UNIVERSITÁ DEGLI STUDI DI MILANO-BICOCCA
Scuola di Economia e Statistica
Corso di laurea Magistrale in
BIOSTATISTICA

APPLICAZIONI NLP E MACHINE LEARNING PER LA VALUTAZIONE DELLA GRAVITÀ DEI PAZIENTI E LA PREVISIONE DI MORTALITÀ. UN'APPLICAZIONE ALLE SCHEDE DI DIMISSIONE OSPEDALIERA.

Relatore: Prof. Paolo Berta

Correlatore: Prof. Lorenzo Malandri

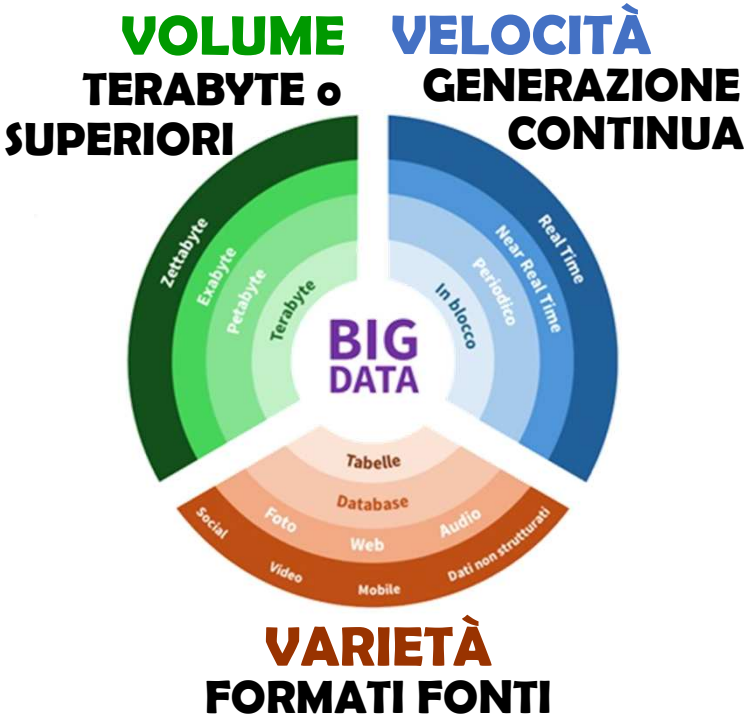
Andrea Millone

Matr. N. 846588

A.A. 2024/2025

INTRODUZIONE

Negli ultimi anni è emersa nel campo della ricerca medica l'importanza dei dati "reali", come quelli provenienti dalle Cartelle Cliniche Elettroniche (EHR).



L'estrazione di conoscenza da questi Big Data caratterizzati dalle 3V – **Volume Velocità e Varietà**

rappresenta una sfida significativa in sanità pubblica per migliorare la qualità delle decisioni cliniche, identificare popolazioni a rischio o prevedere esiti clinici.

Le Cartelle Cliniche Elettroniche contengono **dati STRUTTURATI** e **dati NON STRUTTURATI.**
I dati non strutturati possono arrivare a costituire fino all'80% dei dati EHR.

Schede di Dimissione Ospedaliera (SDO)

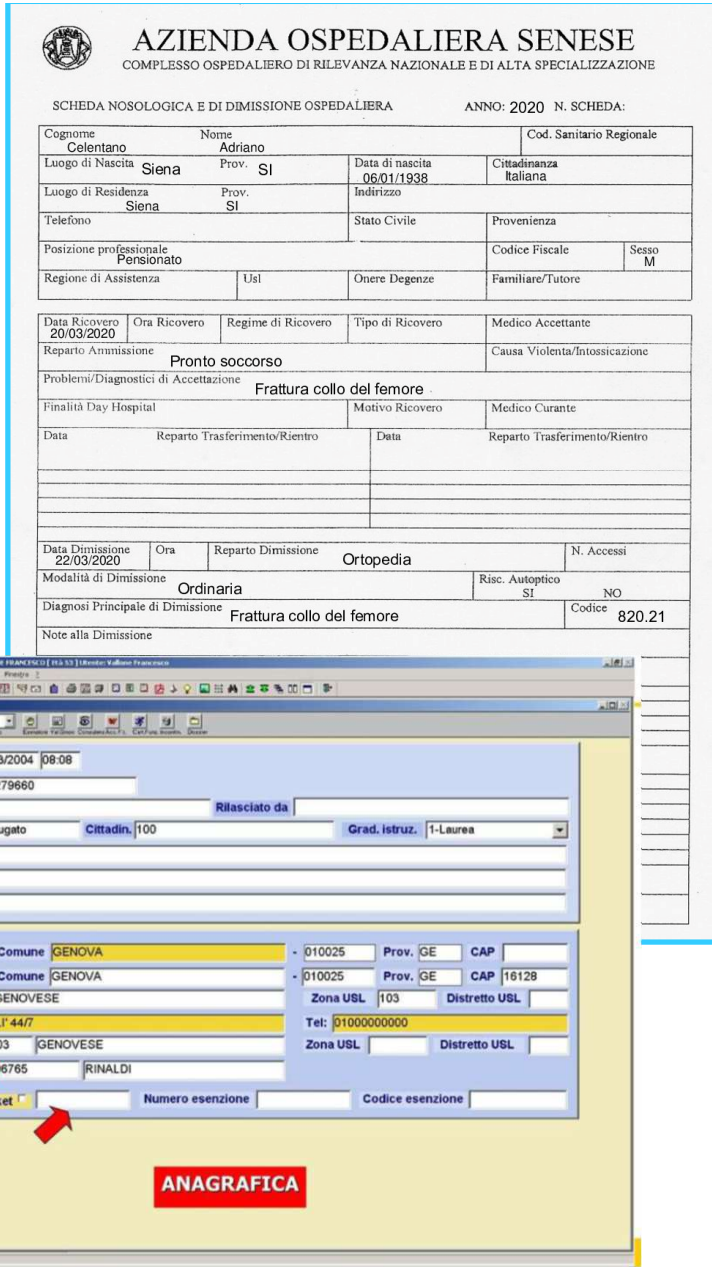
Il database delle SDO italiane raccoglie sistematicamente informazioni relative a tutti gli eventi di ospedalizzazione, nel settore pubblico e privato, rimborsati dal Sistema Sanitario Nazionale (SSN).

