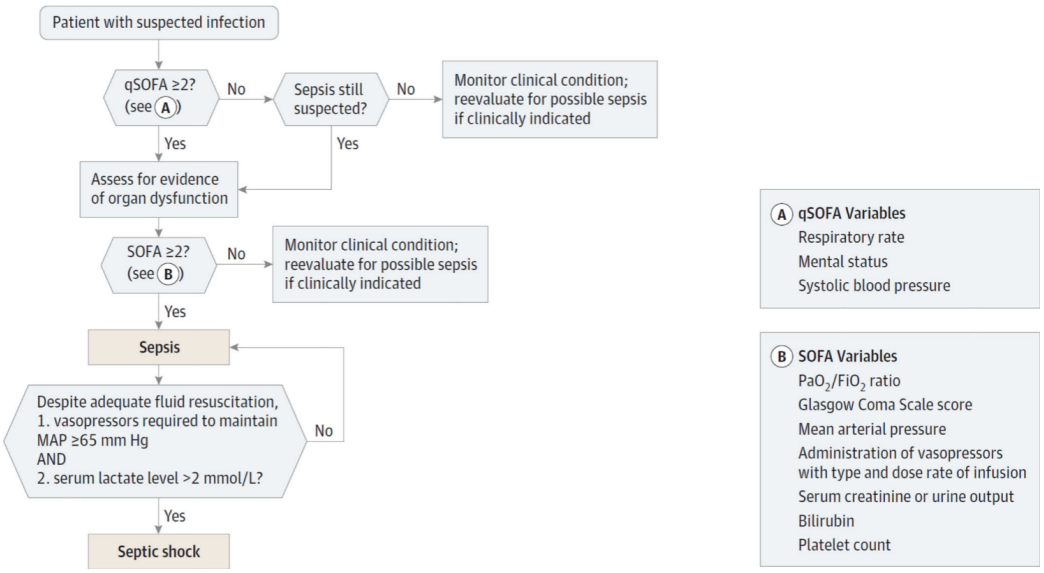


Figura 1.1: Flowchart sepsi

Se il sospetto è fondato si conferma attraverso il punteggio SOFA in base ai parametri riportati alla figura 1.2 [1].

	SOFA score 0	SOFA score 1	SOFA score 2	SOFA score 3	SOFA score 4
Respiratory system: PaO_2/FiO_2 (kPa)	≥53·3	<53·3	<40	<26·7	<13·3
Coagulation system: platelets (×10^3/μL)	≥150	<150	<100	<50	<20
Hepatic system: bilirubin (μmol/L)	<20	20–32	33–101	102–204	>204
Cardiovascular system	MAP >70 mm Hg	MAP <70 mm Hg	Dopamine <5 μg/kg per min, or dobutamine (any dose) administered	Dopamine 5·1–15 μg/kg per min, or epinephrine ≤0·1 μg/kg per min, or norepinephrine ≤0·1 μg/kg per min administered	Dopamine >15 μg/kg per min, or epinephrine >0·1 μg/kg per min, or norepinephrine >0·1 μg/kg per min administered
Central nervous system: Glasgow Coma Scale	15	13–14	10–12	6–9	<6
Renal system					
Creatinine (μmol/L)	<110	111–170	171–299	300–440	>440
Urine output (mL/day)	..	..	..	<500	<200

Scores from 0 to 4 are assigned for each of the six organ systems, with a higher score indicative of worse organ dysfunction in each system. MAP=mean arterial pressure.

Table: **Description of sequential organ failure assessment (SOFA) scoring system**[5]

Figura 1.2: Punteggio SOFA

La differenza principale tra punteggio SOFA e qSOFA risiede nella diversa accuratezza della diagnosi. Nonostante il qSOFA sia un metodo rapido e si basi sull'acquisizione di pochi dati clinici, la sua sensibilità risulta minima e non basta per effettuare una vera e propria diagnosi. Questo punteggio, difatti, va spesso ad identificare la condizione di Sepsi in soggetti che si trovano già in un stato avanzato della condizione ([8]).

1.3.2. Epidemiologia e Fattori di rischio

L'incidenza della sepsi e dello shock settico varia tra continenti e paesi, contribuendo a rendere queste condizioni un problema clinico a livello globale. Sono, difatti, rilevati all'incirca 677,5 casi ogni 100.000 persone ([9]), con i tassi più alti registrati nell'Africa subsahariana, in Oceania e nell'Asia meridionale. In Nord America, si verificano tra 500 e 1000 casi di sepsi ogni 100.000 persone, mentre è stimato che in Europa, l'incidenza sia tra 400 e 800 casi ogni 100.000 persone. Al contrario, nei paesi a basso e medio reddito, la sepsi è più comune, con tassi che, in alcune nazioni asiatiche e africane, superano i 1500 casi ogni 100.000 persone. Anche i tassi di mortalità legati alla sepsi e allo shock settico variano considerevolmente tra le diverse aree geografiche. Nei paesi ad alto reddito, la mortalità della Sepsi varia dal 15% al 25%, mentre la mortalità dello shock settico può raggiungere il 30%-40%. Nei paesi a basso e medio reddito, i tassi di mortalità sono notevolmente più elevati, arrivando a superare il 40% di decessi per la sepsi e il 50% per lo shock settico. Le percentuali così elevate delle due condizioni nei paesi a basso e medio reddito è dovuta ad una più alta prevalenza di malattie infettive, accesso limitato ai servizi sanitari, scarsa igiene e condizioni di vita inadeguate.

Si definiscono di seguito tutti quei fattori che svolgono un ruolo significativo nell'aumentare l'incidenza di sepsi e shock settico: i fattori demografici, come l'età e il sesso, difatti l'età avanzata è un fattore di rischio ben noto, con tassi di incidenza e mortalità che aumentano notevolmente nelle popolazioni più anziane. L'invecchiamento della popolazione in molti paesi, in particolare nei paesi ad alto reddito, contribuisce al crescente carico della sepsi. Ci sono, poi, le malattie croniche: pazienti con malattie croniche sono più soggetti alle infezioni e presentano un rischio maggiore di sviluppare la sepsi. Un altro fattore che ha un ruolo significativo è il sito dell'infezione, il quale può influenzare in modo significativo l'esito della condizione settica. Ne sono un esempio le infezioni respiratorie, come la polmonite, che sono la causa più frequente di sepsi e sono associate a tassi di mortalità più elevati rispetto ad altri siti di infezione. Anche le infezioni addominali, come la peritonite e gli ascessi intra-addominali, sono associate a un tasso di mortalità elevato richiedendo spesso un intervento chirurgico oltre alla sola terapia antimicrobica. Ci sono, infine i traumi, ustioni, altre lesioni acute e i fattori ambientali e socioeconomici: la povertà, la scarsa igiene e l'accesso limitato ad acqua potabile e ai servizi sanitari contribuiscono all'elevata incidenza di sepsi nei paesi a basso e medio reddito. Anche il sovraffollamento, la malnutrizione e l'esposizione ad aria inquinata aumentano il rischio di infezioni e di contrarre la sepsi. Ne segue che anche le condizioni di vita, implicano un decorso più rapido e una facilità nella contrazione della condizione.

1.3.3. Impatto clinico ed economico

Nel 2020 sono stati registrati 49 milioni di casi di sepsi ed è questa condizione che ha causato il 19.7% dei decessi a livello globale. A livello prettamente ospedaliero, un paziente su tre che decede durante il ricovero ha contratto la sepsi. La cura dei pazienti settici rappresenta negli Stati Uniti una spesa annuale di più di 38 miliardi di dollari, diventando di conseguenza la patologia che rappresenta la più alta porzione della spesa sanitaria. Cruciale è la tempestività delle cure; si stima infatti una diminuzione del 7,6% della probabilità di sopravvivere del paziente per ogni ora in cui viene ritardata la somministrazione di farmaci e antibiotici. Per la condizione di shock settico invece è associato un rischio di mortalità ancora più elevato della sepsi stessa e risulta circa il 39% ([1, 2]). Sebbene la definizione di sepsi e la sua gestione clinica siano state sottoposte a continue revisioni, nel mondo clinico c'è accordo solo su alcuni punti chiave nella gestione della sepsi, come il tipo di vasopressori o il tipo di fluidi da somministrare, mentre è ancora presente un'ampia controversia su molti altri, in particolare sulle variabili cliniche chiave su cui concentrarsi ([10]).

2 | Stato dell'Arte

2.1. Modelli di predizione di mortalità in unità di terapia intensiva

Come punto di partenza per lo sviluppo dei modelli di predizione sono stati analizzati vari studi. Nella ricerca di un articolo che fornisca i migliori algoritmi per la predizione,"Bao et al."[3] mettono a disposizione un confronto tra tutti i metodi di Machine Learning per la predizione della sepsi, evidenziando i punti di forza e i punti critici di ogni algoritmo. Lo studio analizza l'affidabilità di diversi modelli su una coorte di circa 13.000 pazienti settici, i cui dati sono stati presi dal database MIMIC-IV utilizzato anche in questo progetto. I modelli utilizzati sono i seguenti:

- Support Vector Machine (SVM)
- Decision Tree Classifier (DTC)
- Random Forest (RF)
- Gradients Boosting (GBM)
- Multiple Layer Perception (MLP)
- XGBoost (XGB)
- Light Gradients Boosting Machines (LGBM)

I vari modelli ML vengono classificati sulla base di parametri statistici come l'AUROC e, i risultati, mostrano come migliori Light GBM e XGBoost: l'AUROC di LGBM Classifier è il maggiore (AUROC = 0.86 $\pm$ 0.12) e XGBoost è il secondo migliore (AUROC = 0.84 $\pm$ 0.12). Molti degli studi pubblicati, riguardanti le unità di terapia intensiva, infatti, sfruttano questi due algoritmi per la predizione di mortalità. Si è scelto quindi di affrontare lo studio utilizzando i migliori due modelli presentati nell'articolo, oltre che ovviamente il modello Logistic Regression: modello che, nonostante presenti una accuratezza inferiore, ha maggiore interpretabilità a livello clinico in quanto è un algoritmo che studia le variabili

tramite classificazione binaria lineare.
Nella letteratura sono numerosi gli articoli che sfruttano modelli di Machine Learning per la predizione di mortalità di pazienti settici nelle UTI. Ne è un esempio l'articolo di "Fei Guo et al." [11], in cui, su una coorte di circa 2500 pazienti, si espone come la combinazione di metodi di machine learning e deep learning permetta di predire la mortalità, oltre che a caratterizzare il fenotipo di ciascun soggetto settico associato alla sua sopravvivenza, valutando, quindi, in che modo la coagulazione del sangue influisce sul decorso della condizione.
La systematic review di "L.M. Fleuren et al." [12] esplora, invece, i diversi metodi diagnostici che sfruttano modelli di machine learning per la predizione della sepsi. Sono stati raccolti un totale di 5280 articoli, di cui, però, solo 28 rispettavano i criteri di eleggibilità selezionati dagli autori. Tra i 28 studi, quello di "Brown S.M. et al." viene definito come l'unico in grado di validare il modello presentato.
Si presenta poi l'articolo di "Zhao et al." [12]. Lo studio si propone di sviluppare un modello di predizione di mortalità basato sull'utilizzo di Logistic Regression, RF, XGBoost e Artificial Neural Network (ANN). In questo studio i dati clinici sono stati estratti dal database eICU-CRD e da questi si é ricavata una coorte di 123929 pazienti. Al termine dello studio si conclude come XGBoost sia il miglior metodo utilizzabile, tra quelli elencati, per la predizione di mortalità con un AUROC di 0.9702 contro 0.9620 di ANN, 0.9559 di RF e 0.9357 di Logistic Regression.
Due, invece, sono gli esempi di studi condotti su pazienti settici, utili per lo sviluppo di questo progetto, ossia gli articoli di "Wang et al."[4] e di "Mollura et al."[5]. Entrambi gli studi esplorano l'uso del Machine Learning per la predizione della mortalità nei pazienti settici in ICU, con un focus particolare sull'individuazione delle variabili critiche. Il primo dei due ha sviluppato un modello di previsione della mortalità precoce da sepsi utilizzando un approccio di machine learning. I dati provengono dal Database MIMIC-III e la coorte estratta è di circa 27.000 pazienti applicando i criteri Sepsis-3. L'obiettivo della ricerca era quello di implementare un modello accurato ma soprattutto interpretabile a livello clinico. Sono state considerate 47 variabili cliniche e gli algoritmi di Machine Learning utilizzati sono Logistic Regression, Support Vector Machine, Deep Neural Network e XGBoost. Tramite SHAP è stato possibile ridurre il numero delle variabili critiche incidenti, ottenendo comunque un buon AUROC di 0.873 per XGBoost e di 0.829 per Logistic Regression. Lo studio ha dimostrato che è possibile basarsi su un numero limitato di variabili senza compromettere l'accuratezza predittiva del modello. Inoltre, l'uso dell'analisi SHAP ha fornito interpretabilità, facilitando così l'adozione di questi modelli da parte del personale clinico.
Il secondo studio ha come obiettivo principale l'identificazione dei parametri clinici che

influenzano maggiormente la sopravvivenza di pazienti affetti da SIRS o sepsi. La coorte su cui è stato condotto lo studio è molto ridotta (circa 1250 pazienti) e si basa su cartelle cliniche provenienti da ospedali in Turchia. L'individuazione delle variabili critiche ha fatto sì che, delle 30 iniziali, solo 5 fossero considerate le più influenti ai fini della predizione di mortalità:

- Punteggio SOFA
- Frequenza cardiaca
- Pressione arteriosa media
- Livello di lattato
- Conta dei globuli bianchi

Il modello mantiene una buona accuratezza predittiva, con un MCC score di 0.53 per la Regressione Logistica e 0.49 per XGBoost, dimostrando che anche con un numero limitato di variabili, si possono ottenere risultati utili per la pratica clinica. Entrambi gli articoli, dunque, dimostrano che è possibile ridurre significativamente il numero di variabili necessarie per predire accuratamente la mortalità nei pazienti settici e migliorare la diagnosi clinica, ottimizzando la gestione di risorse ospedaliere anche con la riduzione dei costi [5].

2.2. Obiettivi

Il progetto utilizza dei modelli ML per la predizione di mortalità in UTI, basandosi sulle metodologie descritte nei lavori di Wang et al. ([6]) e di Mollura et al.([4]).
Gli obiettivi che ci si pone rispetto a questi articoli sono:

- Applicare i metodi di Wang sui dati MIMIC-IV e verificare i risultati
- Espandere il dataset di Wang includendo pazienti diagnosticati sia con la sepsi sia con la SIRS, migliorando la generalizzabilità del modello
- Sviluppare un modello predittivo per la morte in più lassi di tempo: da ingresso a 90 giorni e fino a fine ospedalizzazione
- Verificare la possibile presenza di un bias nella classificazione del rischio dovuto allo stato socio-economico dei pazienti nel database.

3 | Materiali e metodi

3.1. Database MIMIC-IV

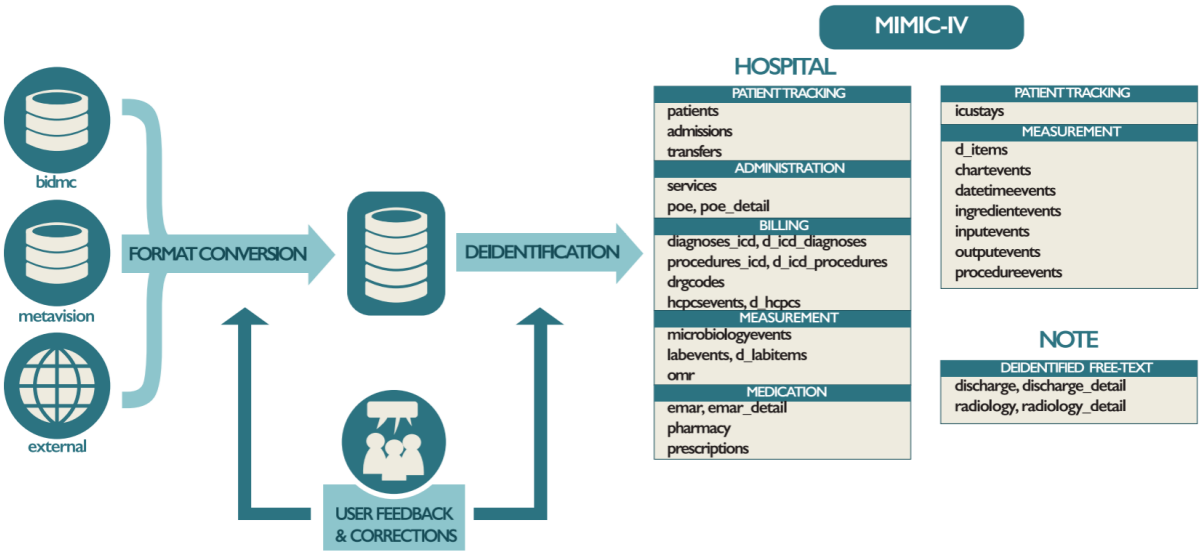


Figura 3.1: Processo di sviluppo per MIMIC

Le cartelle cliniche elettroniche (CCE) sono strumenti fondamentali per gestire le informazioni sanitarie, ma non sono esenti da alcune limitazioni quando si tratta di ricerca medica. I dati contenuti nelle CCE offrono un enorme potenziale: possono infatti contribuire a migliorare l'assistenza ai pazienti e a sviluppare algoritmi utili per il supporto alle decisioni cliniche, oltre a favorire la "Knowledge Discovery", ovvero l'estrazione di informazioni preziose da grandi moli di dati. Tuttavia, l'accesso alle CCE è spesso limitato e controllato, per motivi di privacy e per la natura sensibile dei dati sanitari, rendendo difficoltoso per i ricercatori l'accesso. Di conseguenza, il potenziale delle CCE rimane spesso non sfruttato. Per assolvere a questo problema, negli anni recenti sono state sviluppate diverse basi di dati cliniche pubbliche. Una di queste è il MIMIC-IV, una base di dati contente dati di oltre 40,000 pazienti ammessi nelle UTI del Beth Israel Deaconess Medical Center (BIDMC), Boston, MA. Le informazioni disponibili includono dati clinici dei pazienti,

diagnosi, procedure, trattamenti e note cliniche deidentificate [13]. Per lo sviluppo di questa base di dati, le informazioni cliniche sono state ricavate dall'archivio dati del BIDMC, dal sistema informativo per le UTI (Metavision) e da sorgenti esterne unificate tramite SQL. Successivamente sono applicati degli algoritmi di deidentificazione.

