

Stroke Prediction

Predizione del rischio di Ictus con il Machine Learning

 **Alessio Inglese** - 0512118760

Adriano De Vita - 0512117726

Università degli Studi di Salerno

Anno Accademico 2024/25

Definizione del Problema

1 Problema

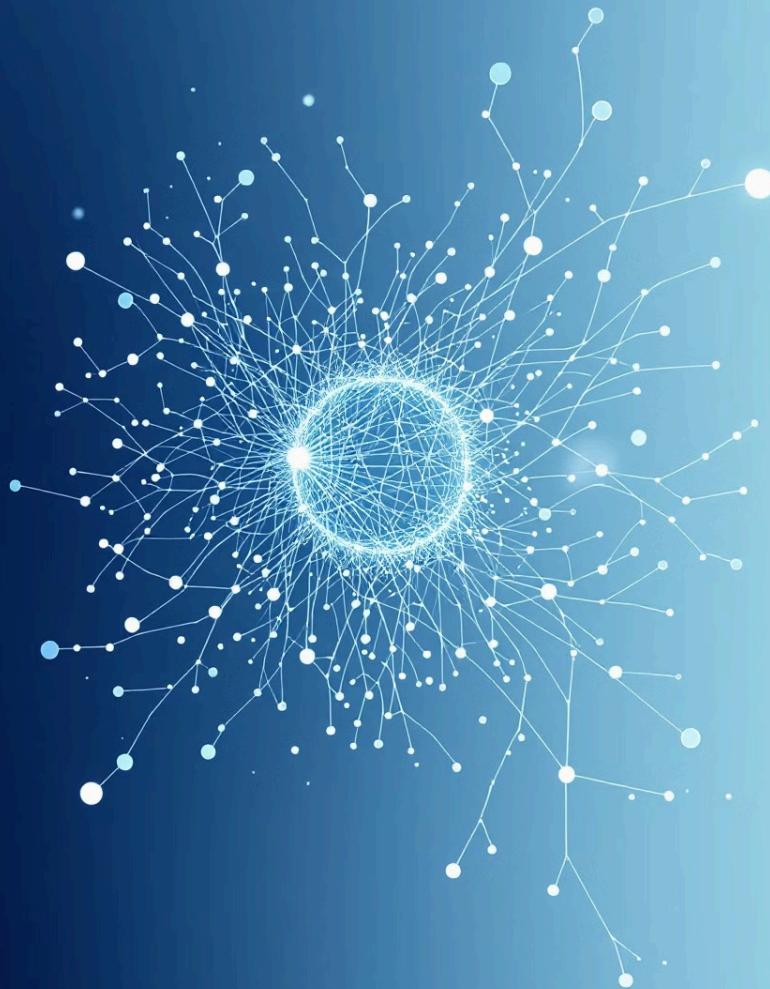
Gli **ictus** sono una delle principali cause di mortalità nel mondo.

Una diagnosi precoce può salvare vite.

2 Sfide principali

- **Dati sbilanciati:** I casi di ictus sono molto meno numerosi rispetto ai non-ictus.
- **Eterogeneità dei pazienti:** Variabili cliniche, comportamentali e demografiche diverse.





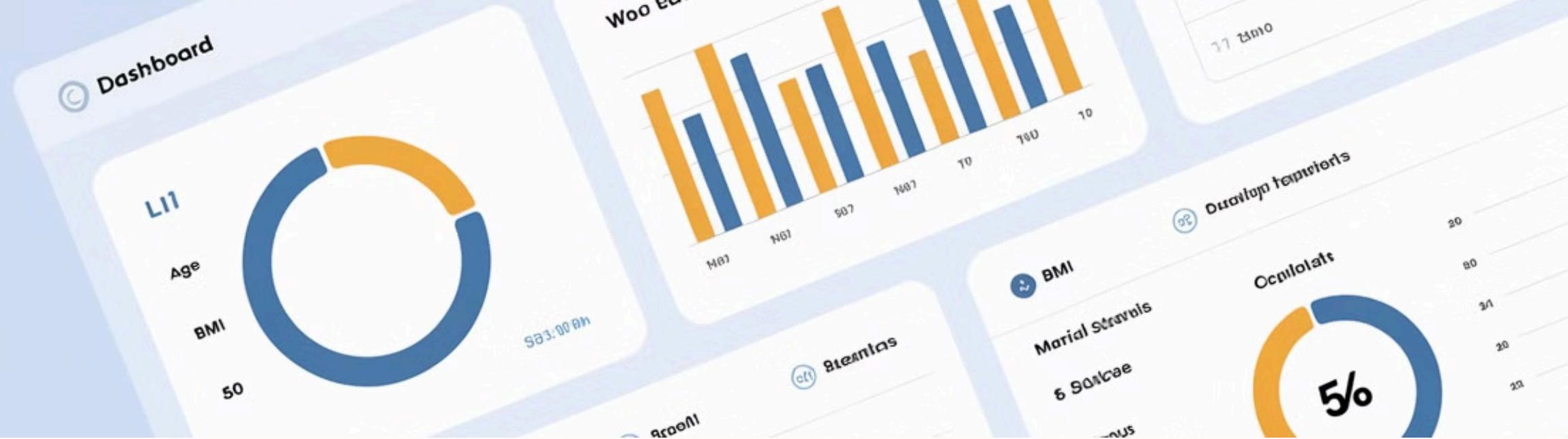
Scopo del Progetto

Obiettivo

Utilizzare il **Machine Learning** per prevedere il rischio di **ictus**, migliorando la diagnosi precoce.