SDO italiane contengono :

- informazioni personali del paziente (sesso, età, ecc.)
- informazioni cliniche (diagnosi, procedure chirurgiche, ecc.)
- informazioni relative alla struttura (regione e tipo di ospedale)

L'analisi diretta di queste informazioni è complessa a causa della loro eterogeneità

Per affrontare queste difficoltà, si può ricorrere a tecniche avanzate di Intelligenza Artificiale (IA) come Machine Learning (ML), Deep Learning (DL) e Natural Language Processing (NLP).



AZIENDA OSPEDALIERA SENESE

COMPLESSO OSPEDALIERO DI RILEVANZA NAZIONALE E DI ALTA SPECIALIZZAZIONE

SCHEDA NOSOLOGICA E DI DIMISSIONE OSPEDALIERA — ANNO: 2020 N. SCHEDA:

Cognome: Celentano	Nome: Adriano		Cod. Sanitario Regionale
Luogo di Nascita: Siena	Prov.: SI	Data di nascita: 06/01/1938	Cittadinanza: Italiana
Luogo di Residenza: Siena	Prov.: SI	Indirizzo	
Telefono		Stato Civile	Provenienza
Posizione professionale: Pensionato			Codice Fiscale / Sesso: M
Regione di Assistenza	Usl	Onere Degenze	Familiare/Tutore

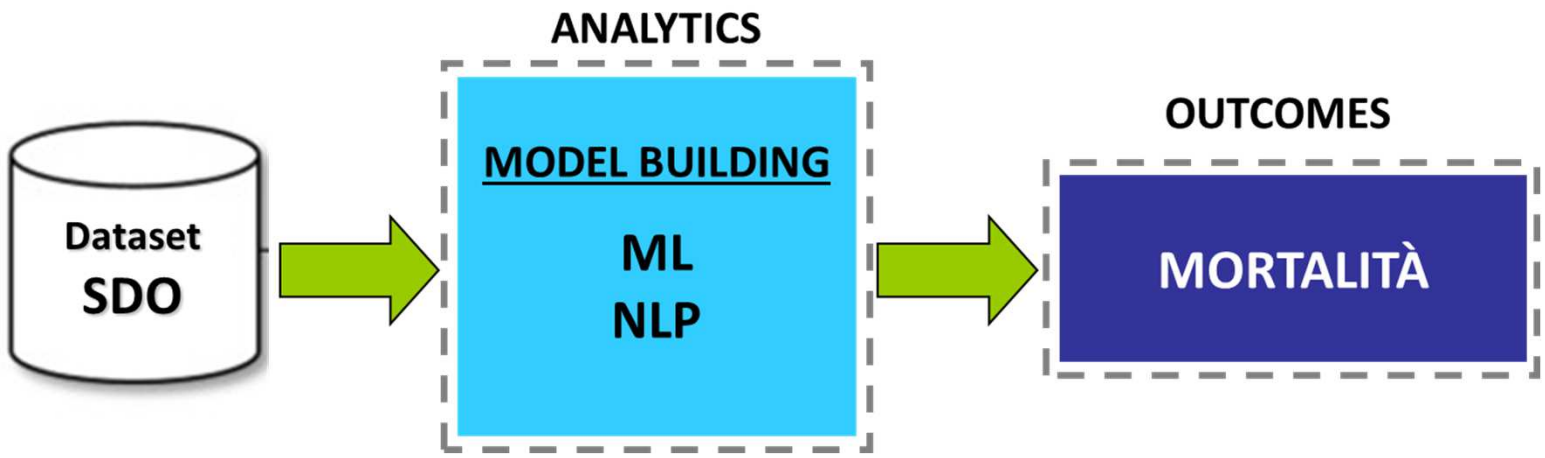
Data Ricovero: 20/03/2020	Ora Ricovero	Regime di Ricovero	Tipo di Ricovero	Medico Accettante
Reparto Ammissione: Pronto soccorso				Causa Violenta/Intossicazione
Problemi/Diagnostici di Accettazione: Frattura collo del femore				
Finalità Day Hospital			Motivo Ricovero	Medico Curante

Data	Reparto Trasferimento/Rientro	Data	Reparto Trasferimento/Rientro

Data Dimissione: 22/03/2020	Ora	Reparto Dimissione: Ortopedia	N. Accessi
Modalità di Dimissione: Ordinaria			Risc. Autoptico: SI / NO
Diagnosi Principale di Dimissione: Frattura collo del femore			Codice: 820.21
Note alla Dimissione			

Obiettivo

- Sviluppare e confrontare differenti algoritmi di Machine Learning (ML) e di Natural Language Processing (NLP) per la previsione della mortalità ospedaliera utilizzando i dati provenienti dal database SDO italiane.



Fonte dei dati

SDO

Cod ID	sex	età	Diagnosi Principale	DIAGNOSI CONCOMITANTI ICD-9-CM					Presenza/assenza 31 comorbilità			INDICE Elixhauser
			CODICE ICD-9-CM	1°	2°	3°	4°	5°	1	2,…	31	Somma (da 1 a 31)
1	F	65	**820.21**	**717.0**	**434.1**				1	0,…	0	3
2	M	55	..									