L'accesso al database è garantito da PhysioNet [14].I dati sono raggruppati in tre moduli:

- Modulo Hosp
- Modulo ICU
- Modulo Note

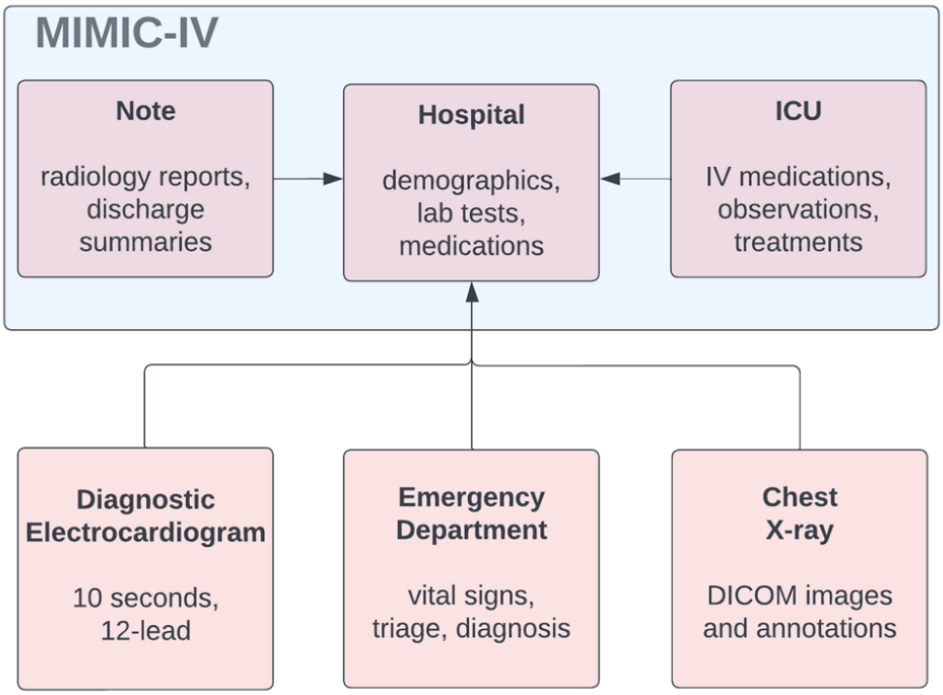


Figura 3.2: Struttura modulare del database

Il modulo **Hosp** contiene tutte le informazioni riguardanti diagnosi cliniche, referti dal laboratorio, misure microbiologiche, trasferimento dei pazienti e gestione delle terapie mediche. Viene inoltre inserita una colonna 'subjectid' che permette di collegare il paziente alle sue demografiche nella tabella "patients". La colonna 'hadmid' rappresenta, invece, la singola ospedalizzazione.
Il modulo **ICU** contiene le sole informazioni dei pazienti ricoverati in terapia intensiva e include tabelle come "chartevents", "ditems", "datetimeevents" "icustays" e altre relative ai vari eventi.
Il modulo **Note**, infine, contiene i referti medici organizzati in dimissioni e radiologia.

Il numero di pazienti ricoverati in unità di terapia intensiva e registrati nel database MIMIC-IV è riportato nella tabella seguente:

	Ammissioni ospedaliere	**Ammissioni in UTI**
Numero di permanenze	431,231	73,181
Pazienti unici	180,733	50,920
Età, media (SD)	58.8 (19.2)	64.7 (16.9)
Pazienti donne, n (%)	224,990 (52.2)	32,363 (44.2)
Assicurazioni, n (%)		
Medicaid	41,330 (9.6)	5,528 (7.6)
Medicare	160,560 (37.2)	33,091 (45.2)
Altre	229,341 (53.2)	34,562 (47.2)
Durata in ospedale, media (SD)	4.5 (6.6)	11.0 (13.3)
Mortalità in ospedale, n (%)	8,974 (2.1)	8,519 (11.6)
Mortalità dopo un anno, n (%)	106,218 (24.6)	28,274 (38.6)

Tabella 3.1: Statistiche per ospedale e ammissione in UTI

3.2. Selezione Dati

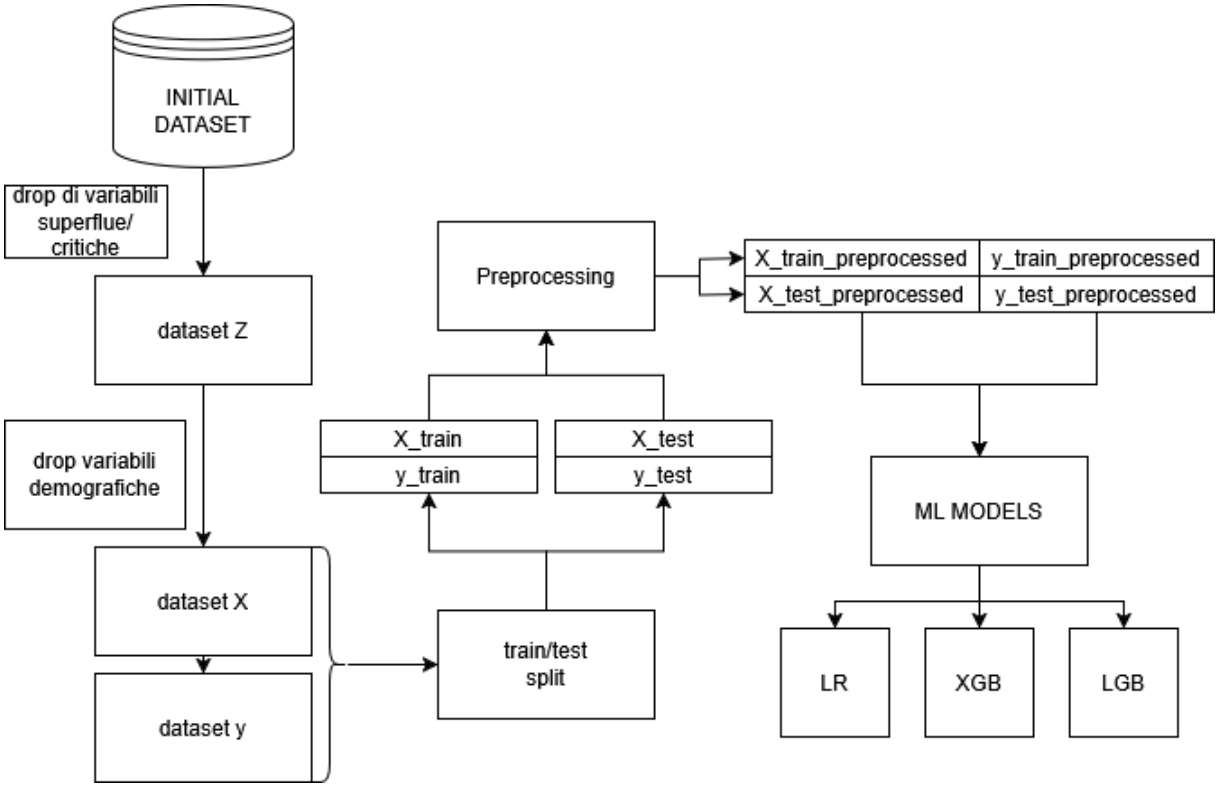


Figura 3.3: Struttura dello studio

Questo studio si basa sul seguente processo: dopo una prima estrazione dei dati da MIMIC IV attraverso SQL si estraggono le variabili diagnostiche e vitali portando alla generazione del dataset iniziale. Una volta creato questo dataset si ha una filtrazione ulteriore di pazienti e variabili superflue, oltre che delle colonne di predizione. Questo nuovo dataset è stato chiamato Z. Una volta rimosse le colonne rappresentanti variabili demografiche (etnia, lingua, tipo di assicurazione), portando alla generazione del dataset X. A seguire, il dataset X è diviso in train e test con una divisione 80/20 ed è sottoposto a un passo di preprocessing. Una volta preprocessati i dati, essi sono utilizzati per allenare e valutare dei modelli ML. Questo processo è rappresentato nell'immagine 3.3.

3.2.1. Identificazione dei Pazienti

Nel database MIMIC-IV, i pazienti sono identificati tramite `subject_id` e `stay_id`, dove `subject_id` rappresenta il codice identificativo del paziente, mentre `stay_id` identifica un singolo ricovero in terapia intensiva (ICU). Il modulo `mimiciv_icu` e `mimiciv_hosp` contiene 180733 `subject_id` distinti. Lo stesso modulo presenta 73176 `stay_id` unici.

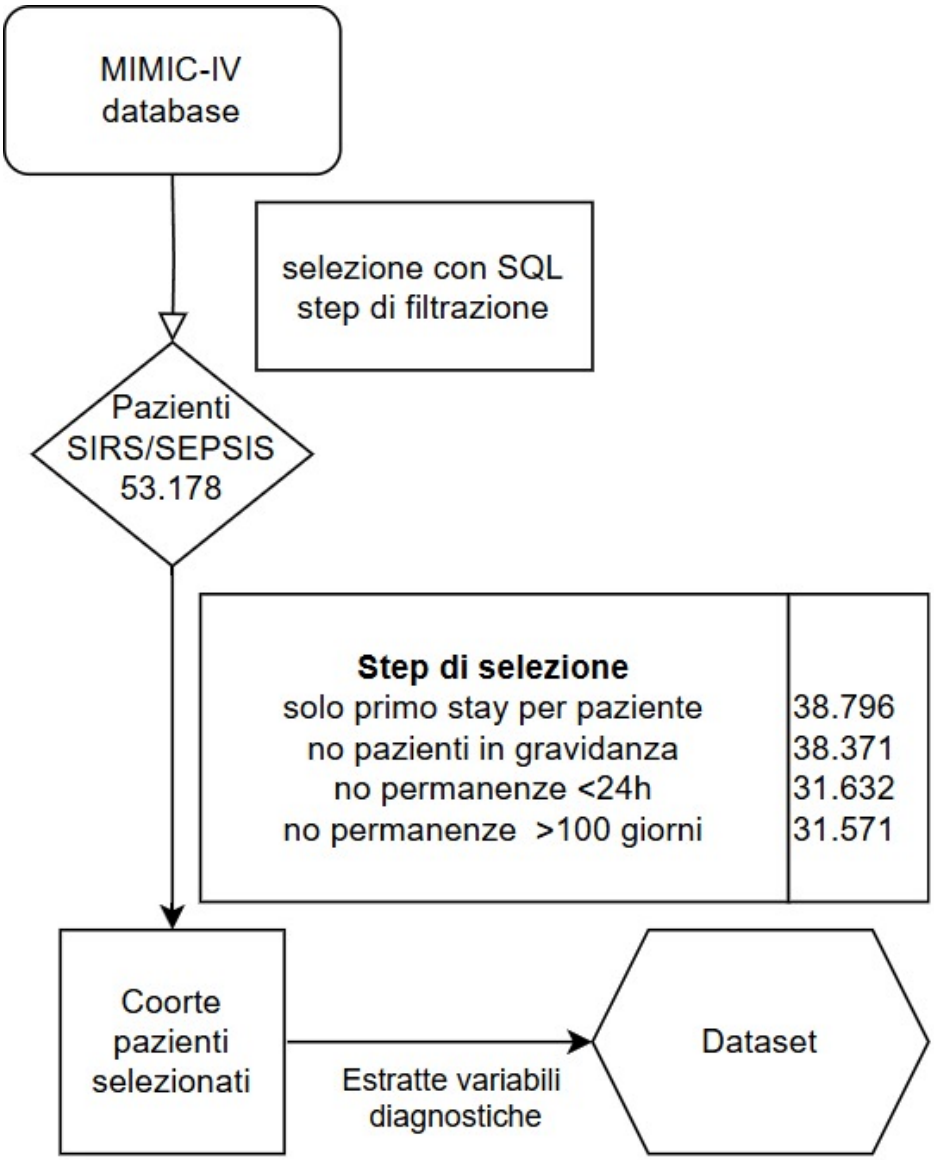


Figura 3.4: Diagramma dei criteri di inclusione

L'identificazione dei pazienti con Sepsi-3 e SIRS segue una pipeline strutturata in più fasi:

- Estrazione delle Variabili Cliniche Rilevanti
- Calcolo dei punteggi diagnostici
- Diagnosi di SIRS e Sepsis-3

3.2.2. Estrazione delle Variabili Cliniche Rilevanti

Le informazioni necessarie per la diagnosi provengono principalmente da due fonti: la tabella Chartevents (`mimiciv_icu.chartevents`), che raccoglie i segni vitali monitorati in terapia intensiva, e la tabella Labevents (`mimiciv_hosp.labevents`), che contiene i risultati degli esami di laboratorio. Le variabili selezionate comprendono i segni vitali, quali frequenza cardiaca, temperatura corporea, pressione arteriosa media e frequenza respiratoria, i valori ematochimici, quali conteggio dei globuli bianchi, piastrine, creatinina, bilirubina e PaO_2, e gli indicatori di risposta infiammatoria, quali $PaCO_2$ e neutrofili a banda. Solo le osservazioni con valori numerici non nulli sono incluse per garantire la qualità dei dati.

3.2.3. Calcolo dei Punteggi Diagnostici: SIRS e SOFA

Dopo l'estrazione, vengono calcolati i punteggi diagnostici per SIRS e Sepsi-3. Per quanto riguarda il punteggio SIRS, esso viene assegnato secondo i seguenti criteri, dove un punteggio totale ≥ 2 indica la presenza di SIRS:

Parametro	**Valore**
Temperatura corporea	< 36°C o > 38°C
Frequenza cardiaca	> 90 bpm
Frequenza respiratoria	> 20 atti/minuto o $PaCO_2 < 32$ mmHg
Conteggio globuli bianchi	< 4.0 o > 12.0 (10^3/uL) o neutrofili immaturi (bande) $> 10\%$

Tabella 3.2: Criteri diagnostici

Il punteggio SOFA, invece, viene assegnato valutando cinque sistemi fisiologici:

Parametro	**Valore**
Funzione respiratoria	$PaO_2/FiO_2 < 400$
Funzione cardiovascolare	MAP < 70 mmHg
Funzione epatica	Bilirubina ≥ 1.2 mg/dL
Coagulazione	Conta piastrinica $< 150 \times 10^3$/uL
Funzione renale	Creatinina ≥ 1.2 mg/dL

Tabella 3.3: Criteri di valutazione della funzione d'organo

Un SOFA score ≥ 2 indica una disfunzione d'organo significativa ed è quindi criterio per la diagnosi di Sepsi-3 e Shock Settico. Dopo l'estrazione iniziale dei dati, la coorte risultante comprende: 38.796 `subject_id` e 53.178 `stay_id`.

Categoria	Criteri
Sepsi-3	SOFA ≥ 2 e SIRS ≥ 2
SIRS (senza sepsi)	SIRS ≥ 2, ma SOFA < 2
Nessuna diagnosi	Non soddisfa i criteri sopra

Tabella 3.4: Criteri diagnostici per Sepsi-3 e SIRS

3.3. Criteri di Inclusione ed Esclusione

Per garantire un dataset affidabile e confrontabile con quelli ottenuti da Wang [4] e Mollura [5] vengono applicati criteri di esclusione specifici. Si conserva solo il primo ricovero in ICU per ciascun paziente, utilizzando la funzione ROW_NUMBER() basata sull'ordine di ingresso in ICU. Le pazienti in gravidanza vengono escluse attraverso l'identificazione dei codici ICD-10, ovvero O, e ICD-9, ovvero V22-V24 e Z33-Z36. Inoltre, si escludono i ricoveri con una durata anomala, eliminando quelli inferiori a 24 ore per evitare ricoveri transitori e quelli superiori a 100 giorni per rimuovere casi atipici. L'ultimo step di selezione è l'esclusione dei pazienti che presentano un numero di valori <20% rispetto al numero di variabili presenti nel dataset.

Tabella 3.5: Criteri di selezione dei pazienti per il dataset finale.

Criterio	**subject_id**	**stay_id**
Dati iniziali in MIMIC-IV	180.733	73.176
Applicazione criteri SIRS e Sepsis-3	38.796	53.178
Selezione solo prima ammissione per paziente	38.796	38.796
Esclusione donne in gravidanza	38.371	38.371
Esclusione ricoveri <24h o >100 giorni	31.571	31.571
Esclusione valori presenti < 20%	26.400	26.400

3.3.1. Selezione Coorte

La selezione della coorte è stata effettuata con l'obiettivo di garantire l'affidabilità dei modelli predittivi ed evitare problematiche di data leakage, oltre a escludere le variabili diagnostiche di bassa utilità. L'approccio adottato si allinea con le strategie degli studi di Wang et al. [4] e Mollura et al. [5], i quali hanno dimostrato che una selezione mirata delle variabili può migliorare le prestazioni del modello senza compromettere la robustezza predittiva. In particolare, è stata effettuata un'analisi dei dati mancanti, escludendo variabili con più del 90% di valori mancanti, e sono stati rimossi attributi ridondanti o non significativi dal punto di vista clinico e statistico. Per prevenire il rischio di fuga di informazioni, sono state escluse feature direttamente collegate all'outcome, come identificativi del paziente e variabili temporali post-evento. L'adozione di tecniche di selezione delle feature come la rimozione di variabili con basso peso sulla predizione e non utili dal punto di vista diagnostico, coerentemente con Wang et al. [4], ha permesso di mantenere solo le variabili con maggiore impatto predittivo, garantendo un AUROC competitivo. Allo stesso modo, Mollura et al. [5] hanno dimostrato che una drastica riduzione delle feature, unita all'uso di metodi interpretabili come SHAP, consente di migliorare l'equità e l'interpretabilità dei modelli. Di conseguenza, la coorte finale è stata costruita selezionando variabili clinicamente e statisticamente rilevanti, in modo da ottimizzare l'equilibrio tra accuratezza predittiva e interpretabilità, come visibile nella tabella A.1, la quale presenta in maniera schematica un confronto delle coorti del nostro studio e degli studi di referenza.

3.4. Preprocessing

3.4.1. Divisione in Train/Test

La prima fase del preprocessing è una divisione del dataset in un training set (80%) e un test set (20%), una divisione nota come train-test split 80/20 che rappresenta una prassi comune in ambito di machine learning. L'80% dei dati viene utilizzato per addestrare il modello, permettendogli di apprendere le relazioni presenti all'interno dei dati, mentre il restante 20% viene riservato per valutare le prestazioni del modello, garantendo una stima affidabile della sua capacità di generalizzazione su dati non visti. Tale strategia aiuta a minimizzare il rischio di overfitting, assicurando che il modello mantenga buone performance anche su nuovi campioni.

3.4.2. K-Nearest Neighbors (KNN)

Una volta diviso il dataset si passa all'imputazione dei valori mancanti, una fase chiave del preprocessing, essenziale per evitare che dati incompleti compromettano le prestazioni del modello. A questo scopo, è stato impiegato il metodo K-Nearest Neighbors (KNN) Imputer, basato sul principio secondo cui i valori mancanti possono essere stimati utilizzando i dati più simili presenti nel dataset. In pratica, per ogni valore mancante l'algoritmo individua i K punti più vicini nello spazio delle feature e calcola la loro media per sostituire il valore assente. La scelta del KNN Imputer si è basata sui risultati riportati dallo studio [15], che ne ha evidenziato l'efficacia nella gestione dei valori mancanti in dataset clinici. In particolare, è stato selezionato il valore di $n\_neighbors = 50$ per garantire una stima stabile e meno sensibile al rumore. È importante sottolineare che un valore di K troppo basso porta a un fenomeno di overfitting, in cui l'imputazione si basa su pochi vicini ed è pertanto molto sensibile a outlier e rumore (basso bias ma alta varianza), mentre un valore di K troppo elevato comporta underfitting, poiché l'imputazione si basa su un gran numero di vicini, riducendo la varianza ma aumentando il bias e perdendo la capacità di catturare le peculiarità della singola variabile.
Prima di applicare qualsiasi tecnica di preprocessing, il dataset è stato suddiviso in training set (X_{train}) e test set (X_{test}); questa separazione è fondamentale per garantire che l'imputazione dei valori mancanti venga appresa esclusivamente sui dati di training, evitando così la contaminazione del test set con informazioni future e prevenendo fenomeni di data leakage. Il processo di imputazione è stato realizzato seguendo tre passaggi principali: innanzitutto, il dataset è stato diviso in X_{train} e X_{test}; successivamente, il modello KNN è stato fittato esclusivamente su X_{train}, calcolando i valori mancanti in base

ai 50 vicini più prossimi; infine, l'imputer appreso su X_{train} è stato utilizzato per imputare i dati mancanti sia in X_{train} che in X_{test}, mantenendo la coerenza tra i due set e assicurando che il test set rimanga indipendente.

3.4.3. Trattamento Outlier

Si continua con un passo di trattamento outlier, i quali possono possono influenzare negativamente le performance dei modelli di machine learning, soprattutto in dataset clinici, dove la presenza di valori estremi può derivare da errori di misura o condizioni patologiche rare. Per il trattamento degli outlier è stata adottata la metodologia dell'*Interquartile Range (IQR)*. In particolare, per ciascuna variabile numerica sono stati calcolati il primo quartile ($Q1$) e il terzo quartile ($Q3$), definendo il range interquartile come

$$IQR = Q3 - Q1.$$

I valori che cadono al di fuori dell'intervallo

$$[Q1 - 1.5 \times IQR,\ Q3 + 1.5 \times IQR]$$

sono stati considerati outlier. Infine, gli outlier rilevati sono stati sostituiti con valori mancanti (NaN), in modo da poterli trattare successivamente con l'imputazione basata sul metodo K-Nearest Neighbors (KNN).

3.4.4. Mappatura delle Variabili Categoriali

Per rendere il dataset compatibile con gli algoritmi di machine learning, le variabili categoriali sono state convertite in formato numerico durante il passo di inizializzazione dati che avviene prima del preprocessing. Questa trasformazione è necessaria affinché i modelli possano elaborare correttamente le informazioni senza introdurre ambiguità.

Nel caso del dataset **X**, le seguenti variabili sono state mappate:

Tabella 3.6: Diagnosi

Valore	Condizione
1	Sepsi
0	SIRS

Tabella 3.7: Genere

Valore	Genere
1	M
0	F

Nel caso del dataset **Z**, oltre alle trasformazioni sopra descritte, sono state incluse ulteriori categorizzazioni per garantire una rappresentazione più dettagliata delle informazioni:

Tabella 3.8: Razza

Codice	Categoria
1	Hispanic
2	Black
3	Asian
4	White
0	altri

Tabella 3.9: Tipo di Assicurazione

Codice	Categoria
3	PRIVATE
2	MEDICARE
1	MEDICAID
0	altri

3.5. Algoritmi di Machine Learning

Durante la creazione del modello di previsione sono stati impiegati tre algoritmi di machine learning: Regressione Logistica (LR), XGBoost (XGB) e LightGBM (LGBM). Ogni metodo presenta caratteristiche specifiche che li rendono adatti all'analisi di dataset clinici ad alta dimensionalità.

- Logistic Regression (LR): modello lineare utilizzato per la classificazione binaria, particolarmente adatto per problemi in cui è importante l'interpretabilità del modello. LR calcola le probabilità di appartenenza alle classi utilizzando la funzione sigmoide ed è stato implementato in Python tramite la libreria `scikit-learn` [16]. In particolare LR viene utilizzato come riferimento per ottimizzare o valutare i modelli utilizzati ai passi successivi del processo.
- XGBoost (XGB): algoritmo di boosting basato su alberi decisionali ottimizzato per efficienza e accuratezza. XGB utilizza una combinazione di alberi deboli, il metodo di boosting e una gestione efficiente della memoria per migliorare le prestazioni rispetto ad altri algoritmi basati su alberi [17]. L'implementazione è stata effettuata tramite la libreria `xgboost` in Python.
- LightGBM (LGB): altra tecnica di boosting, simile a XGBoost ma ottimizzata per gestire dataset di grandi dimensioni in modo più veloce ed efficiente tramite una tecnica di crescita ad istogramma (*histogram-based learning*). LightGBM è particolarmente efficace quando si lavora con dataset molto sbilanciati e con numerose feature [16]. Il modello è stato implementato utilizzando la libreria `lightgbm` in Python.

Tutti i modelli sono stati addestrati e testati utilizzando `scikit-learn` per la gestione dei dati, `xgboost` e `lightgbm` per le rispettive implementazioni, con un'ottimizzazione degli iperparametri basata su `RandomSearchCV`, `Optuna` e `GridSearchCV`.

3.5.1. Logistic Regression

Dopo aver completato il preprocessing, il workflow prevede l'addestramento della regressione logistica (LR) adottando tre diversi approcci per la trasformazione dei dati: dati pre-processati (encoded), dati normalizzati e dati standardizzati.

LR sui Dati Encoded: In questo primo approccio, il modello di regressione logistica viene addestrato utilizzando i dati pre-processati, che hanno già subito operazioni di imputazione e trattamento degli outlier. Impiegando la funzione `LogisticRegression` di `scikit-learn` (con solver "saga" e `max_iter=10000`), il modello riesce a gestire anche dataset di grandi dimensioni. Questa configurazione rappresenta il punto di riferimento (baseline) contro il quale valutare l'impatto delle trasformazioni applicate nelle modalità successive.

LR sui Dati Normalizzati: La normalizzazione consiste nel ridimensionare le features in un intervallo definito (tipicamente $[0, 1]$). Questa trasformazione risulta particolarmente utile quando le variabili hanno scale molto differenti, poiché riduce la varianza tra le feature, stabilizzando il processo di ottimizzazione. Nello studio [18] viene dimostrato come il ridimensionamento dei dati possa accelerare la convergenza degli algoritmi e migliorare le performance complessive. Inoltre, viene sottolineato che la normalizzazione può portare a significativi incrementi in termini di accuratezza e AUC, specialmente in contesti in cui le feature presentano scale eterogenee come nel nostro caso. Nel nostro studio le variabili sono normalizzate attraverso distribuzioni ottenute attraverso il test Kolmogorov–Smirnov (kstest) [19], raccolte in appendice alla tabella A.2. Le distribuzioni delle variabili sono ottenute durante il passo di inizializzazione dati precedenti al preprocessing.

LR sui Dati Standardizzati: La standardizzazione trasforma i dati affinché ciascuna feature abbia una media pari a zero e una deviazione standard pari a uno. Questa tecnica non solo favorisce la convergenza degli algoritmi di ottimizzazione, ma rende anche più interpretabili i coefficienti stimati nei modelli lineari. Come evidenziato da Bishop et al. [20], la standardizzazione aiuta a mitigare l'effetto delle diverse scale delle variabili, migliorando la robustezza nella stima dei parametri. Diversi studi hanno confermato che l'addestramento della regressione logistica su dati standardizzati porta a risultati più coerenti e, in molti casi, a prestazioni superiori in termini di metriche come F1-score e MCC, in particolare quando si lavora con dataset caratterizzati da variabili eterogenee.

3.5.2. XGBoost

L'algoritmo *XGBoost* (eXtreme Gradient Boosting) si basa su un metodo di boosting che costruisce alberi decisionali in modo iterativo per correggere gli errori commessi nelle stime precedenti Un elemento distintivo di XGBoost è l'introduzione di un termine di regolarizzazione che ne migliora la generalizzazione, riducendo il rischio di overfitting [17].

Inoltre, diversi studi hanno evidenziato come *XGBoost* sia particolarmente efficace nella gestione di dataset altamente sbilanciati, tipici in ambito clinico per la previsione diagnostica e della mortalità. La ricerca di Bao et al. [3] ha dimostrato che l'uso di *XGBoost* permette di ottenere prestazioni superiori rispetto ad altri metodi nel predire eventi di mortalità e diagnosi in insiemi di dati caratterizzati da un forte squilibrio tra le classi. Tale ricerca conferma come l'approccio basato sul boosting e l'accurata ottimizzazione degli iperparametri possano fare la differenza in contesti clinici complessi.

Nella pipeline di codice il modello XGB viene implementato tramite la classe `XGBClassifier` della libreria XGBoost, che si basa sul boosting degli alberi decisionali. Il processo di modellizzazione si sviluppa in due fasi principali. In un primo momento l'addestramento del modello può essere eseguito in due modalità: in modalità standard il modello viene addestrato direttamente utilizzando i parametri specificati, ad esempio quelli definiti in `best_params_xgb` visibili nella tabella A.3a, mentre in modalità ottimizzazione, se il flag `run_xgb_optimization` è abilitato, si procede al processo di ottimizzazione degli iperparametri. Successivamente, il modello viene valutato sul set di test attraverso metriche quali accuracy, AUC-ROC, F1-score e Matthews Correlation Coefficient (MCC), e vengono generati diversi grafici diagnostici, come la curva ROC e i plot SHAP, che risultano utili per interpretare il contributo delle variabili predittive.

3.5.3. LightGBM

Il modello LightGBM è stato impiegato per la classificazione binaria finalizzata alla diagnosi di mortalità. Il processo si divide in due modalità operative, che differiscono per la gestione degli iperparametri e convergono in una valutazione del modello. Nella modalità di addestramento diretto il modello viene addestrato utilizzando un set di iperparametri predefiniti, specificati nel dizionario `best_params_lgb` osservabili nella tabella A.3b. Questi parametri includono configurazioni essenziali quali il numero massimo di foglie (`num_leaves`), la profondità massima degli alberi (`max_depth`), il tasso di apprendimento (`learning_rate`), la frazione delle feature utilizzate (`feature_fraction`) e altri parametri relativi alla regolarizzazione. L'addestramento diretto consente di ridurre i tempi di elaborazione, in quanto non viene effettuata una fase aggiuntiva di ottimizzazione degli gli iperparametri, ed è particolarmente indicato quando si desidera ottenere rapidamente un modello robusto basato su configurazioni già validate dalla letteratura oppure una volta ottenuti i parametri ottimali da un passo pregresso di ottimizzazione.

Invece se il flag `run_lgb_optimization` è abilitato, il modello viene addestrato dopo aver eseguito una procedura automatizzata per il tuning degli iperparametri. In questo caso il processo prevede una ricerca nei parametri che consente di affinare la configurazione del modello per massimizzare le performance predittive. Il processo è descritto alla sezione 3.6.

Indipendentemente dalla modalità scelta, il modello viene definito come un'istanza di `lgb.LGBMClassifier` con le seguenti caratteristiche: l'argomento `objective` è impostato su `'binary'`, rendendo il modello adatto a problemi di classificazione binaria come la diagnosi di mortalità; un seme casuale (`random_state`) è impostato per garantire la riproducibilità dei risultati, permettendo di ottenere risultati consistenti nelle diverse esecuzioni.

Una volta allenato, il modello viene valutato in base a parametri di valutazione comuni ai modelli di regressione logistica (LR) e XGBoost già affrontati.

3.6. Ottimizzazione iperparametri

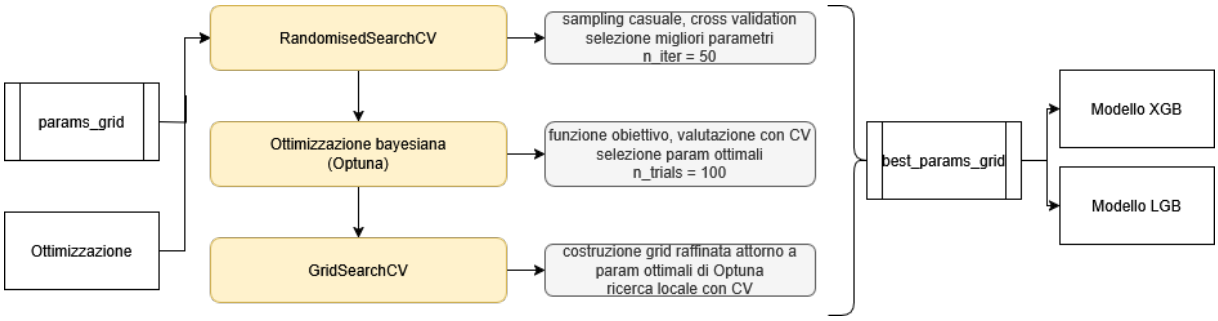


Figura 3.5: Diagramma della pipeline del preprocessing

3.6.1. Randomized SearchCV

Il processo di ottimizzazione degli iperparametri è stato diviso in tre fasi. La prima fase di ottimizzazione si basa su Randomized SearchCV. In questa fase viene testato solo un sottoinsieme di possibili combinazioni per un numero predefinito di iterazioni (`n_iter_random`). Come evidenziato da [21], questo approccio risulta più efficiente rispetto a una grid search completa. La definizione dello spazio di ricerca segue il seguente processo: viene innanzitutto definito per ogni iperparametro un intervallo, continuo per valori numerici e discreto per valori categorici. In seguito, per ogni iterazione viene scelta una configurazione casuale, che viene valutata mediante cross-validation. Al termine delle iterazioni, viene selezionata la combinazione di iperparametri che massimizza la metrica di scoring (ad esempio l'AUROC o l'F1-Score).

3.6.2. Bayesian Optimization con Optuna

La seconda fase consiste nell'utilizzo della Bayesian Optimization tramite Optuna per perfezionare ulteriormente la configurazione ottenuta in precedenza. Optuna sfrutta il Tree-structured Parzen Estimator per costruire un modello probabilistico dello spazio degli iperparametri, aggiornandolo ad ogni prova in base ai risultati ottenuti e indirizzando la ricerca verso le aree più promettenti. La funzione obiettivo, richiesta da Optuna, riceve in input una configurazione proposta dall'oggetto `trial` e valuta il modello mediante `cross_val_score`, restituendo il punteggio della metrica da ottimizzare come l'AUROC o l'F1-Score. Optuna costruisce una distribuzione probabilistica per ogni iperparametro, aggiornandola ad ogni iterazione per concentrare la ricerca nelle regioni che hanno mostrato le migliori prestazioni. Dopo un numero definito di iterazioni (definite attraverso la variabile

`n_trials_optuna`), viene selezionata la configurazione che massimizza la funzione obiettivo, la quale verrà sottoposta a un ulteriore raffinamento nella fase successiva.

3.6.3. Grid Search Raffinato

La terza ed ultima fase prevede l'esecuzione di una Grid Search ristretta attorno ai migliori iperparametri individuati con Optuna, con l'obiettivo di stabilizzare il modello e ridurne la varianza. Per la costruzione della griglia locale si definisce un intorno per ciascun iperparametro: per un parametro continuo x_{opt} vengono inclusi i valori compresi tra $0.8\,x_{\text{opt}}$ e $1.2\,x_{\text{opt}}$; per un parametro intero y_{opt} si testano i valori $y_{\text{opt}} - 1$, y_{opt} e $y_{\text{opt}} + 1$, mentre per i parametri categorici vengono verificate eventuali varianti simili. La Grid Search esplora tutte le combinazioni possibili nella griglia locale, impiegando una validazione più rigorosa (tipicamente con `cv=5`) per ottenere una stima accurata della performance del modello. Alla conclusione della Grid Search viene scelto il modello ottimizzato in base alla metrica di scoring, garantendo il miglior compromesso tra accuratezza e capacità di generalizzazione.

3.6.4. Analisi mediante SHAPLEY

Lo studio [22] ha dimostrato che i valori di Shapley offrono una base teorica solida per l'interpretazione dei modelli predittivi, garantendo trasparenza e consistenza. Inoltre, essi evidenziano come l'adozione di tecniche come SHAP possa migliorare significativamente la comprensibilità dei modelli, specialmente in ambito clinico, dove la spiegazione delle decisioni è fondamentale. Una volta estratti i risultati dei modelli è possibile analizzarli attraverso l'analisi di SHAP, come negli studi [4] e [5], per ottenere maggiore comprensibilità del modello e un'analisi delle variabili determinanti nella diagnosi di mortalità positiva. La possibilità di visualizzare e quantificare l'impatto di variabili cliniche consente di verificare la coerenza clinica delle predizioni, identificare eventuali bias o anomalie mediante l'individuazione di variabili particolarmente influenti sulla predizione, e favorire l'adozione dei modelli in contesti reali, rendendo più semplice integrare i modelli nel processo decisionale clinico grazie a una maggiore interpretabilità.