1° DATASET
DATI STRUTTURATI

ICD-9-CM

2° DATASET
DATI NON STRUTTURATI

Descrizione testuali della diagnosi	Codice ICD-9-CM
Frattura del Femore….	**820.21**

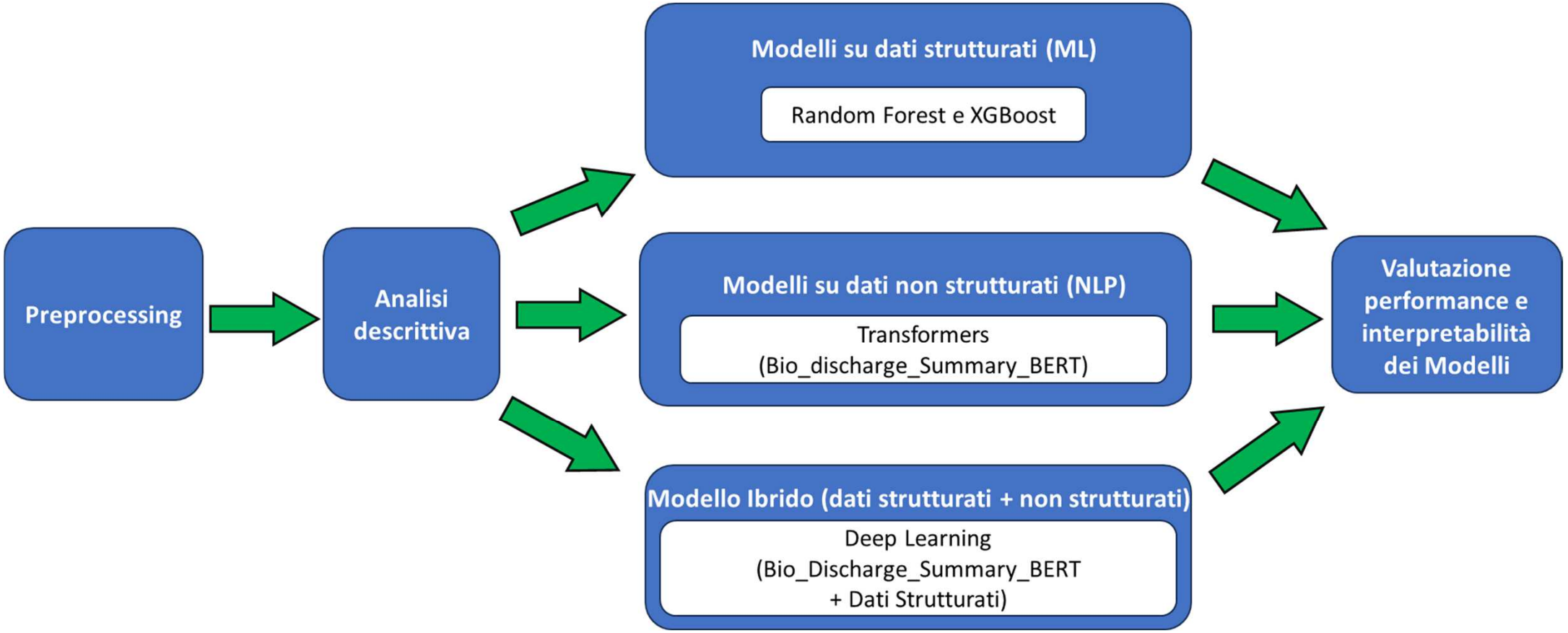
Cod ID	sex	età	Diagnosi Principale	Descrizioni testuali Diagnosi Principale + Diagnosi Concomitanti	31comorbilità + INDICE Elixhauser
			CODICE ICD-9-CM		1…31+ SOMMA
1	F	65	**820.21**	Frattura del Femore…	1,0,0…1 **3**
2	M	55	….		