Il valore di Shapley per una feature calcola, in media, quanto la presenza di quella feature incrementa (o decrementa) la predizione del modello, considerando tutte le possibili combinazioni (o coalizioni) delle altre variabili. In termini matematici, il contributo di una feature viene determinato confrontando la predizione del modello con e senza quella feature, mediante una media ponderata su tutte le possibili permutazioni.

Questo approccio garantisce i seguenti principi: accuratezza locale, coerenza e mancanza. Il criterio della accuratezza locale, noto anche come completezza, prevede che la somma dei valori di SHAP per tutte le feature, sommata al valore di base (di solito la media delle predizioni sul training set), corrisponda esattamente alla predizione del modello per ogni singolo campione. Inoltre, il principio di coerenza stabilisce che se il contributo di una feature aumenta passando da un modello a un altro, il suo valore SHAP non deve diminuire. Infine, il criterio di mancanza impone che le feature non presenti o con valori mancanti assumano un contributo pari a zero.

Il calcolo esatto dei valori di Shapley richiede di valutare tutte le possibili combinazioni di feature, un compito computazionalmente oneroso (NP-completo) per dataset ad alta dimensionalità. Per questo motivo, nella pratica si utilizzano metodi approssimativi come *Tree SHAP*, che sfrutta la struttura degli alberi decisionali per calcolare i valori di Shapley in tempo polinomiale. Questa tecnica è particolarmente efficiente e precisa quando il modello predittivo si basa su strutture ad albero, come XGBoost e LightGBM.

I force plot rappresentano uno strumento grafico fornito dalla libreria `shap` per visualizzare in maniera intuitiva i valori di SHAP. In questi grafici, ogni freccia o barra indica il contributo di una feature: quelle che spingono la predizione verso un valore più elevato, suggerendo un rischio maggiore di mortalità, vengono rappresentate con colori caldi, mentre quelle che riducono il valore predittivo sono indicate con colori freddi. La somma di tutti questi contributi corrisponde alla differenza tra la predizione per il campione e il valore di base, permettendo così di comprendere come ogni variabile influenzi l'output finale del modello.

3.7. Analisi del Bias

3.7.1. Analisi dei Missing Values

Il primo passo della valutazione del bias consiste nell'analizzare il pattern dei dati mancanti all'interno del dataset. È fondamentale verificare se la distribuzione dei missing values sia casuale o meno, poiché una distribuzione non casuale potrebbe indicare bias nella raccolta dei dati, influenzando negativamente il comportamento del modello. Il Test MCAR (Missing Completely At Random) è stato effettuato impiegando la funzione `test_mcar_with_visualization`. Il codice produce varie rappresentazioni grafiche come matriciali e grafici a barre per ottenere un'idea iniziale della distribuzione dei dati mancanza usare grafici visivi (come il diagramma a calore e il dendrogramma). In seguito i dati vengono suddivisi per caratteristiche demografiche come sesso, etnia e lingua e si calcola la media dei valori mancanti per ciascun gruppo. Successivamente si esegue un test statistico di indipendenza (test del chi-quadrato) per controllare l'influenza delle caratteristiche demografiche delle persone sul modello. Un p-value inferiore a 0.05 suggerisce che i dati non siano completamente casuali (MCAR), evidenziando la possibilità che esistano correlazioni sistematiche tra il *missingness* e le caratteristiche demografiche.
Il Test MNAR (Missing Not At Random) è stato effettuato parallelamente; la funzione `test_mnar` valuta se la presenza di valori mancanti in ciascuna colonna possa essere spiegata dalle altre feature. Per ogni variabile, viene creato un indicatore binario della presenza di un dato mancante e si addestra un modello di regressione logistica sui restanti attributi per predire questo indicatore. La qualità della predizione, misurata in termini di AUC, consente di intuire se il *missingness* (mancanza di valori) sia sistematico. Un valore elevato di AUC indica che il modello è in grado di prevedere la mancanza del dato, il che rappresenta un segnale di pattern non casuale (MNAR) e suggerisce la presenza di potenziali bias a livello di raccolta dati. In particolare, la rimozione dei valori mancanti MNAR può comportare una perdita di informazione nel passo di preprocessing e, di conseguenza, un peggioramento del modello addestrato sul dataset.

3.7.2. Forzatura delle Variabili Demografiche

In seguito si procede a studiare l'impatto delle variabili demografiche sulle predizioni del modello attraverso una pratica di "forzatura", in linea con il metodo proposto da Kusner et al. [23], secondo il quale un modello è considerato equo se le sue predizioni rimangono invariate in scenari fittizi in cui le variabili sensibili vengono modificate, mentre tutte le altre caratteristiche restano uguali.

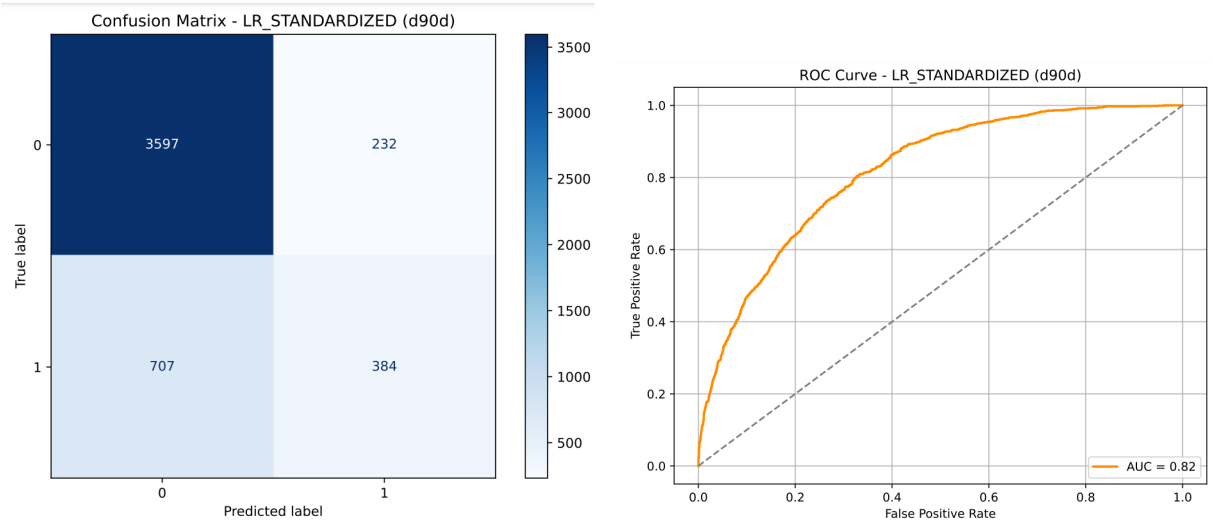
In questa prima fase il modello viene sottoposto a condizioni "what-if" modificando artificialmente i valori delle variabili demografiche: utilizzando le funzioni `mortality_prediction_distribution` o `analyze_demographic_bias` si creano dataset in cui, per ciascuna variabile (come `gender`, `race`, `language` o `insurance`), il valore viene fissato per tutte le osservazioni (ad esempio, impostando tutte le osservazioni a "0" per il genere femminile oppure a "1" per il genere maschile). Questa operazione genera scenari fittizi che permettono di valutare come il cambiamento di una singola variabile sensibile influisca sulle predizioni del modello. Su questi dataset modificati, il modello genera le probabilità di mortalità e le distribuzioni delle stesse, rappresentate tramite istogrammi o curve KDE, vengono poi confrontate tra i diversi scenari: se, ad esempio, forzando il valore di `race` a "Black" si ottiene una distribuzione significativamente diversa rispetto a quella ottenuta forzando il valore a "White", ciò indica che il modello attribuisce un peso rilevante a quella variabile, suggerendo la presenza di bias interno.
La tecnica di forzatura viene ulteriormente integrata con metodi esplicativi come l'analisi tramite SHAP; utilizzando funzioni quali `logistic_regression_model` ed `evaluate_model` si generano SHAP summary plots e force plots che evidenziano il contributo di ciascuna feature sul processo decisionale del modello. Se le variabili sensibili forzate risultano tra quelle con valori SHAP elevati, questo segnala che esse influenzano significativamente le predizioni, contribuendo in maniera marcata al bias del modello.
Infine, la funzione `stratify_performance` divide il dataset in base ai gruppi demografici e, per ciascun gruppo, calcola metriche chiave quali accuracy, AUC-ROC, MCC e F1 score, consentendo di confrontare le performance del modello tra i diversi gruppi e di evidenziare eventuali disparità, che rappresentano un possibile segnale di bias.

4 | Risultati

I modelli sviluppati hanno effettuato una predizione di mortalità in più istanti di tempo: 90 giorni dopo l'ammissione in unità di terapia intensiva (d90d) ed entro fine ospedalizzazione (dinhosp). Per ogni istante di tempo sono stati applicati i tre algoritmi (LR, XGB e LGB), allenando sul trainset e valutando sul testset, ricavando matrici di confusione e curva ROC, dalle quali sono estratti parametri di valutazione per i modelli modelli quali accuratezza, coefficiente di correlazione di Matthew (MCC), AUROC, sensibilità (SE), specificità (SP), precisione, recall e F1 score. Inoltre mediante la analisi di SHAPLEY (SHAP) sono state selezionate le feature più imporanti nella predizione di mortalità da parte dei modelli. Questi risultati sono classificati in base alla predizione: classe 0 nel caso di predizione di sopravvivenza, classe 1 nel caso di predizione di mortalità.
Per l'identificazione di un possibile bias socio-economico è stata effettuata una analisi in più step (4.4). Dapprima le distribuzioni dei valori mancanti sono state studiate per osservare una possibile discriminazione sulla quantità di misure per pazienti di diversi gruppi demografici, successivamente è stato valutato il peso delle variabili demografiche sul modello; in conclusione è stata messa in pratica una forzatura delle colonne demografiche all'interno del dataset e un'analisi delle performance dei modelli sul nuovo dataset trasformato. In seguito sono state effettuate delle analisi sulle performance dei modelli utilizzati, stratificando per ogni variabile osservata (lingua, sesso, etnia e tipo di assicurazione), per verificare la presenza di una situazione diversificata a seconda della condizione sociale, economica e demografica dei pazienti.

4.1. Risultati a 90 giorni

4.1.1. Logistic Regression



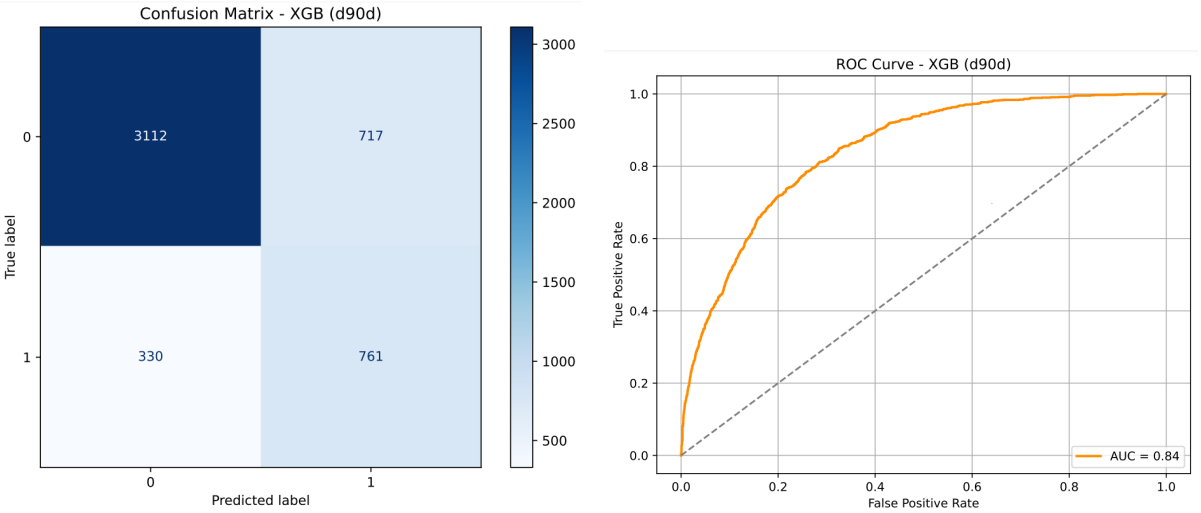
(a) Matrice di Confusione LR (90 giorni)

(b) Curva ROC LR8 (90 giorni)

Figura 4.1: Confronto tra matrice di confusione e curva ROC per Logistic Regression (90 giorni)

Nella figura 4.1a la matrice di confusione riporta che il modello classifica correttamente circa 4000 casi. Riguardo la classe 0 si ha una precisione di 0.84. Il recall è di 0.94 e l'F1-Score è di 0.88. Per la classe 1 invece si hanno dei valori minori per tutti i parametri. La precisione è 0.62, il recall è 0.35 e l'F1-Score è 0.45. Si è calcolato anche l'MCC che risulta essere 0.37. Nella figura 4.1b è riportata l'accuratezza del modello sotto forma di curva ROC che evidenzia un'area sotto la curva (AUROC) di 0.82. La sensibilità del modello è 0.35 e la specificità di 0.94.

4.1.2. XGBoost

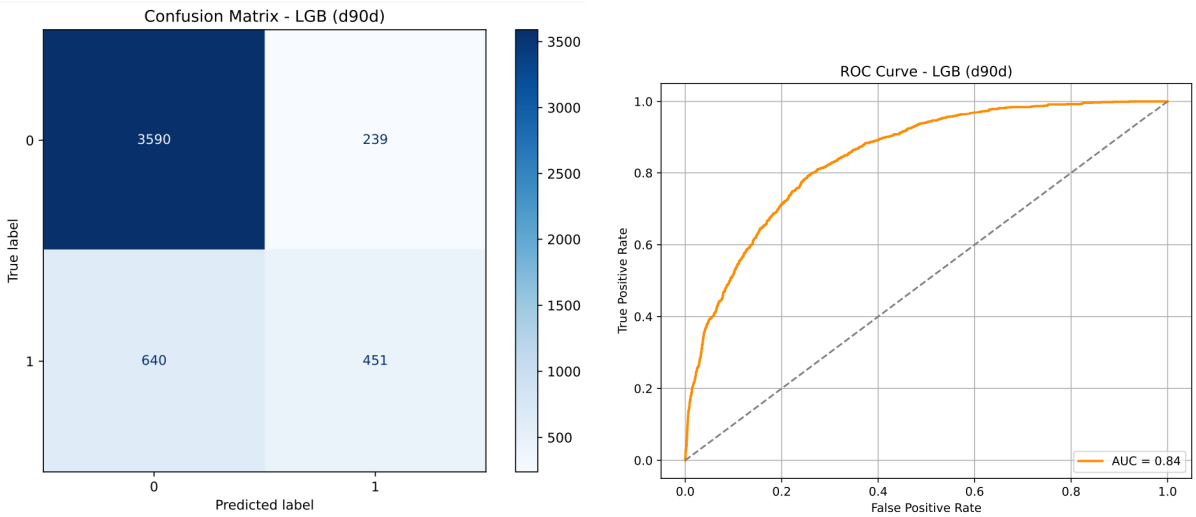


(a) Matrice di Confusione XGBoost (90 giorni) (b) Curva ROC XGBoost (90 giorni)

Figura 4.2: Confronto tra matrice di confusione e curva ROC per XGBoost (90 giorni)

La figura 4.2a riporta la matrice di confusione per XGBoost. I risultati per questo modello riportano il numero dei falsi negativi dimezzato rispetto al modello 4.1a che porta ad avere per la classe 1 un Recall di 0.70, un F1-Score di 0.59 e la precisione di 0.51. Per la classe 0 aumenta la precisione a 0.90 ma diminuiscono Recall (0.81) e F1-Score (0.86). L'MCC per il modello sale a 0.46 e, come riportato in figura 4.2b, l'AUROC è 0.84, con un'accuratezza di 0.79. Questi valori sono modificati dal fatto che diminuiscono, oltre ai falsi negativi, anche i veri positivi. La sensibilità è di 0.70, nettamente migliore di Logistic Regression e la specificità di 0.81.

4.1.3. LGBM



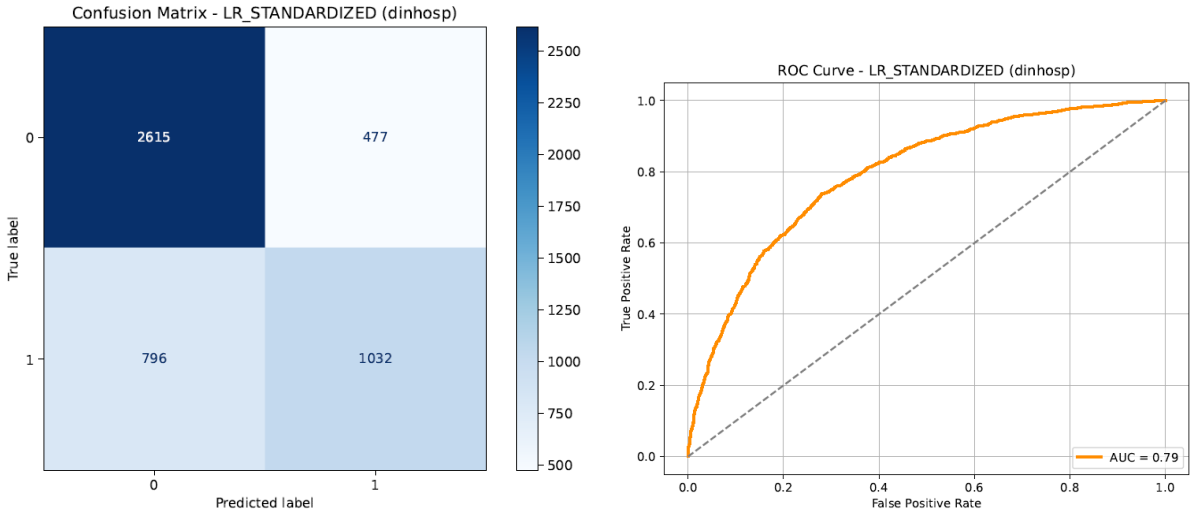
(a) Matrice di Confusione LGBM (90 giorni) (b) Curva ROC LGBM (90 giorni)

Figura 4.3: Confronto tra matrice di confusione e curva ROC per LGBM (90 giorni)

La matrice di confusione per LightGBM 4.3a evidenzia che il modello ha un'accuratezza maggiore degli altri due modelli e corrisponde a 0.82. La classe 1 riporta un F1-Score di 0.51, un recall di 0.41 e una precisione di 0.65. La classe 0 invece ha precisione di 0.85, un recall uguale a 0.94 e un F1-Score di 0.89 evidenziando un miglioramento sulla classe 0 rispetto a XGBoost tranne per la precisione. Discorso opposto riguarda la classe 1 in cui solo la precisione migliora mentre gli altri indicatori peggiorano. L'MCC è 0.42, quindi leggermente inferiore a XGBoost. Nella figura 4.3b è riportata la curva ROC del modello con la relativa area sottesa (AUROC = 0.84) che rimane invariata rispetto a XGBoost. La sensibilità è 0.41 e la specificità è 0.94.

4.2. Risultati a fine ospedalizzazione

4.2.1. Logistic Regression



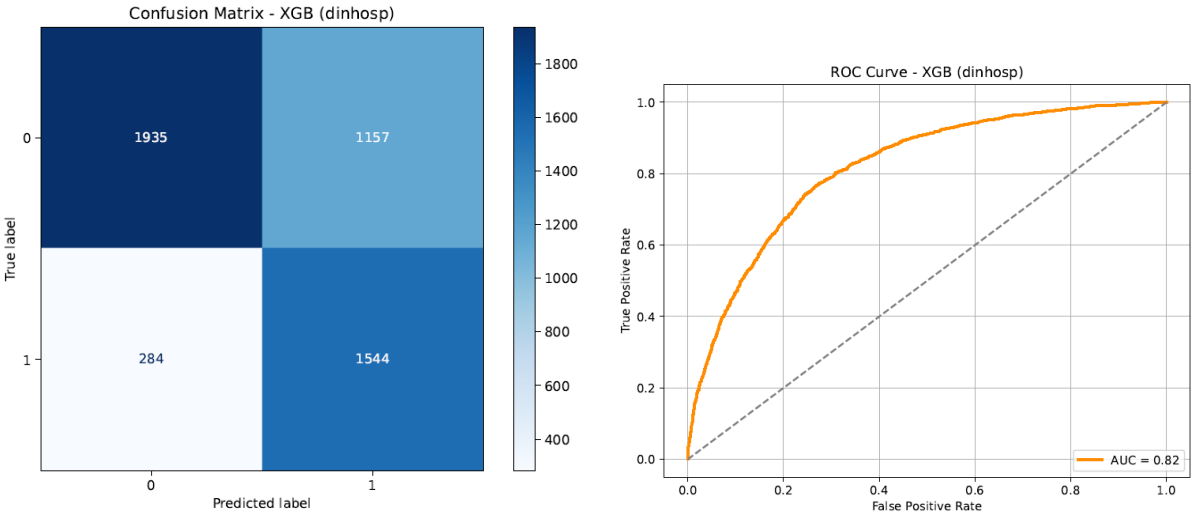
(a) Matrice di Confusione Logistic Regression

(b) Curva ROC Logistic Regression

Figura 4.4: Confronto tra matrice di confusione e curva ROC per Logistic Regression

Il modello che sfrutta Logistic Regression può essere rappresentato dalla matrice soprastante (Figura 4.4a). Essa mostra come il modello classifichi esattamente circa 3600 pazienti (1032 Veri Positivi e 2615 Veri Negativi) avendo un'accuratezza di 0.74. Il modello ha maggiore difficoltà a trattare la classe 1 e ciò è evidente dai parametri statistici. Per questa classe, infatti, si ha una precisione di 0.68, un recall di 0.56 e un F1-Score di 0.62. In confronto, la classe 0 riporta un valore di precisione di 0.77, un recall di 0.85 e un F1-Score di 0.80. L'MCC è 0.43 che indica una correlazione tra predizione del modello e realtà positiva. La difficoltà nella classificazione della classe 1 è evidente dal valore dei falsi negativi (796). Nel grafico seguente (Figura 4.4b) è riportata la curva ROC con conseguente AUROC di 0.79 che risulta peggiorato di 0.03 rispetto alla classificazione analoga con la predizione a 90 giorni.

4.2.2. XGBoost

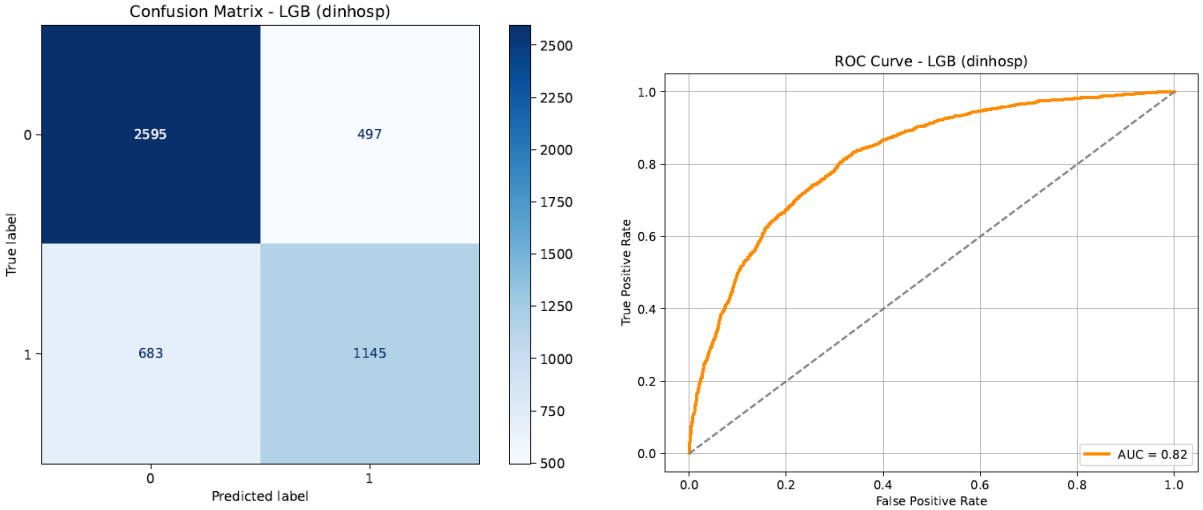


(a) Matrice di Confusione XGBoost (b) Curva ROC XGBoost

Figura 4.5: Confronto tra matrice di confusione e curva ROC per XGBoost

La figura 4.5a riporta la matrice di confusione per XGBoost. Nella classificazione del dataset si nota come diminuisca rispetto a Logistic regression (figura 4.4a) il numero dei falsi negativi da 796 a 284 con però un aumento dei falsi positivi ed una diminuzione dei veri positivi. Questi fattori vanno ad incidere sulle metriche di valutazione portando l'accuratezza a 0.71. La maggiore classificazione della classe 1 porta ad un aumento rispetto a LR sia del Recall (0.84) che dell'F1-Score (0.68) nonostante una diminuzione della precisione (0.57). Il modello in questione ha più difficoltà a classificare la classe 0 perché nonostante la precisione aumenti a 0.87, il Recall cala a 0.63 e l'F1-Score arriva a 0.73. Abbiamo un leggero aumento invece dell'MCC che arriva a 0.46. La figura sottostante (Figura 4.5b) riporta la curva ROC che presenta un AUROC di 0.82, leggermente migliore di LR (4.4b).

4.2.3. LGBM



(a) Matrice di Confusione LGBM

(b) Curva ROC LGBM

Figura 4.6: Confronto tra matrice di confusione e curva ROC per LGBM

Analizzando la figura 4.6a si nota che il numero di falsi negativi è maggiore rispetto a quello del modello XGBoost, anche se minore di Logistic Regression; si nota inoltre una diminuzione dei falsi positivi. Questi valori spiegano come LGB sia più conservativo e meno incline a dare falsi positivi, ma possa perdere alcuni veri positivi. L'accuratezza è migliorata a 0.76 e, dato l'aumento dei falsi negativi, è evidente uno sbilanciamento rispetto alla classe 1. Questa riporta una precisione di 0.70, un Recall di 0.63 e un F1-Score di 0.66. Dall'altra parte vi è un miglioramento delle metriche riguardanti la classe 0, il cui Recall arriva a 0.84 e l'F1-Score a 0.82 su un gruppo di supporto di 1828 elementi. La curva ROC seguente (Figura 4.6b) riporta l'AUROC migliore tra i tre modelli per predizione a fine ospedalizzazione e vale 0.82. L'MCC risulta ulteriormente migliorato di 2 centesimi: da 0.46 (XGBoost) a 0.48 (LGBM).

4.3. Risultati analisi SHAPLEY

4.3.1. Analisi XGB per mortalità a 90 giorni

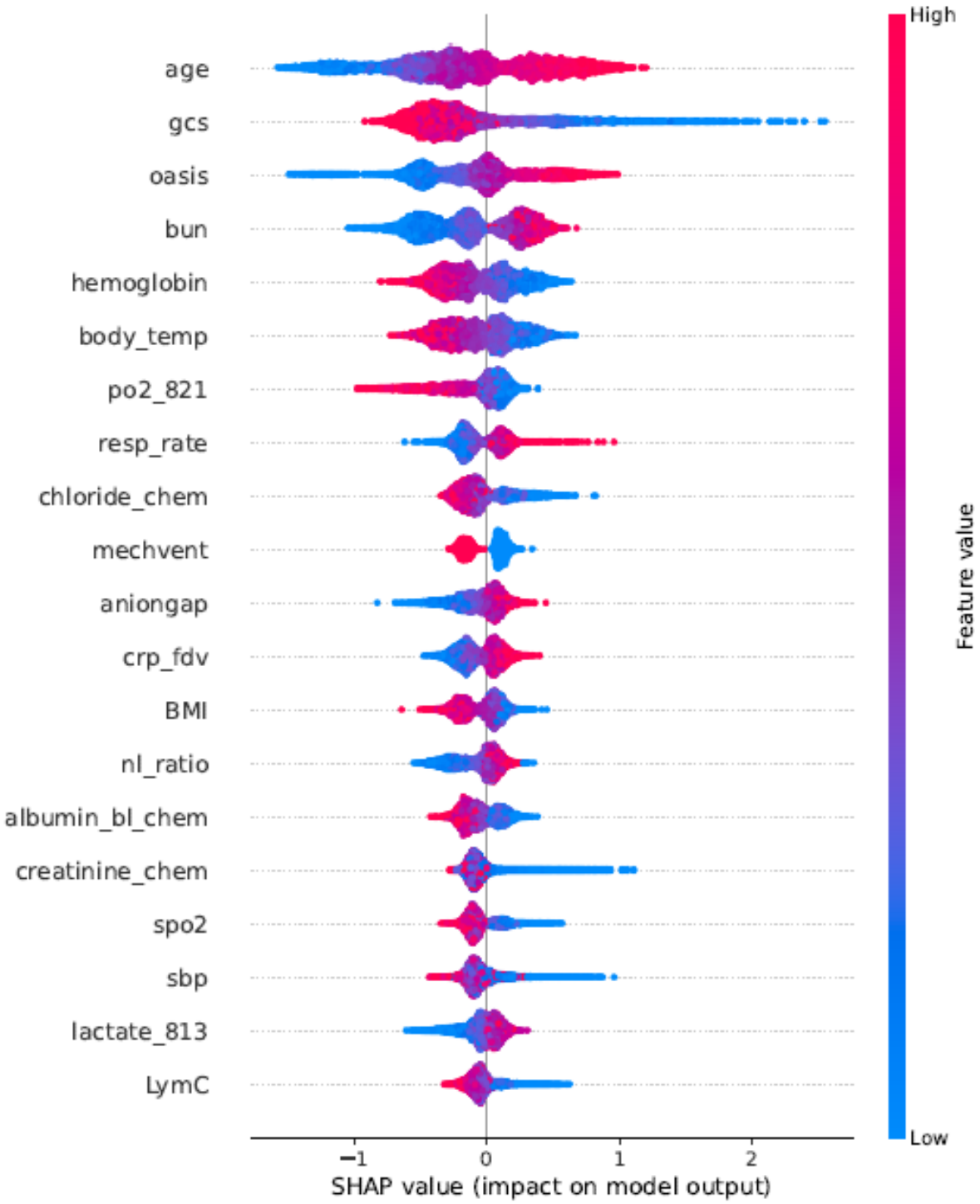


Figura 4.7: SHAP explainer per XGB d90d

L'analisi SHAP sul modello di predizione a 90 giorni evidenzia che le variabili con il maggiore impatto sono il conteggio dei linfociti (LymC), in cui valori più bassi determinano un marcato incremento del rischio di mortalità, il livello di lattato, per il quale valori elevati orientano la predizione verso esiti negativi, la pressione arteriosa sistolica (sbp), poiché una sua riduzione è associata a instabilità emodinamica e ad un conseguente aumento del rischio, e la saturazione ossigenica (spo2), il cui decremento comporta una maggiore probabilità di esito negativo. Inoltre, biomarcatori di disfunzione renale e infiammatoria, come l'elevata creatinina e i valori aumentati di proteina C reattiva (CRP), insieme al rapporto neutrofili/linfociti (nl_ratio), rappresentano ulteriori fattori critici che, in sinergia, orientano la predizione verso una maggiore probabilità di mortalità a 90 giorni, come visibile in figura 4.10. Altri fattori inclusi nei parametri critici dalla analisi SHAP sono età, Glasgow Coma Scale (GCS), azoto ureico nel sangue (BUN) e insieme di informazioni sui risultati e sulla valutazione (OASIS), i quali formano una base per la predizione di mortalità in pazienti critici, osservabili nei force plot a immagine 4.8 e 4.9.

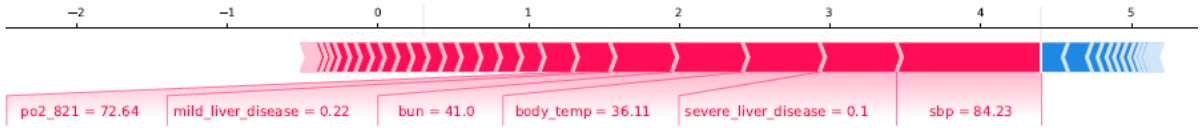


Figura 4.8: Force plot per pazienti ad alto rischio per d90d

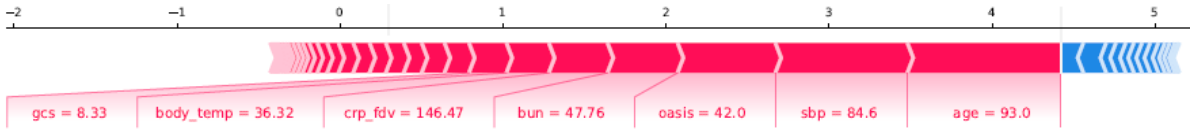


Figura 4.9: Force plot per pazienti ad alto rischio per d90d

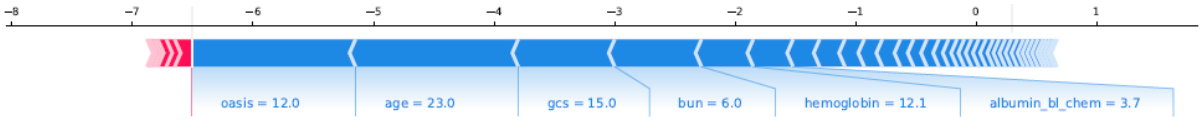


Figura 4.10: Force plot per pazienti a basso rischio per d90d

4.3.2. Analisi XGB per mortalità in ospedale

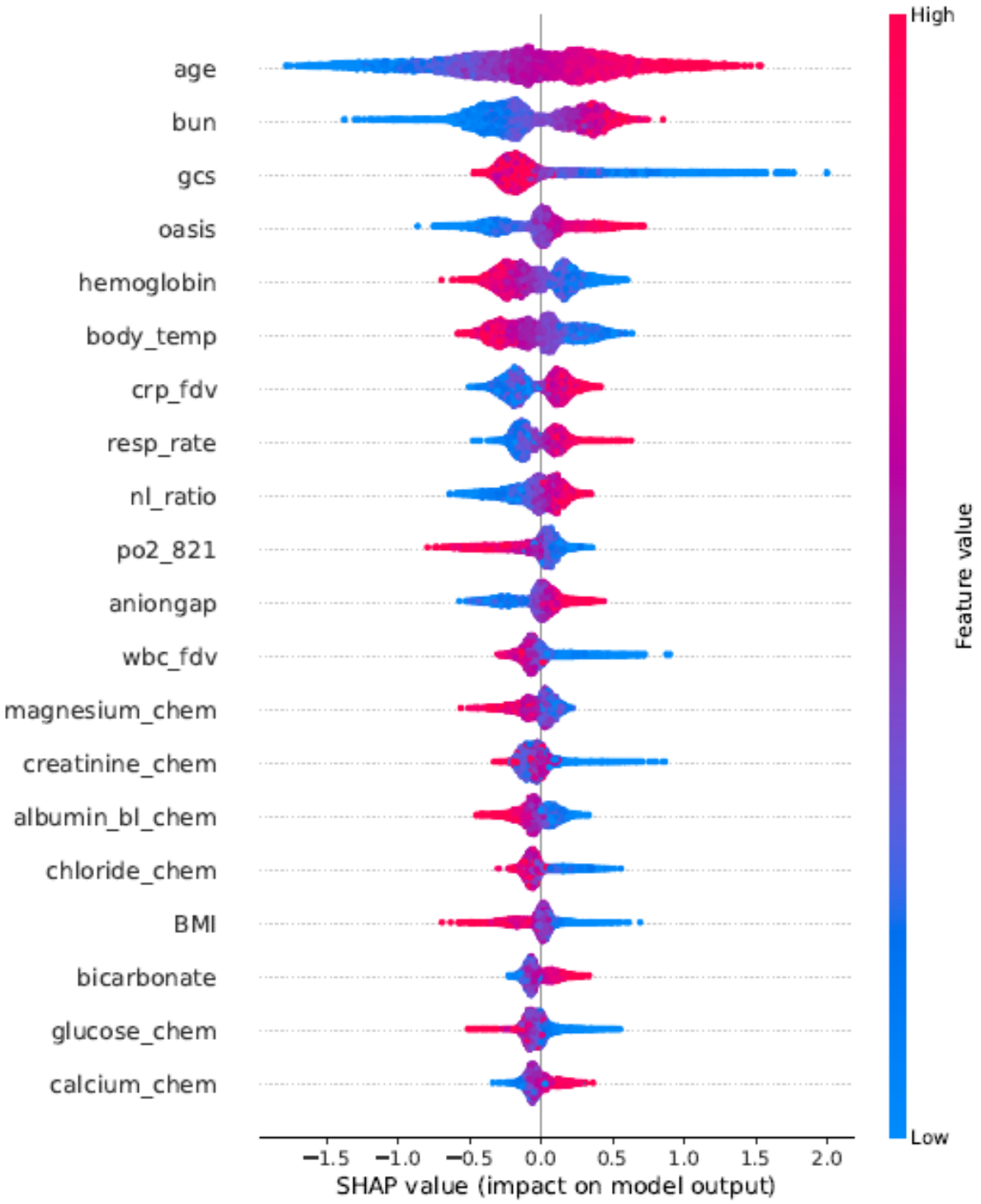


Figura 4.11: SHAP explainer per XGB dinhosp

Per quanto riguarda la mortalità in ospedale, l'analisi SHAP evidenzia un profilo differente di variabili critiche. In particolare, la CRP gioca un ruolo centrale, con livelli elevati che contribuiscono in modo significativo all'incremento del rischio predetto. Valori anomali di glucosio e di bicarbonato indicano un'alterazione del metabolismo, risultando determinanti nella predizione, mentre un livello di creatinina elevata, segno di disfunzione renale, emerge come un fattore importante nell'aumentare il rischio. Gli squilibri elettrolitici, come le variazioni nei livelli di calcio e cloruro, insieme al rapporto neutrofili/linfociti, completano il quadro, suggerendo che un insieme di alterazioni concomitanti sposta la previsione verso un maggior rischio di mortalità in ospedale, come osservabile in figura 4.12 e 4.14. Anche nel caso del modello di predizione della mortalità in ospedale parametri come età, GCS, BUN e OASIS risultano critici, i quali insieme costituiscono una base per prevedere la mortalità nei pazienti osservati, come rilevabile in figura 4.13.