DATASET RELAZIONALE

DATI STRUTTURATI + DATI NON STRUTTURATI

***Dati SDO circa 3,5 milioni di ricoveri provenienti da tutte le regioni italiane relative al periodo 2012–2016**

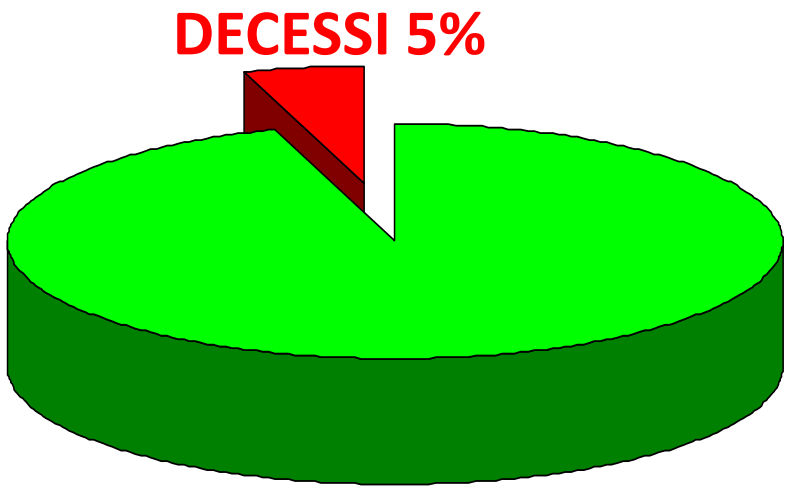
Analisi dei dati



Le analisi sono state svolte con risorse Google Coolab pro +

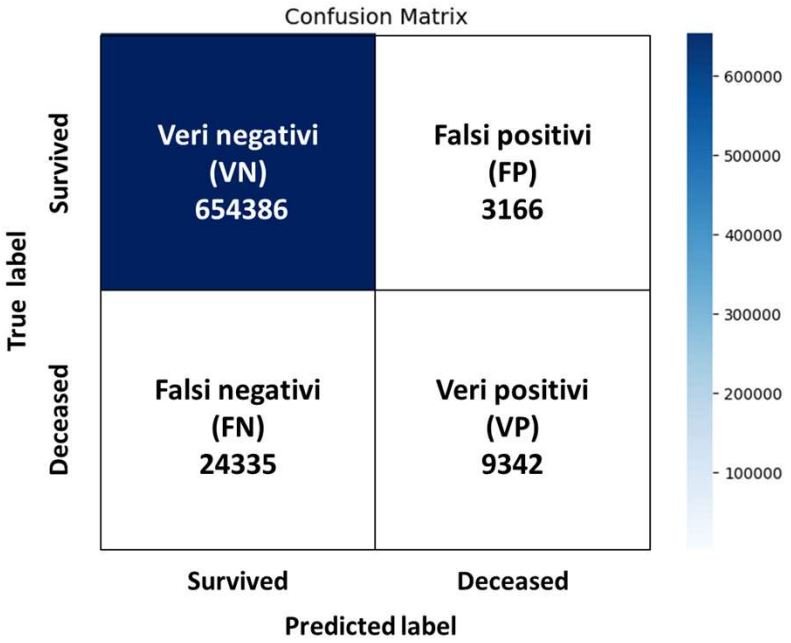
Caratteristiche del campione

	Totale N=3.453.570	Decessi N=168.395	Sopravvisuti N=3.285.175
Femmine (%)	44,63	39,73	44,89
Maschi (%)	55,36	60,27	55,11
Età—anni medi (DS)	68,2 (13,02)	73,02 (12,61)	67,93 (12,99)
Durata mediana della degenza ospedaliera (Q1-Q3)	7 (4-12)	9 (3-18)	7 (4-12)
Punteggio Elixhauser, media (DS)	1,39 (1,49)	2,18 (1,54)	1,35 (1,47)
Numero parole diagnosi, media (DS)	15,01 (9,62)	20,76 (9,36)	14,7 (9,53)



Comorbilità	Totale N=3.453.570	Decessi N=168.395	Sopravvisuti N=3.285.175
aritmie cardiache	5,41%	8,46%	5,26%
insufficienza cardiaca congestizia	3,78%	9,26%	3,50%
perdita di peso	1,42%	12,52%	0,85%
cancro metastatico	16,51%	41,05%	15,26%
tumore solido senza metastasi	38,52%	63,50%	37,22%

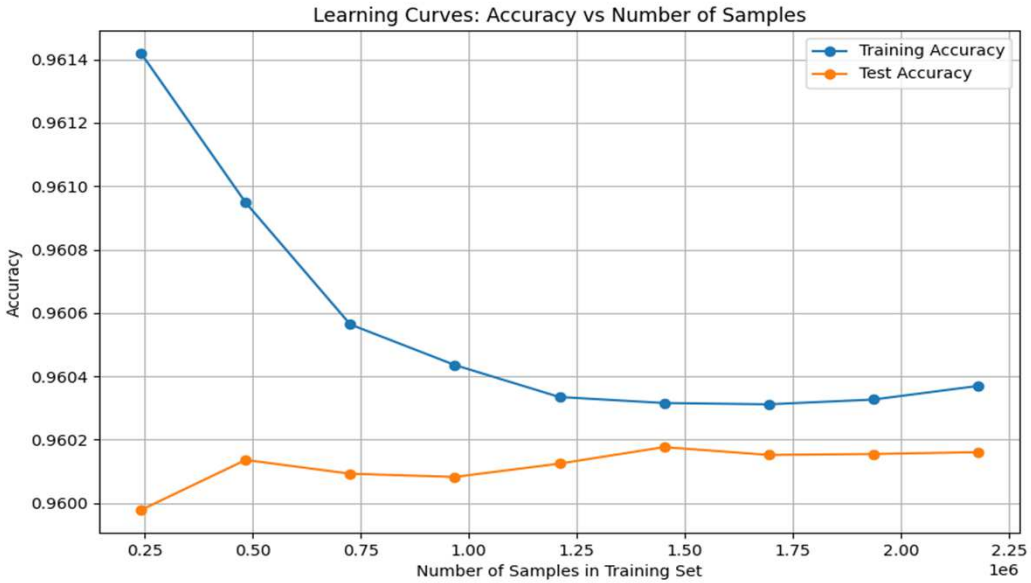
MODELLI SU DATI STRUTTURATI (ML): Random Forest