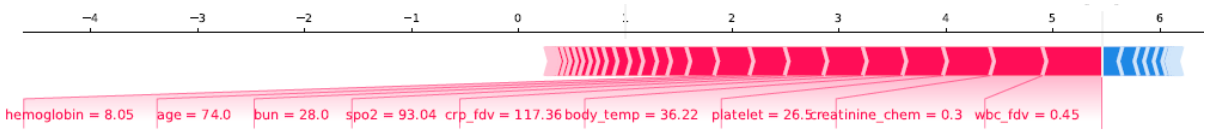


Figura 4.12: Force plot per pazienti ad alto rischio per dinhosp

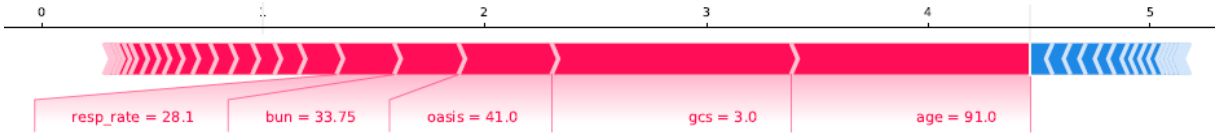


Figura 4.13: Force plot per pazienti ad alto rischio per dinhosp

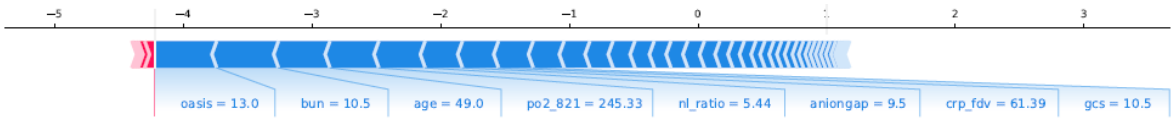


Figura 4.14: Force plot per pazienti a basso rischio per dinhosp

4.4. Risultati al Bias-Socio Economico

Un secondo obiettivo sviluppato nel corso di questo studio riguarda la valutazione dell'effetto di alcuni parametri sui modelli utilizzati per la predizione.

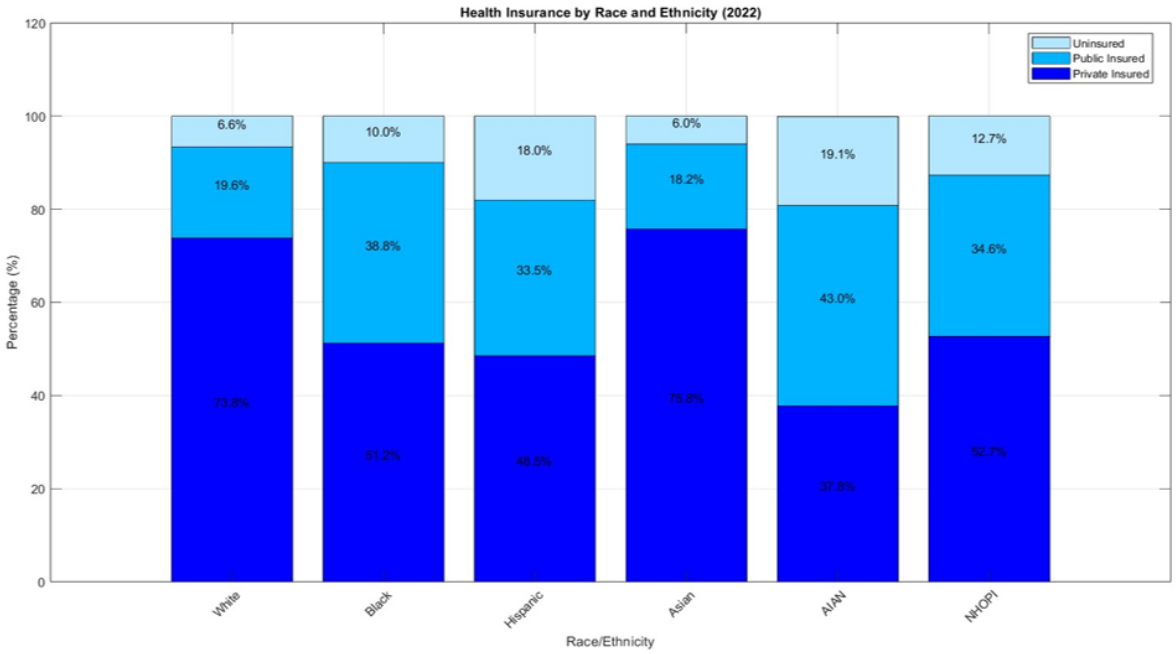


Figura 4.15: Distribuzione dell'assicurazione sanitaria negli Stati Uniti nel 2022

Nella 4.15, è mostrata la distribuzione dell'assicurazione sanitaria negli Stati Uniti del 2022. Lungo l'asse delle ascisse sono rappresentate le principali etnie (il termine AIAN indica American Indian and Alaska Native Resources, mentre il termine NHOPI significa Native Hawaiian and Other Pacific Islanders) presenti negli Stati Uniti, mentre lungo l'asse delle ordinate sono rappresentate le percentuali di una tipologia specifica di assicurazione per una determinata etnia. Le categorie assicurative si dividono in assicurazione privata, rappresentata con il colore blu scuro; assicurazione pubblica, in azzurro, e, infine, non assicurati in azzurro chiaro. Dal grafico risulta evidente come vi sia una prevalenza di assicurati privati tra etnie asiatiche e bianche, con valori che oscillano tra il 73% e il 75%. Al contrario, per quanto riguarda individui neri, ispanici e nativi americani si ha una prevalenza di copertura sanitaria pubblica e mancata copertura sanitaria. Queste prime informazioni possono derivare da uno squilibrio a livello di possibilità economiche per etnie diverse ([24]). Questo, dunque, rende legittima la volontà di verificare la resistenza del modello predittivo al bias socio-economico.
Prima di introdurre i metodi di verifica utilizzati, è necessario specificare di che assicurazioni si costituisce il sistema sanitario degli Stati Uniti:

- Assicurazione privata: è un tipo di assicurazione offerta da diverse aziende statunitensi e permettono la copertura annuale di trattamenti medici su soggetti che pagano una quota mensile. La quota da pagare varia in base al piano scelto dall'individuo, che si differenzia sulla base dei servizi sanitari offerti.
- Medicaid: è un tipo di assicurazione offerta dallo stato e di cui possono usufruire famiglie o individui con basso reddito.
- Medicare: indica il programma assicurativo offerto dallo stato a persone anziane e soggetti con disabilità.
- Non-assicurati/ altre tipologie di assicurazioni

([25]). Passando, dunque, all'analisi del bias socio-economico sul modello predittivo, tramite l'iniziale eliminazione e la successiva forzatura delle colonne demografiche del dataset all'interno del modello è stato evidenziato come il grafico della distribuzione di probabilità della predizione non vari in maniera sostanziale.

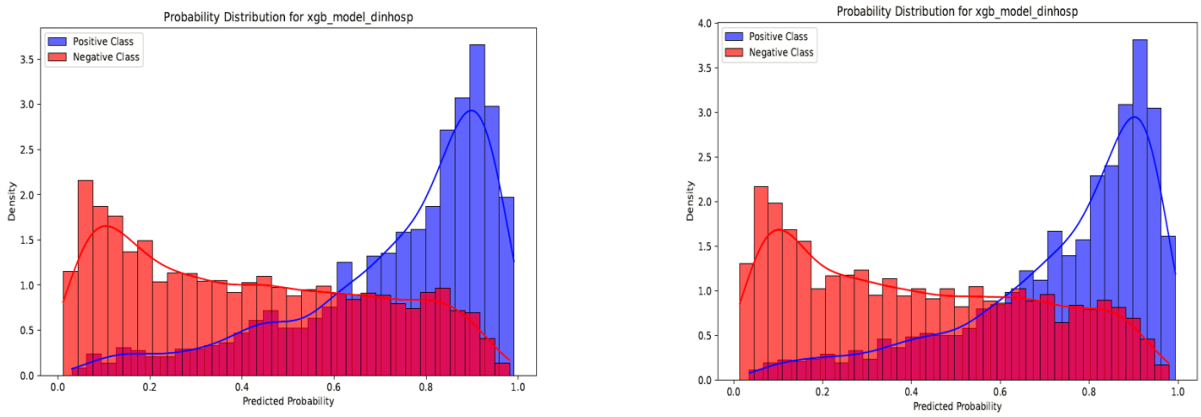


Figura 4.16: Distribuzione di probabilità della predizione senza colonne demografiche (grafico a sinistra) e con colonne demografiche (grafico a destra)

Andando più nello specifico, quando si approfondisce lo studio della distribuzione di probabilità nei 90 giorni post-ricovero si può osservare come i grafici siano simili fra loro, senza critiche differenze in ambito assicurativo per etnie diverse e sesso dei soggetti. Al contrario si iniziano a notare delle disuguaglianze nella distribuzione di probabilità entro fine ospedalizzazione.

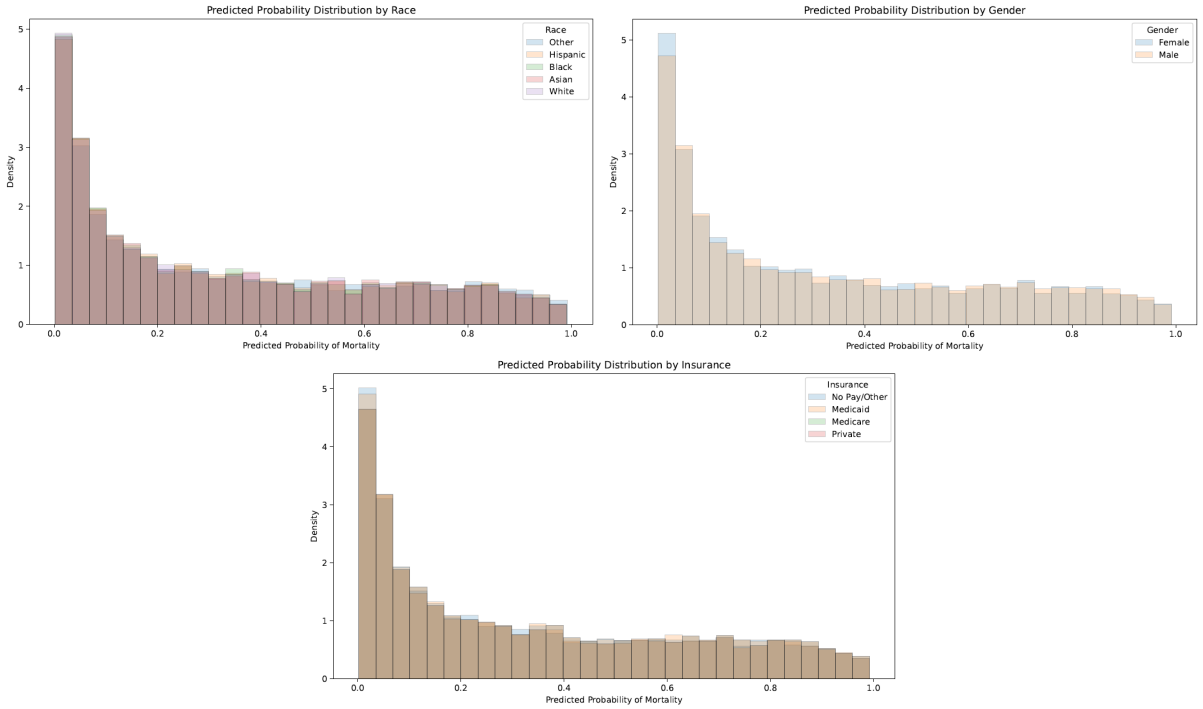


Figura 4.17: Distribuzione di probabilità della predizione sulla base del sesso, dell'etnia e dell'assicurazione del paziente nei 90 giorni post-ricovero

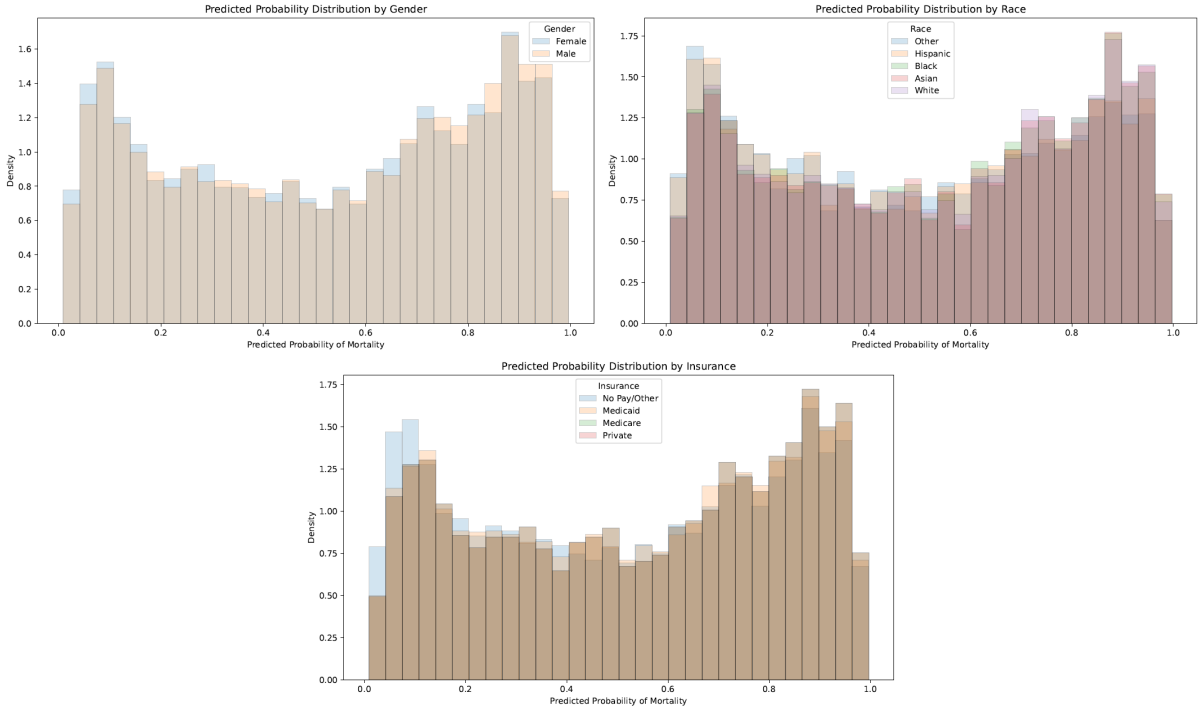


Figura 4.18: Distribuzione di probabilità della predizione sulla base del sesso, dell'etnia e dell'assicurazione del paziente entro fine ospedalizzazione

Nelle seguenti tabelle sono rappresentati i risultati dei parametri del test set valutati in termini di accuratezza, AUC, MCC e F1-Score in relazione a genere, etnia e tipologia di assicurazione del paziente. Il confronto tra i diversi valori per diverse categorie permette di identificare una possibile dipendenza del modello dal bias socio-economico. La somiglianza dei valori, invece, conferma la robustezza del modello.
La differenza nei risultati osservabile attraverso i parametri di valutazione del modello rispetto alla variabile assicurazione sono attribuibili alla differenza di valori mancanti a seconda delle assicurazioni: infatti nel caso di un paziente con assicurazione privata o altre sono presenti rispettivamente 58.601 e 65.179 valori mancanti, mentre nel caso di assicurazione con Medicare o Medicaid vi sono 9.961 valori mancanti. Questa grande differenza di valori mancanti porta a una migliore performance del modello per pazienti con assicurazione di tipo Medicare o Medicaid.

Genere	Accuratezza	AUC-ROC	MCC	F1
Uomo	96-97%	0.57	0.27	0.24
Donna	97%	0.58	0.26	0.25

Etnia	Accuratezza	AUC-ROC	MCC	F1
Altre (Race 0)	94-97%	0.58	0.28	0.26
Hispanic (Race 1)	97%	0.61	0.34	0.32
Black (Race 2)	97%	0.59	0.25	0.24
Asian (Race 3)	95%	0.62	0.34	0.33
White (Race 4)	97%	0.57	0.24	0.22

Assicurazione	**Accuratezza**	**AUC-ROC**	**MCC**	**F1**
Others (Insurance 0)	97%	0.57	0.24	0.22
Medicaid/Medicare (Insurance 1)	98%	0.65	0.41	0.39
Private (Insurance 2)	96%	0.58	0.27	0.25

Tabella 4.1: Risultati delle metriche per genere, etnia e assicurazione

4.5. Sintesi dei risultati

I risultati di XGBoost, ricavati da una valutazione basata sul test set e poi migliorati attraverso l'ottimizzazione degli iperparametri, sono risultati i migliori per questo tipo di predizione. Questo modello di predizione di mortalità infatti riporta questi valori a 90 giorni dall'ammissione e a fine ospedalizzazione:

Metrica	**XGBoost - 90 giorni**	**XGBoost - Fine Ospedalizzazione**
F1 Score	0.59	0.68
AUC-ROC	0.84	0.82
Recall	0.70	0.84
Accuracy	0.79	0.71
MCC	0.46	0.46

Tabella 4.2: Metriche di XGBoost per la predizione della mortalità a 90 giorni e a fine ospedalizzazione.

Dalle analisi di SHAP condotte sono risultati due diversi gruppi di variabili a seconda che la predizione fosse a 90 giorni o a fine ospedalizzazione:

- 90 giorni = conta dei linfociti, lattato sierico, pressione arteriosa e saturazione
- Fine ospedalizzazione = squilibri elettrolitici, valori di glucosio e bicarbonato, creatinina e CRP

Successivamente, nella ricerca di un bias socio-economico nel modello, si è arrivati a risultati pressoché simili non attribuibili in definitiva al modello, ma a possibili comorbidità e errori di misurazione. L'unica situazione anomala si ha analizzando i risultati per le assicurazioni in cui si nota che il modello ha risultati migliori se il paziente è assicurato Medicare o Medicaid come riportato in Tabella 4.3. Questo è giustificabile in quanto nel caso Medicare o Medicaid si ha un numero minimo di valori mancanti (9961) in confronto agli altri tipi assicurativi (65179 per "Others" e 58601 per "Private"); di conseguenza, la performance del modello è migliore nel caso dei pazienti assicurati con Medicare o Medicaid.

Assicurazione	**Accuratezza**	**AUC-ROC**	**MCC**	**F1**
Others (Insurance 0)	97%	0.57	0.24	0.22
Medicaid/Medicare (Insurance 1)	**98%**	**0.65**	**0.41**	**0.39**
Private (Insurance 2)	96%	0.58	0.27	0.25

Tabella 4.3: Rilevanza delle metriche per pazienti assicurati Medicare/Medicaid

5 | Discussione

5.1. Modelli di predizione

Nei risultati a 90 giorni dall'ammissione in Unità di Terapia Intensiva (4.1) il modello di Logistic Regression (LR)(4.1a) riporta un Recall di 0.35 e una Precisione di 0.62 che possono essere dovuti allo sbilanciamento dei dati. Le metriche della classe 0 hanno inoltre dei valori maggiori rispetto alla classe 1, evidenziando una migliore capacità del modello nella loro classificazione. Il discorso migliora con l'implementazione di XGBoost (4.2a). Il modello migliora F1-Score e Recall (che passano rispettivamente da 0.45 a 0.59 e da 0.35 a 0.70) data la diminuzione dei falsi negativi ma peggiora nel riconoscimento della classe 0 in quanto aumentano i falsi positivi. Clinicamente tra i due modelli si è più propensi a scegliere XGBoost rispetto a Logistic Regression in quanto il valore dei falsi negativi è minore ma bisogna anche considerare che l'utilizzo di un modello che classifichi un numero così alto di falsi positivi rappresenta per il sistema un alto spreco di risorse. Migliorando anche l'AUROC, XGBoost si può considerare migliore per il nostro studio. Comparandolo con il modello che sfrutta LightGBM (4.3a), l'accuratezza aumenta rispetto agli altri due modelli ma si è vista l'importanza dell'F1-Score nella gestione di dati clinici e per il modello in questione si ha una leggera riduzione da 0.59 a 0.51. Fondamentale è anche il recall che risulta inferiore (0.41 rispetto a 0.70 di XGBoost). Questo modello ha però una migliore precisione sulla classe 1 rispetto a XGBoost e Logistic Regression (0.62).

Tabella 5.1: Confronto tra le classi a 90 giorni

Modello	Precisione		Recall		F1-Score	
	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
LR	0.84	0.62	**0.94**	0.35	0.88	0.45
XGB	**0.90**	0.51	0.81	**0.70**	**0.86**	**0.59**
LGB	0.85	**0.65**	**0.94**	0.41	0.89	0.51

Tabella 5.2: Confronto metriche a 90 giorni

Modello	Accuratezza	MCC	AUROC	SE	SP
LR	0.81	0.37	0.82	0.35	**0.94**
XGB	0.79	**0.46**	**0.84**	**0.70**	0.81
LGB	**0.82**	0.42	**0.84**	0.41	**0.94**

Riguardo al confronto tra classi nella tabella soprastante (5.1) si può evidenziare maggiormente la motivazione di tale sbilanciamento ponendo le matrici di confusione per i tre modelli a confronto:

Tabella 5.3: Confronto Matrici a 90 giorni

Modello	Veri Positivi	Falsi Positivi	Veri Negativi	Falsi Negativi
LR	384	232	3597	707
XGB	761	717	3112	330
LGB	451	239	3590	640

Guardando la colonna "Falsi Negativi" si nota infatti come XGBoost abbia un valore dimezzato rispetto agli altri due modelli, facendo sì che metriche come Recall e F1-Score siano migliori e che determinino questo modello come il migliore a livello clinico. Facendo un discorso analogo, si è eseguito lo stesso protocollo per i modelli di predizione di mortalità a fine ospedalizzazione (4.2). LightGBM risulta il modello con la migliore accuratezza (0.82) indicando una buona capacità di bilanciamento delle predizioni avendo anche un numero intermedio tra XGBoost e LR di falsi positivi e negativi. Il Recall non è ottimale, rendendo così il modello poco efficace se l'obiettivo è l'individuazione dei positivi. Logistic Regression ha un andamento delle metriche simile a LightGBM ma con valori minori. Il Recall è il peggiore tra i tre modelli e tende quindi a non classificare in maniera ottimale i casi positivi, obiettivo fondamentale in ambito clinico. Il vantaggio che ha rispetto agli altri modelli è la maggiore interpretabilità e la trasparenza delle decisioni da prendere. Il modello che ha un comportamento migliore con la Classe 1 è XGBoost che riesce a individuare meglio i pazienti positivi avendo il Recall e l'F1-Score migliori degli altri modelli. Il lato negativo di XGBoost rappresenta la classificazione dei falsi positivi che sono molto più elevati di LightGBM e di LR. La scelta del modello va fatta sempre in relazione all'applicazione che si va ad adottare con esso, quindi se l'obiettivo è quello di individuare principalmente i casi positivi si andrà a selezionare il modello con F1-Score e Recall migliori (XGBoost nel nostro caso), mentre se si è interessati a contenere la spesa

per falsi positivi a discapito di una percentuale di positivi non classificati come tali (falsi negativi) sarà meglio adottare LightGBM. Se d'altro canto si cerca un modello molto semplice e maggiormente interpretabile, che comunque presenti una buona accuratezza, si può ricorrere a Logistic Regression. Nel caso in questione si preferisce utilizzare un modello che identifichi i positivi senza considerare la spesa sanitaria per i falsi positivi e XGBoost risulta la scelta migliore. Le analisi a 90 giorni e a fine ospedalizzazione hanno riportato metriche diverse: si prova a effettuare un confronto tra le matrici di XGBoost per entrambi i lassi di tempo.

Tabella 5.4: Confronto tra matrici di confusione per XGB (dinhosp) e XGB (d90d)

	XGB (dinhosp)		XGB (d90d)	
	Pred 0	Pred 1	Pred 0	Pred 1
True 0	1935	1157	3112	717
True 1	284	1544	330	761

Per il modello a fine ospedalizzazione (4.5a) le previsioni corrette, ovvero "Veri Positivi" e "Veri Negativi" sono 4437, una buona parte del totale. Tuttavia i "Falsi Positivi" sono 1521 simboleggiando una tendenza del modello a classificare erroneamente i pazienti appartenenti alla classe negativa. Il numero dei "Falsi Negativi" è contenuto (356) dimostrando il perché di un Recall così elevato. Il modello per previsione a 90 giorni ha invece una migliore capacità nell'individuare i veri negativi e un numero di falsi positivi ridotti a quasi la metà rispetto all'altro modello (817) riducendo eventualmente le spese superflue dovute alla classificazione della classe negativa. Il numero di "Veri Positivi" però è inferiore e i "Falsi Negativi" aumentano attribuendo un Recall inferiore.
La differenza tra i modelli determina quindi una scelta in base agli obiettivi:

- XGB (dinhosp) se si vuole massimizzare il Recall riducendo al minimo i falsi negativi, sacrificando la precisione sui negativi
- XGB (d90d) se si vuole avere una precisione migliore evitando di creare falsi positivi, sacrificando però alcune positività

Tabella 5.5: Confronto tra le classi per dinhosp

Modello	Precisione		Recall		F1-Score	
	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
LR	0.77	0.68	**0.85**	0.56	0.80	0.62
XGB	**0.87**	0.57	0.63	**0.84**	0.73	**0.68**
LGB	0.79	**0.70**	0.84	0.63	**0.81**	0.66

Tabella 5.6: Confronto metriche globali per dinhosp

Modello	Accuratezza	MCC	AUROC	SE	SP
LR	0.74	0.43	0.79	0.56	**0.85**
XGB	0.71	0.46	**0.82**	**0.84**	0.63
LGB	**0.76**	**0.48**	**0.82**	0.63	0.84

5.2. Bias socio-economico

Dai risultati riportati precedentemente si può osservare come il modello non venga influenzato dalla presenza di variabili socio-economiche all'interno del dataset. Nella Figura 4.17 e nella Figura 4.18, sono mostrate le distribuzioni di probabilità di mortalità per le diverse categorie di genere, di etnia e assicurazione. Queste permettono di verificare la presenza di disuguaglianze, le quali, se evidenti e elevate, possono dimostrare una diversificazione nel trattamento dei pazienti, da parte del personale medico. Poiché, però, anche forzando le variabili socio-economiche nel modello, non vi sono differenze sostanziali tra le distribuzioni, le piccole variazioni presenti possono essere attribuite a diverse situazioni economiche, che indicano stili di vita diversi e, quindi, un decorso della sepsi e una reazione alle cure che varia da soggetto a soggetto. Non devono, inoltre, essere ignorate le possibili comorbidità, ossia tutti quei decessi classificati come settici, ma che in realtà non hanno nulla a che vedere con la condizione. Infine, osservando la Tabella 5.7, che mostra i risultati del modello, per ciascuna categoria di genere, etnia e assicurazione, si nota come i risultati in termini di accuratezza, AUC-ROC, MCC e F1 siano simili, suggerendo un'indipendenza del modello dal bias. Come si può vedere nella tabella, i risultati sulla base del genere del paziente sono pressoché identici, mentre per quanto riguarda le etnie e le assicurazioni, i valori di accuratezza oscillano tra il 95 e il 97%, l'AUC-ROC tra 0.58 e il 0.65, mentre MCC e F1 presentano valori compresi fra 0.24 e 0.41.

Le variazioni più alte si trovano nei risultati riguardanti le assicurazioni, ma, come già specificato in precedenza, questo può essere relativo ad una diversa situazione economica dei soggetti e quindi a variabili indipendenti dal trattamento sanitario. La differenza nei risultati rispetto alla variabile assicurativa è attribuibile alla differenza nel numero di valori mancanti: infatti, nel caso di pazienti con assicurazione privata o altre, sono presenti rispettivamente 58.601 e 65.179 valori mancanti, mentre nel caso di assicurazione con Medicare o Medicaid vi sono 9.961 valori mancanti. Questa differenza sostanziale porta a una migliore performance del modello nel caso in cui l'assicurazione sia Medicare o Medicaid.

Il modello può, potenzialmente, essere applicato a tutti i pazienti con un rischio basso di discriminare in base allo status socio-economico.

Assicurazione	**Accuratezza**	**AUC-ROC**	**MCC**	**F1**
Others (Insurance 0)	97%	0.57	0.24	0.22
Medicaid/Medicare (Insurance 1)	98%	**0.65**	**0.41**	**0.39**
Private (Insurance 2)	96%	0.58	0.27	0.25

Tabella 5.7: Risultati delle metriche per assicurazione

5.3. Analisi SHAP e confronto con i modelli di riferimento

Mortalità a 90 giorni (d90d)

Nel modello XGB per la mortalità a 90 giorni, come illustrato in figura 4.7, emergono diverse variabili critiche. Tra queste, il conteggio dei linfociti (LymC) riveste un ruolo importante: un suo decremento segnala una compromissione della risposta immunitaria, associata a un incremento significativo del rischio di morte. Anche il livello di lattato (lactate_813), quando elevato, indica stress metabolico e possibili stati di shock, orientando la predizione verso esiti negativi. La pressione arteriosa sistolica (sbp), se bassa, evidenzia instabilità emodinamica, contribuendo a un ulteriore aumento del rischio di morte, mentre la riduzione della saturazione d'ossigeno (spo2) suggerisce una compromissione dell'ossigenazione, favorendo anch'essa esiti sfavorevoli. A questi indicatori si affiancano biomarcatori di disfunzione renale e infiammatoria, come la creatinina, l'albumina e la CRP (crp_fdv), nonché il rapporto neutrofili/linfociti (nl_ratio). Inoltre, parametri clinici fondamentali quali l'età (age), il BUN, l'indice OASIS, la Glasgow Coma Scale (GCS) e il livello di emoglobina (hemoglobin) completano il quadro diagnostico, fornendo una base per la valutazione dello stato del paziente. I force plot relativi a d90d presenti in immagine 4.8 e 4.9 mostrano chiaramente come valori critici, ad esempio lattato elevato o sbp bassa, possano spostare la predizione del modello verso un esito negativo.

Mortalità in ospedale (dinhosp)

Nel modello XGB per la mortalità in ospedale l'analisi SHAP delinea un quadro diagnostico in parte diverso, come evidenziato in figura 4.11. In questo contesto la CRP risulta particolarmente rilevante poiché livelli elevati di questo marcatore infiammatorio innalzano sensibilmente la probabilità di un esito negativo. Valori anomali di glucosio e bicarbonato forniscono indicazioni importanti sullo stato metabolico, giocando un ruolo decisivo nella predizione dell'esito. Un incremento della creatinina segnala possibili disfunzioni renali, mentre squilibri elettrolitici, per esempio variazioni nei livelli di calcio e cloruro, insieme al rapporto neutrofili/linfociti completano il quadro diagnostico, aiutando a selezionare le variabili diagnostiche rilevanti nella predizione di un esito di un paziente soggetto allo studio. Anche i parametri di base, come età (age), BUN, indice OASIS, GCS ed emoglobina (hemoglobin), forniscono un contributo fondamentale nella valutazione del rischio del paziente. I force plot riferiti a dinhosp, mostrati in figura 4.12 e figura 4.13, illustrano come variazioni significative dei parametri critici spostino la predizione verso un esito più critico, chiarendo l'impatto di ciascuna variabile sul risultato finale.

Confronto con i modelli di Mollura e Wang

Confrontando i nostri risultati con gli studi di Mollura e Wang, si osserva che: Mollura [5] ha evidenziato che, nei pazienti con sepsi, i principali predittori includono l'APACHE II, il mean platelet volume (MPV), il conteggio degli eosinofili (EoC) e la CRP, mentre per i pazienti con SIRS sono risultati critici il SOFA, l'APACHE II, il conteggio piastrinico e la CRP. Il nostro modello conferma il ruolo centrale della CRP, ma amplia il quadro includendo marker di stress metabolico (lattato, glucosio, bicarbonato) e squilibri elettrolitici (calcio, cloruro) che non sono stati enfatizzati nello studio di Mollura. La differenza sostanziale è data dai diversi dataset: pochi pazienti ma valori continui per Mollura, mentre molti pazienti ma con pochi valori nel nostro caso.
Wang [4] ha utilizzato l'analisi di SHAP per evidenziare l'importanza in particolare di parametri emodinamici, oltre che all'importanza di parametri comuni ai nostri quali GCS, BUN e età. I nostri risultati si allineano con quelli di Wang, evidenziando l'importanza di sbp, spo2, creatinina e nl_ratio. Tuttavia, il nostro modello sottolinea ulteriormente l'impatto dei parametri immunitari (LymC) e dello stress metabolico (lattato), contribuendo a una stratificazione del rischio più dettagliata.

I risultati suggeriscono come il nostro modello predittivo si confronti favorevolmente con i benchmark impostati negli studi di Mollura e Wang: il nostro modello, infatti, sottolinea l'importanza dei marker diagnostici tradizionali per la sepsi (quali BUN, GCS, CRP), integrandoli con altri indicatori di stress metabolico che consentono una migliore predizione per i pazienti critici presenti nel dataset.

5.4. Innovazioni del Progetto

Il nostro lavoro si distingue dalla letteratura esistente grazie a una serie di innovazioni metodologiche e applicative che ne migliorano la robustezza, la trasparenza e l'applicabilità clinica.

5.4.1. Ottimizzazione degli Iperparametri

In primo luogo, è stato implementato un processo di ottimizzazione degli iperparametri articolato in più fasi. A differenza degli studi analizzati nello Stato dell'Arte, che si basano esclusivamente su una grid search o una random search isolata, il nostro approccio integra:

1. RandomizedSearchCV: una fase iniziale che consente un'esplorazione ampia e casuale dello spazio degli iperparametri, evitando la computazione esaustiva di tutte le possibili combinazioni.
2. Ottimizzazione Bayesiana tramite Optuna: una fase intermedia che indirizza la ricerca verso le configurazioni più promettenti sfruttando un modello probabilistico dello spazio dei parametri.
3. GridSearchCV Mirata: una fase finale di analisi di intorni di variabili già ottimizzate mediante RandomizedSearchCV e ottimizzazione mediante Optuna, finalizzata ad affinare ulteriormente i valori ottimali ottenuti, garantendo modelli più robusti e con una migliore capacità di generalizzazione, e riducendo il rischio di overfitting.