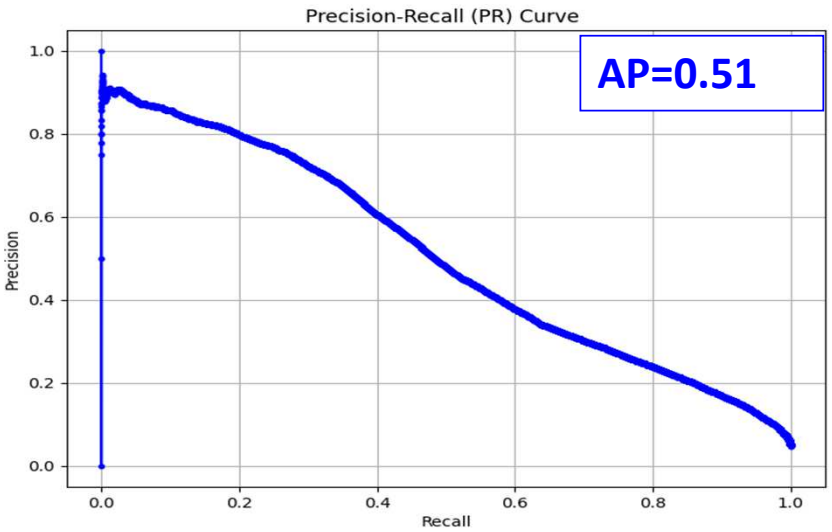


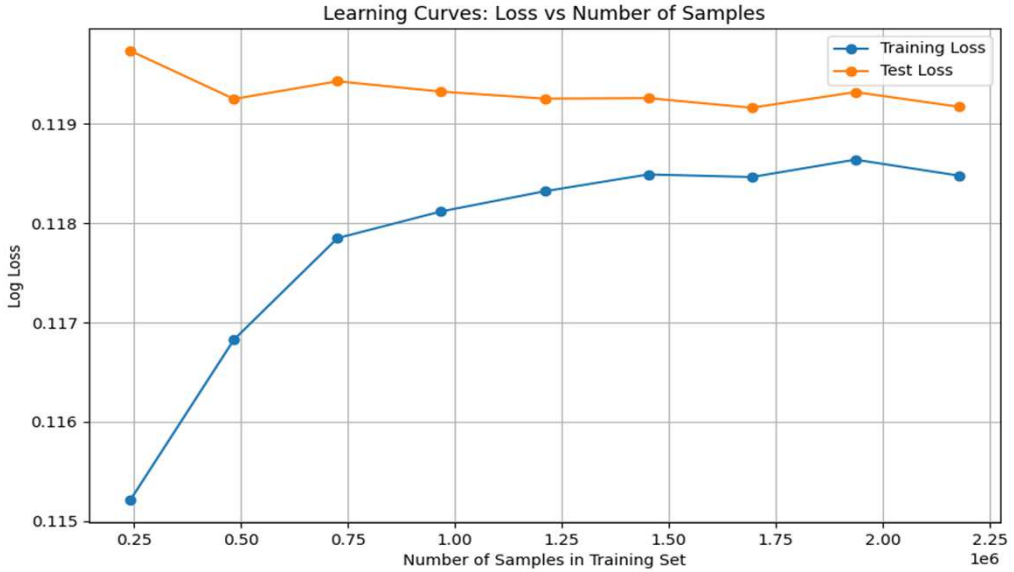
Acc=96%

Spec=99%

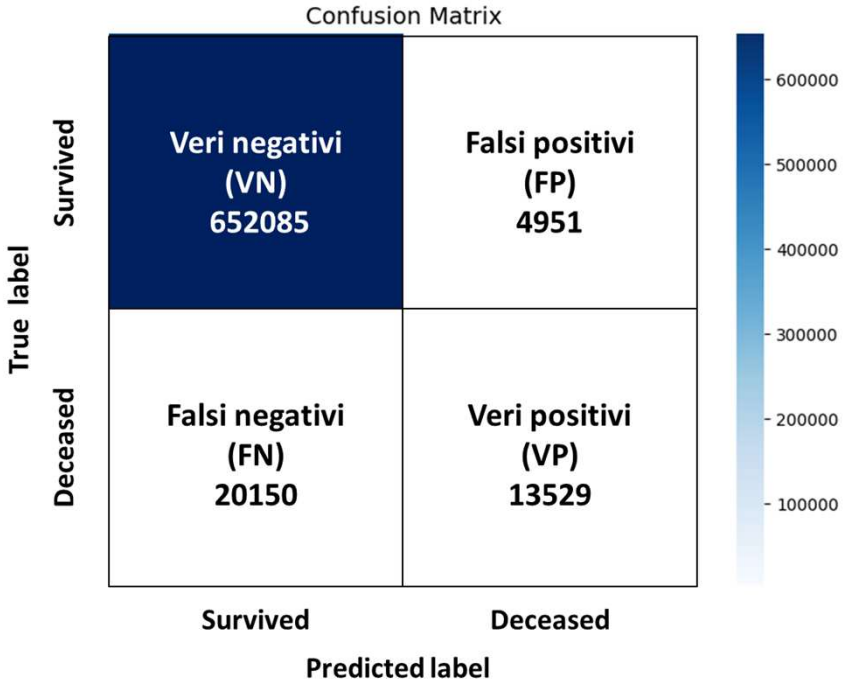
Recall=28%







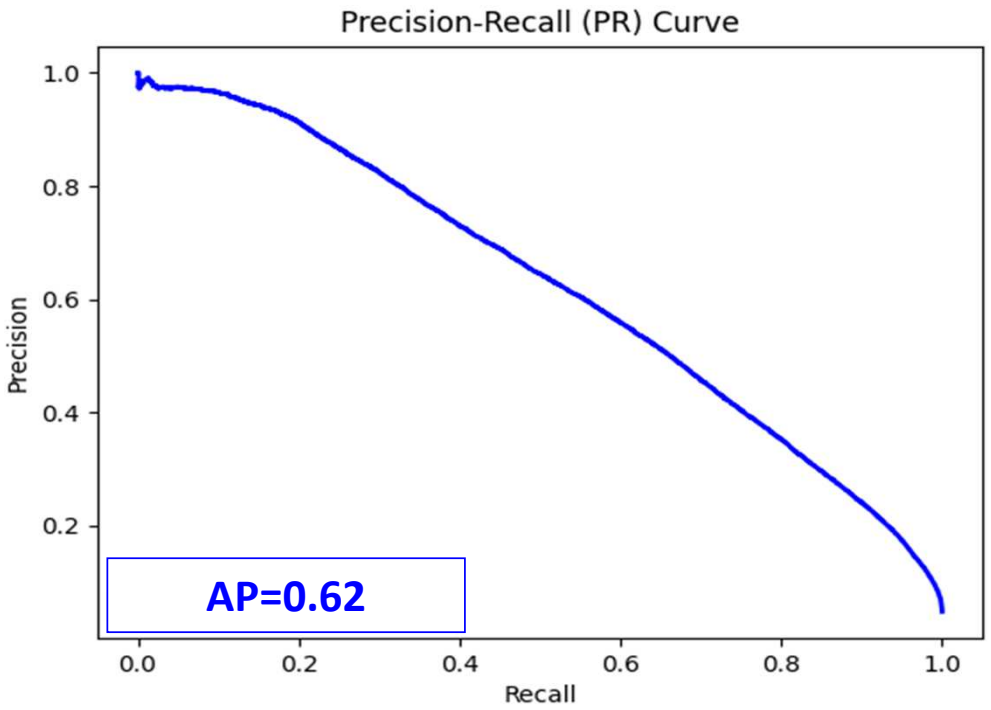
MODELLO SU DATI NON STRUTTURATI (NLP): Bio_Discharge_Summary_BERT *



Acc=96%

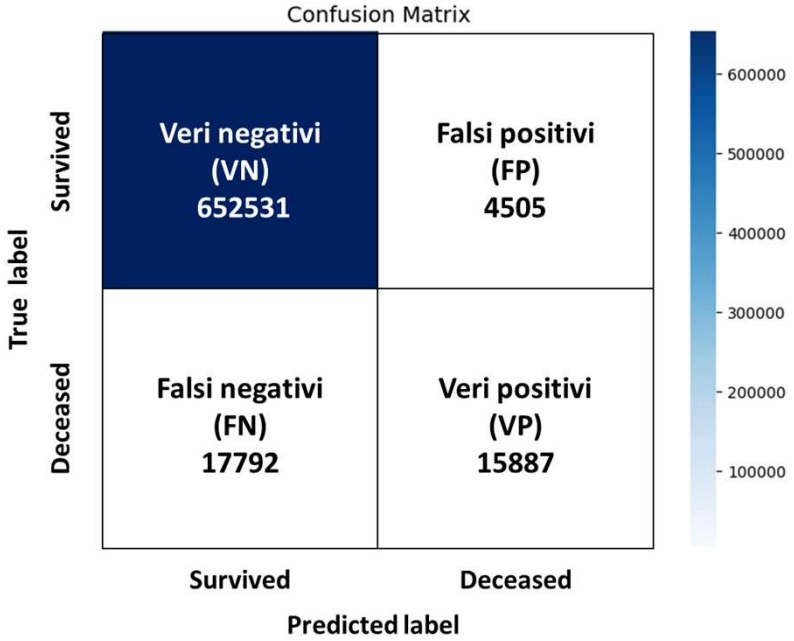
Spec=99%

Recall=41%



*Il modello Bio_Discharge_Summary_BERT è stato pre-addestrato su un corpus specifico di schede di dimissione clinica estratte dal dataset MIMIC-III.

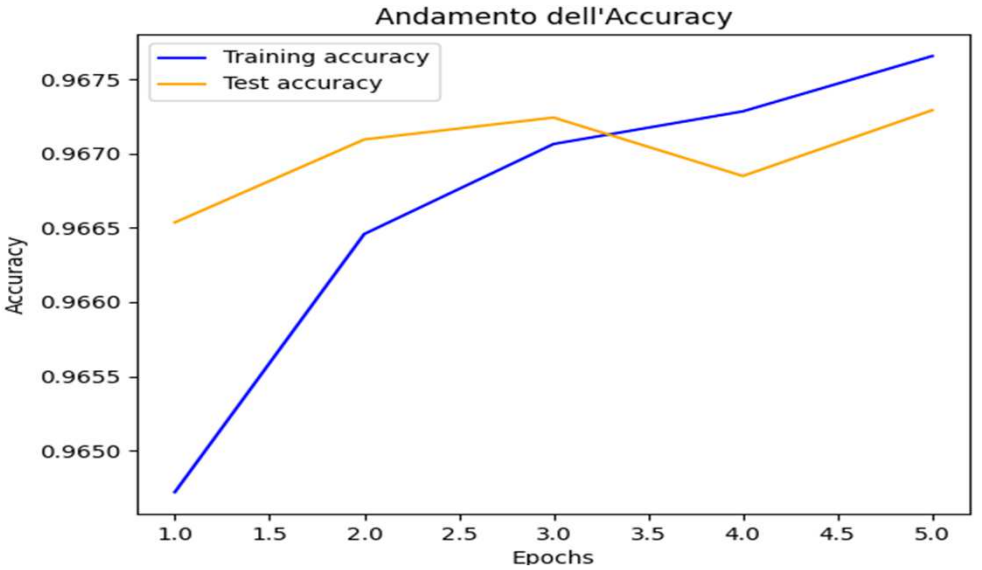
MODELLO IBRIDO (DATI STRUTTURATI + DATI NON STRUTTURATI)