5.4.2. Analisi del Bias Socio-economico

Il progetto dedica particolare attenzione all'analisi del bias socio-economico e demografico. Considerando che la presenza di bias nelle predizioni può compromettere l'affidabilità nell'uso clinico dei modelli, sono state adottate tecniche per: analizzare i pattern dei missing values per identificare eventuali correlazioni sistematiche, verificando che le predizioni non siano influenzate in maniera influente da bias nella raccolta dei dati, valutare l'impatto delle variabili socio-economiche (tipo di assicurazione) e demografiche (come etnia e genere) sulle performance predittive. Infine è stata effettuata una "forzatura" delle variabili sensibili, creando scenari per verificare la stabilità delle predizioni dei modelli al variare di tali fattori demografici.

5.4.3. Utilizzo del Database MIMIC-IV

Un ulteriore elemento distintivo del nostro studio è l'utilizzo del database MIMIC-IV. Diversamente dagli studi precedenti, che si sono basati su MIMIC-III, MIMIC-IV offre dati più aggiornati. Ciò consente di beneficiare di una maggiore quantità di informazioni cliniche, contribuendo a migliorare la validità e la generalizzabilità dei risultati ottenuti. Inoltre, fornisce una base per studi futuri sullo sviluppo di modelli di predizione di mortalità in pazienti provenienti dal database MIMIC-IV.

5.4.4. Ampliamento della Coorte

Il nostro lavoro si caratterizza anche per l'ampliamento della coorte di pazienti. Mentre studi come quelli di Mollura [5] e Wang [4] si sono concentrati rispettivamente ed esclusivamente sui pazienti con sepsi e SIRS. Nel nostro studio abbiamo esteso l'analisi includendo pazienti diagnosticabili con entrambe le patologie, portando a un dataset di pazienti settici e affetti da SIRS; questa scelta ha permesso di incrementare il numero complessivo di pazienti analizzati, considerando una gamma più ampia di condizioni cliniche, rendendo i modelli predittivi applicabili a scenari più variegati e rappresentativi della realtà clinica.

6 | Conclusioni e Sviluppi Futuri

L'analisi comparativa dei modelli di machine learning applicati alla predizione della mortalità in pazienti critici ha evidenziato la superiorità di XGBoost rispetto agli altri approcci considerati. Nello specifico, XGBoost si è dimostrato il modello più adatto sia per la predizione della mortalità a 90 giorni dall'ammissione in terapia intensiva, sia per quella alla fine dell'ospedalizzazione. Tale modello riesce infatti a bilanciare in modo ottimale recall e precisione, minimizzando il rischio di falsi negativi, aspetto cruciale in ambito clinico.

I risultati ottenuti evidenziano che XGBoost garantisce un recall elevato (0.70 a 90 giorni, 0.84 alla fine dell'ospedalizzazione), fondamentale per l'identificazione tempestiva dei pazienti ad alto rischio. Sebbene LightGBM abbia raggiunto un'accuratezza leggermente superiore (0.82) nella predizione a 90 giorni, il suo recall inferiore lo rende meno efficace dal punto di vista diagnostico. Analogamente, la Logistic Regression, seppur più interpretabile, ha mostrato prestazioni predittive inferiori.

Interpretabilità mediante SHAP

L'analisi delle variabili più influenti, condotta tramite SHAP, ha confermato che XGBoost basa le proprie decisioni su indicatori clinicamente rilevanti. In particolare, per la predizione a 90 giorni il modello risulta fortemente influenzato da parametri quali la conta dei linfociti, il lattato sierico, la pressione arteriosa sistolica (SBP) e la saturazione ossigenica (SpO2), mentre per la predizione alla fine dell'ospedalizzazione le decisioni del modello si fondano principalmente su variabili come la CRP (proteina C reattiva), il glucosio, il bicarbonato, la creatinina e gli squilibri elettrolitici. Tali risultati, pienamente in linea con la letteratura clinica, rafforzano la validità del modello nel riconoscere i fattori di rischio associati a esiti negativi.

Bias Socio-Economico e Demografico

L'analisi del bias ha evidenziato che XGBoost non venga influenzato significativamente da genere, etnia o tipo di assicurazione. L'unica discrepanza osservata riguarda i pazienti

coperti da Medicare/Medicaid, che hanno ottenuto metriche leggermente migliori, probabilmente dovute al maggior numero di valori disponibili nel caso di questo tipo assicurativo, aumentando le metriche del modello per questa classe assicurativa rispetto alle altre.

Confronto con Studi Precedenti

Confrontando i nostri risultati con lo studio di Mollura [5] si osserva come il nostro modello confermi il ruolo centrale della CRP, ma ampli il quadro includendo marker di stress metabolico (lattato, glucosio, bicarbonato) e squilibri elettrolitici (calcio,cloruro), non rilevanti nello studio di confronto. Questa differenza è data dai diversi dataset, molti pazienti ma pochi valori nel nostro caso, pochi pazienti ma molti valori per Mollura. Nel confronto con Wang [4] si osserva come il nostro abbia in comune con la referenza parametri emodinamici e biomarcatori di disfunzione d'organo, tuttavia il nostro studio evidenzia maggiormente il ruolo di parametri immunitari come LymC e dello stress metabolico. Il nostro modello predittivo si confronta favorevolmente con gli studi di Mollura e Wang: esso conferma l'importanza dei marker diagnostici tradizionali per la sepsi, integrandoli con ulteriori indicatori che consentono una migliore stratificazione del rischio nei pazienti critici.

Conclusioni

In conclusione, XGBoost ha dimostrato di essere il modello migliore per le predizioni di mortalità in pazienti con sepsi o SIRS , in particolare per la riduzione dei falsi negativi rispetto sia a Logistic Regression che a LightGBM. Un alto numero di falsi negativi, infatti, porta a sottodiagnosticare i pazienti, mentre un numero elevato di falsi positivi indica una maggiore spesa sanitaria sacrificando la correttezza delle diagnosi e aumentando il numero di falsi allarmi. Il modello LightGBM rimane comunque attendibile in caso di predizione di mortalità, nonostante abbia una minore accuratezza diagnostica causata dal maggior numero di falsi negativi, ma potrebbe essere utilizzabile per ridurre i costi di gestione ospedaliera. L'analisi del bias socio-economico non ha evidenziato criticità significative all'interno del modello. Le minime differenze delle curve di distribuzione di probabilità non appartengono al modello, ma dipendono da preesistenti comorbidità. XGBoost ha delle metriche migliori per i modelli assicurativi Medicaid/Medicare per la presenza di meno valori mancanti.

Sviluppi futuri

L'obiettivo futuro è quindi quello di diminuire sia il numero di falsi positivi che di falsi negativi attraverso diversi metodi: una migliore accuratezza nell'inserimento delle variabili misurate in modo da avere una più precisa rappresentazione dei casi clinici all'interno del dataset e una coorte selezionata da database esterni e cartelle cliniche per ottenere una validazione esterna. Un altro possibile sviluppo sarebbe utilizzare un modello ensemble tra XGB e LGB o altri modelli ML.

Bibliografia

[1] S. et al, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Journal of the American Medical Association*, 2016.

[2] R. et al., "Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the global burden of disease study," *The Lancet*, pp. 200–211, 2020.

[3] B. et al., "Machine-learning models for prediction of sepsis patients mortality," *Medicina Intensiva*, 2023.

[4] W. et al., "Early sepsis mortality prediction model based on interpretable machine learning approach: development and validation study," *Internaland Emergency Medicine*, 2024.

[5] M. et al., "Identifying prognostic factors for survival in intensive care unit patients with sirs or sepsis by machine learning analysis on electronic health records," *PLOS Digital Health*, 2024.

[6] G. et al., "Civetta, taylor, kirby's critical care," *WW Norton Company*, 2009.

[7] C. et al., "Sepsis and septic shock," *The Lancet*, vol. 392, pp. 75–87, 2018.

[8] G. et al., "Sepsis: Diagnosis and management," *American Family Physician*, 2020.

[9] L. V. et al., "The global burden of sepsis and septic shock," *Epidemiologia*, 2024.

[10] Gotts, "Sepsis: pathophysiology and clinical management," *BMJ*, 2016.

[11] F. Guo, "Clinical applications of machine learning in the survival prediction and classification of sepsis: coagulation and heparin usage matter," *Journal of Translation Medicine*, 2022.

[12] L. M. Fleuren, "Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy," *Intensive Care Med*, 2020.

[13] J. et al., "Mimic-iv, a freely accessible electronic health record dataset," *Scientific Data*, 2023.

[14] G. et al., "Physiobank, physiotoolkit, and physionet: components of aa new research resource for complex physiologic signals," *Circulation*, 2000.

[15] S. et al, "A comparison of imputation methods for categorical data," *Informatics in Medicine Unlocked*, 2023.

[16] G. et al., *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.", 2022.

[17] C. et al., "Xgboost: A scalable tree boosting system," *Publication History*, 2016.

[18] S. et al., *Exploring the effect of normalization on medical data classification*. Springer, 2021.

[19] D. Yadolah, *Kolmogorov–Smirnov Test*. Springer New York, 2008.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning*. Information Science and Statistics, New York: Springer, 2006.

[21] B. et a., "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

[22] I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., *A Unified Approach to Interpreting Model Predictions*, vol. 30. Curran Associates, Inc., 2017.

[23] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," 2018.

[24] J. et al., "Race and ethnicity and cardiometabolic risk profile: Disparities across income and health insurance in a national sample of us adults," *Journal of Public Health Management and Practice*, 2022.

[25] S. et al., "Health insurance coverage in the united states: 2014," *U.S. Census Bureau*, 2015.

A | Appendice A

A.1. Strumenti e Pacchetti Utilizzati nel Progetto

Il progetto è stato sviluppato mediante un workflow che prevede l'estrazione, il preprocessing, la modellizzazione e la valutazione dei dati e dei modelli. Per superare i vincoli computazionali dei PC locali, è stato impiegato BigQuery sfruttando SQL per estrarre i dati dal database MIMIC-IV. Di seguito si riportano inoltre i pacchetti Python principalmente utilizzati.

Pacchetti Python Principali

Il codice sfrutta numerosi pacchetti, oltre ai moduli built-in, per garantire l'efficienza dell'intero processo. Di seguito le principali librerie e relative versioni:

- pandas (v2.2.2) – Gestione e analisi dei dati strutturati.
- seaborn (v0.13.2) e matplotlib (v3.10.0) – Creazione di grafici e visualizzazioni.
- numpy (v1.26.4) – Calcolo numerico e gestione di array multidimensionali.
- optuna (v4.2.1) – Ottimizzazione degli iperparametri.
- scikit-learn (v1.6.1) – Preprocessing, validazione incrociata e valutazione dei modelli.
- shap (v0.46.0) – Interpretazione dell'impatto delle variabili predittive.
- xgboost (v2.1.4) e lightgbm (v4.5.0) – Modelli di boosting per classificazione e regressione.
- scipy (v1.13.1) – Operazioni statistiche e test di ipotesi.
- missingno (v0.5.2) – Visualizzazione dei dati mancanti.
- tableone (v0.9.1) – Generazione di tabelle descrittive.
- missingpy (v0.2.0) – Imputazione dei valori mancanti.

A.2. Confronto delle Coorti

Variabile	Nostro studio	Wang et al.	Mollura et al.
Età	✓	✓	✓
Sesso	✓	✓	✓
SOFA Score	✓	✓	✓
SIRS Score	✓	×	×
BMI	✓	×	×
GCS (Glasgow Coma Scale)	✓	✓	×
MAP (Pressione Arteriosa Media)	✓	✓	×
Frequenza Cardiaca	✓	✓	✓
Frequenza Respiratoria	✓	✓	✓
Saturazione O2 (SpO2)	✓	✓	×
Lattato	✓	✓	✓
Proteina C Reattiva (CRP)	×	×	✓
Globuli Bianchi (WBC)	✓	✓	✓
Albumina	✓	×	×
Emoglobina	✓	✓	×
INR	×	✓	×
BUN (Azoto Ureico)	✓	✓	×
Creatinina	✓	✓	✓
Piastrine	✓	✓	✓
Ventilazione Meccanica	✓	✓	×
Uso Vasopressori	✓	✓	✓
Diagnosi ICD	×	✓	×

Tabella A.1: Confronto tra le variabili selezionate nel nostro studio, Wang et al. (2024) [4] e Mollura et al. (2024) [5].

A.3. Distribuzioni stimate

Tabella A.2: Distribuzione e risultati ks test per le variabili

variable	distribution	ks_stat	p_value
bmi	lognorm	0.02	1.71e-06
crp_fdv	beta	0.09	1.74e-04
wbc_fdv	t	0.05	1.29e-58
heart_rate	lognorm	0.02	3.98e-09
sbp	lognorm	0.02	6.32e-10
dbp	lognorm	0.02	9.78e-07
body_temp	t	0.04	1.17e-29
spo2	beta	0.04	8.13e-33
resp_rate	lognorm	0.02	4.46e-08
fio2	beta	0.07	7.18e-80
albumin_bl_chem	weibull_min	0.03	4.21e-07
aniongap	lognorm	0.04	1.99e-38
bun	lognorm	0.05	4.22e-56
calcium_chem	t	0.02	2.35e-11
chloride_chem	t	0.04	2.07e-27
creatinine_chem	lognorm	0.11	1.45e-242
glucose_chem	lognorm	0.05	7.02e-48
sodium_chem	norm	0.08	5.43e-134
potassium_chem	lognorm	0.03	5.69e-24
magnesium_chem	t	0.05	1.21e-46
lactate_813	lognorm	0.03	3.83e-15
bicarbonate	gamma	0.07	1.67e-91
ph	t	0.05	5.91e-41
po2_821	beta	0.02	9.26e-07
pco2_818	lognorm	0.06	6.37e-71
hematocrit	gamma	0.01	5.53e-05
hemoglobin	beta	0.02	2.01e-05
platelet	lognorm	0.02	1.67e-11
neuc	beta	0.05	5.11e-26
lymc	gamma	0.03	5.95e-07
nl_ratio	lognorm	0.03	5.50e-11
age	beta	0.02	1.25e-10
oasis	weibull_min	0.05	1.58e-45

A.4. Griglie di iperparametri

A seguito del processo di ottimizzazione descritto nel capitolo 3.6, sono state individuate le configurazioni ottimali per gli iperparametri dei modelli XGBoost e LightGBM, osservabili nelle griglie presenti in seguito.

Parametro	Valore
max_depth	5
min_child_weight	7
gamma	0.75
subsample	0.8
colsample_bytree	0.6
learning_rate	0.05
scale_pos_weight	2
reg_alpha	1
reg_lambda	3
n_estimators	350
eval_metric	auc

(a) XGBoost

Parametro	Valore
boosting_type	dart
num_leaves	80
max_depth	5
learning_rate	0.05
n_estimators	700
min_data_in_leaf	25
max_bin	127
feature_fraction	0.6
bagging_fraction	0.6
bagging_freq	5
reg_alpha	0.1
reg_lambda	2
min_gain_to_split	0.2

(b) LightGBM

Tabella A.3: Griglie dei parametri per XGBoost e LightGBM

Elenco delle figure

1.1 Flowchart sepsi 5
1.2 Punteggio SOFA 5

3.1 Processo di sviluppo per MIMIC 13
3.2 Struttura modulare del database 14
3.3 Struttura dello studio 16
3.4 Diagramma dei criteri di inclusione 17
3.5 Diagramma della pipeline del preprocessing 27

4.1 Confronto tra matrice di confusione e curva ROC per Logistic Regression (90 giorni) 34
4.2 Confronto tra matrice di confusione e curva ROC per XGBoost (90 giorni) 35
4.3 Confronto tra matrice di confusione e curva ROC per LGBM (90 giorni) . 36
4.4 Confronto tra matrice di confusione e curva ROC per Logistic Regression . 37
4.5 Confronto tra matrice di confusione e curva ROC per XGBoost 38
4.6 Confronto tra matrice di confusione e curva ROC per LGBM 39
4.7 SHAP explainer per XGB d90d 40
4.8 Force plot per pazienti ad alto rischio per d90d 41
4.9 Force plot per pazienti ad alto rischio per d90d 41
4.10 Force plot per pazienti a basso rischio per d90d 41
4.11 SHAP explainer per XGB dinhosp 42
4.12 Force plot per pazienti ad alto rischio per dinhosp 43
4.13 Force plot per pazienti ad alto rischio per dinhosp 43
4.14 Force plot per pazienti a basso rischio per dinhosp 43
4.15 Distribuzione dell'assicurazione sanitaria negli Stati Uniti nel 2022 44
4.16 Distribuzione di probabilità della predizione senza colonne demografiche (grafico a sinistra) e con colonne demografiche (grafico a destra) 45
4.17 Distribuzione di probabilità della predizione sulla base del sesso, dell'etnia e dell'assicurazione del paziente nei 90 giorni post-ricovero 46

4.18 Distribuzione di probabilità della predizione sulla base del sesso, dell'etnia e dell'assicurazione del paziente entro fine ospedalizzazione 46

Elenco delle tabelle

3.1 Statistiche per ospedale e ammissione in UTI 15
3.2 Criteri diagnostici 18
3.3 Criteri di valutazione della funzione d'organo 18
3.4 Criteri diagnostici per Sepsi-3 e SIRS 19
3.5 Criteri di selezione dei pazienti per il dataset finale. 19
3.6 Diagnosi 22
3.7 Genere 22
3.8 Razza 23
3.9 Tipo di Assicurazione 23

4.1 Risultati delle metriche per genere, etnia e assicurazione 47
4.2 Metriche di XGBoost per la predizione della mortalità a 90 giorni e a fine ospedalizzazione. 48
4.3 Rilevanza delle metriche per pazienti assicurati Medicare/Medicaid 49

5.1 Confronto tra le classi a 90 giorni 51
5.2 Confronto metriche a 90 giorni 52
5.3 Confronto Matrici a 90 giorni 52
5.4 Confronto tra matrici di confusione per XGB (dinhosp) e XGB (d90d) . . . 53
5.5 Confronto tra le classi per dinhosp 54
5.6 Confronto metriche globali per dinhosp 54
5.7 Risultati delle metriche per assicurazione 55

A.1 Confronto tra le variabili selezionate nel nostro studio, Wang et al. (2024) [4] e Mollura et al. (2024) [5]. 68
A.2 Distribuzione e risultati ks test per le variabili 69
A.3 Griglie dei parametri per XGBoost e LightGBM 70