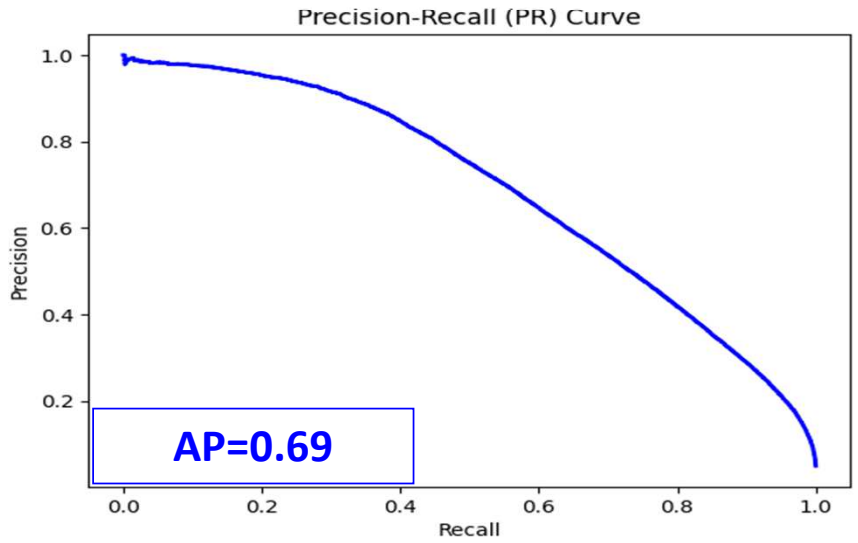


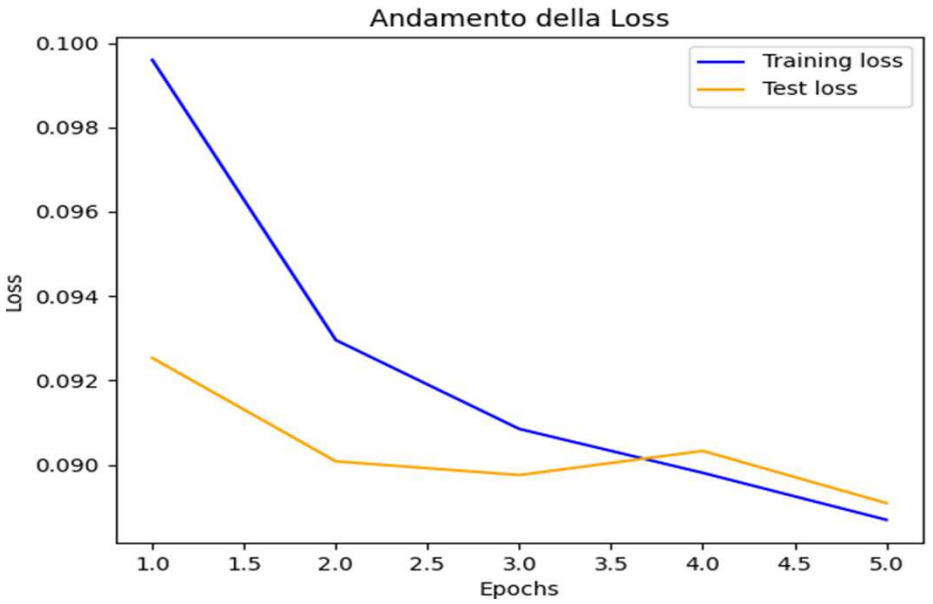
Acc=97%

Spec=99%

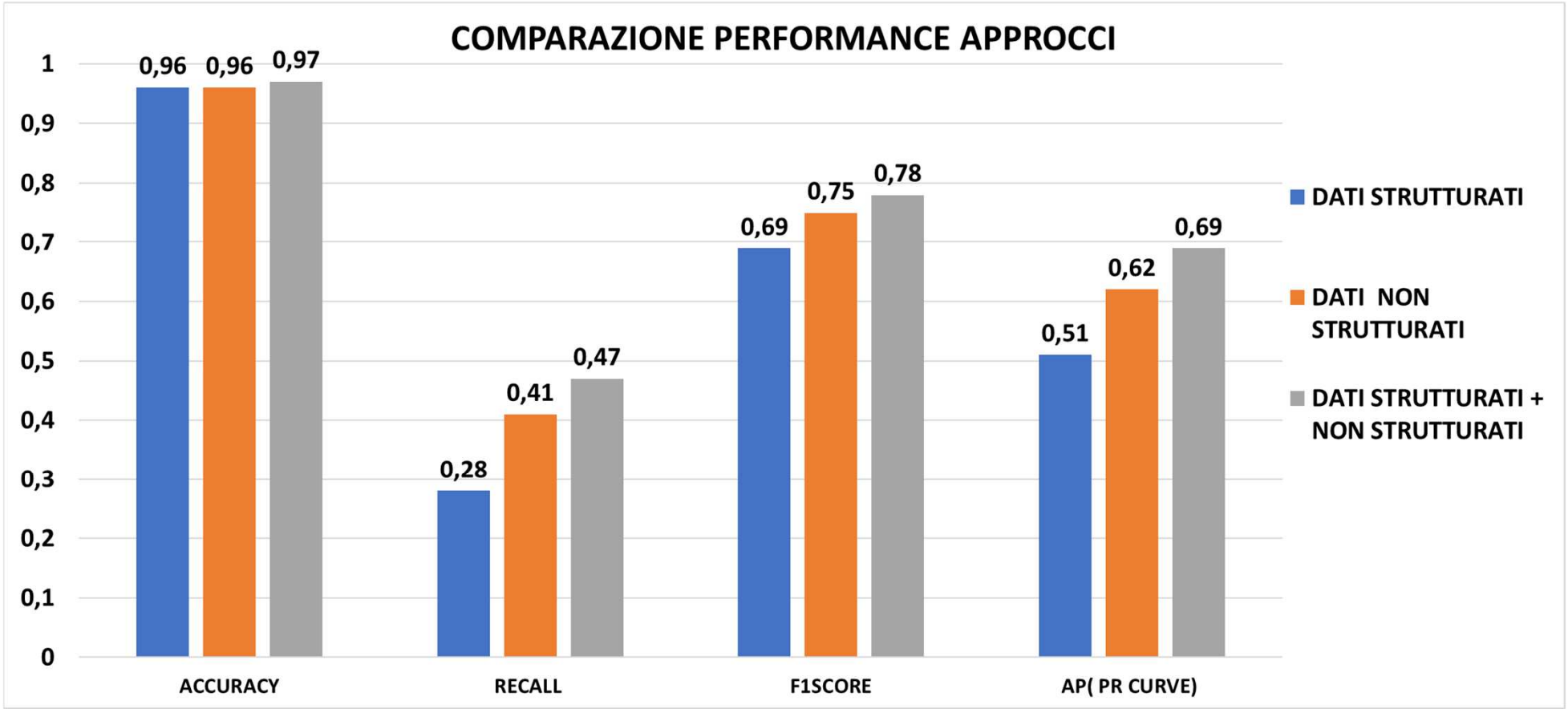
Recall=47%







SINTESI DEI RISULTATI



Dati Strutturati: Modello Random Forest con le variabili strutturate (eta , disciplina ospedaliera, durata del ricovero, 4 comorbilità e indice Elixhauser) selezionate tramite LASSO, ottimizzato con split dt stratificato (train 80% test 20%) cross-validation a 3 fold e **Bayesian Optimization: N° di alberi** 250, **Profondità massima dell'albero** 10, **Minimo n° di campioni per split** 2, **Minimo n° di campioni per foglia** 3.

Dati Non Strutturati: Modello preaddestrato Bio_Discharge_Summary_BERT con la variabile non strutturata (descrizione testuale delle diagnosi), ottimizzato con split dt stratificato (train 80% test 20%), **Tokenizzazione:** Max_length = 90, Padding = Max_length, Truncation = True, **Batch size** 32, 1 **epoca**, ottimizzatore **AdamW**, l**earning rate** 1×10^{-5} e **Weight decay** 0.01.

Modello Ibrido (Dati Strutturati + Non Strutturati): Il modello ibrido ha combinato la variabile non strutturata (descrizione testuale delle diagnosi) e le variabili strutturate (durata del ricovero, sesso, età, regione, disciplina ospedaliera, indice Elixhauser e 31 comorbilità), è stato ottimizzato con split dt stratificato (train 80% test 20%), **Tokenizzazione:** Max_length = 90, Padding = Max_length , Truncation = True, **Bayesian Optimization: Batch size** 32, 5 **epoche**, **Adam**, **dropout** 0.2, l**earning rate** $2{,}72\times10^{-4}$ **Lstm_units** 128, **dense_units** 32 e **batch normalization**.

Interpretabilità tramite XAI : LIME

Vero Positivo

Classe	Prediction Probabilities
SURVIVED	0,39
DECEASED	0,61

Parola	Valore Importanza LIME	Tipo Contributo
CACHEXIA	0,4	POSITIVO
MALIGNANT	0,09	POSITIVO
BREAST	0,08	POSITIVO
LIVER	0,06	POSITIVO
NEOPLASM	0,04	POSITIVO
JAUNDICE	0,04	NEGATIVO
UNSPECIFIED	0,03	POSITIVO

Testo pulito: *malignant neoplasm breast female unspecified malignant neoplasm liver secondary malignant neoplasm bone marrow jaundice unspecified newborn cachexia*

Falso Positivo

Classe	Prediction Probabilities
SURVIVED	0,28
DECEASED	0,72

Parola	Valore Importanza LIME	Tipo Contributo
CACHEXIA	0,45	POSITIVO
FAILURE	0,16	POSITIVO
ACUTE	0,11	POSITIVO
MALIGNANT	0,11	POSITIVO
UPPER	0,05	NEGATIVO
LUNG	0,04	POSITIVO
LOBE	0,04	NEGATIVO

Testo pulito: *malignant neoplasm upper lobe bronchus lung cachexia acute respiratory failure*

Punti di Forza e Limiti

Punti di forza :

- Le analisi, condotte su 3,5 milioni di osservazioni provenienti dalle SDO italiane, garantiscono robustezza e generalizzabilità dei risultati grazie a un solido approccio statistico che riduce il rischio di overfitting.
- L'applicazione di tecniche di Deep Learning (DL) ha permesso di integrare in modo efficace dati strutturati e non strutturati, sfruttando avanzate rappresentazioni semantiche delle diagnosi e le informazioni contenute nelle SDO.

Limiti :

- Nonostante il grande volume di dati , lo sbilanciamento della variabile target, con una netta disparità tra sopravvissuti e deceduti (95% vs. 5%), ha parzialmente limitato l'apprendimento dei modelli e la capacità predittiva.
- I modelli avanzati, come *Bio_Discharge_Summary_BERT* e le architetture ibride, richiedono elevate risorse computazionali, rendendone l'applicazione meno praticabile in contesti con capacità limitate.
- Nonostante l'impiego di LIME, i modelli di DL rimangono complessi e difficili da interpretare, il loro processo decisionale, spesso definito "scatola nera", non può essere osservato né interpretato direttamente.

Conclusioni

- L’approccio sperimentale attraverso l’uso di tecniche classiche e innovative di ML e NLP ha dimostrato che il modello ibrido che combina dati strutturati e non strutturati garantisce le performance migliori nel predire il rischio clinico.

- I risultati di questo studio suggeriscono che l’uso di tecniche di Intelligenza Artificiale applicate ai Big Data sanitari possa migliorare le decisioni cliniche, fornendo un efficace strumento di supporto per i professionisti della salute.