Strategia adottata

- Modello di classificazione basato su **Random Forest**, ottimizzato per dati sbilanciati.
- **Pipeline** strutturata con EDA, pre-processing, addestramento e validazione, valutazione.
- Focus su **Recall**, un aspetto cruciale nel contesto medico dove è preferibile ridurre i falsi negativi.



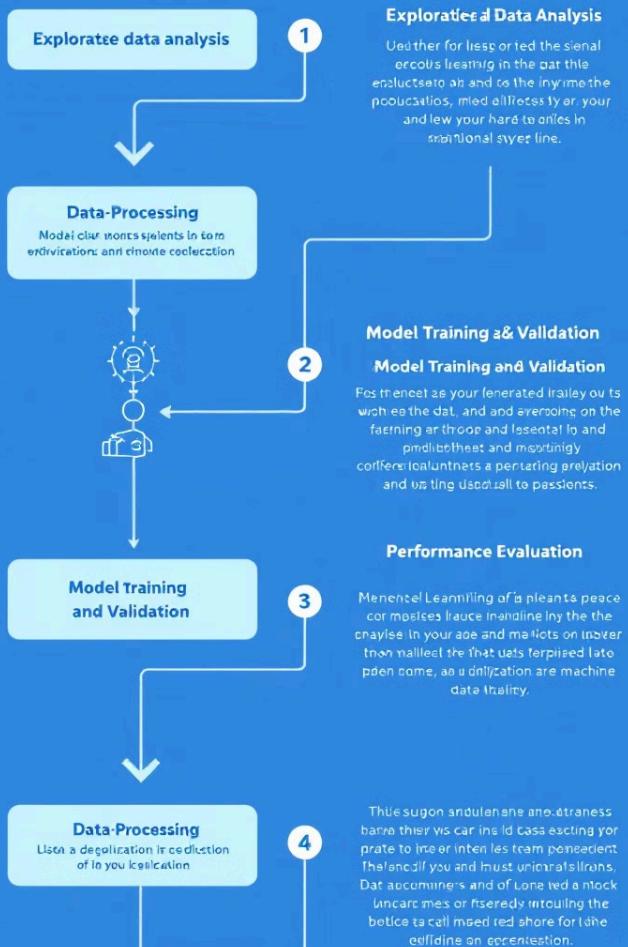
Il Dataset

- Origine:** proviene da **Kaggle – Stroke Prediction Dataset**, contiene informazioni cliniche e demografiche di pazienti.
- Struttura:** 5110 osservazioni e 12 features, sia **numeriche** che **categoriche** (età, BMI, glicemia, stato civile, tipo di lavoro...)
- Sbilanciamento:** solo il **5%** dei campioni rappresenta pazienti con **ictus**, richiedendo tecniche per riequilibrare le classi.

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
32221	Male	60	0	0	1 Yes	Private	Urban	91.92	35.9	smokes	1
10548	Male	66	0	0	0 Yes	Private	Rural	76.46	21.2	formerly smoked	1
52282	Male	57	0	0	0 Yes	Private	Rural	197.28	34.5	formerly smoked	1
45535	Male	68	0	0	0 Yes	Private	Rural	233.94	42.4	never smoked	1
40460	Female	68	1	1	1 Yes	Private	Urban	247.51	40.5	formerly smoked	1
17739	Male	57	0	0	0 Yes	Private	Rural	84.96	36.7	Unknown	1
49669	Female	14	0	0	0 No	children	Rural	57.93	30.9	Unknown	1
27153	Female	75	0	0	0 Yes	Self-employed	Rural	78.8	29.3	formerly smoked	1
34060	Male	71	1	0	0 Yes	Self-employed	Rural	87.8	N/A	Unknown	1
43424	Female	78	0	0	0 Yes	Private	Rural	78.81	19.6	Unknown	1
30669	Male	3	0	0	0 No	children	Rural	95.12	18	Unknown	0
30468	Male	58	1	0	0 Yes	Private	Urban	87.96	39.2	never smoked	0
16523	Female	8	0	0	0 No	Private	Urban	110.89	17.6	Unknown	0
56543	Female	70	0	0	0 Yes	Private	Rural	69.04	35.9	formerly smoked	0
46136	Male	14	0	0	0 No	Never_worked	Rural	161.28	19.1	Unknown	0
32257	Female	47	0	0	0 Yes	Private	Urban	210.95	50.1	Unknown	0
52800	Female	52	0	0	0 Yes	Private	Urban	77.59	17.7	formerly smoked	0
41413	Female	75	0	1	1 Yes	Self-employed	Rural	243.53	27	never smoked	0
15266	Female	32	0	0	0 Yes	Private	Rural	77.67	32.3	smokes	0
28674	Female	74	1	0	0 Yes	Self-employed	Urban	205.84	54.6	never smoked	0
10460	Female	79	0	0	0 Yes	Govt_job	Urban	77.08	35	Unknown	0
64908	Male	79	0	1	1 Yes	Private	Urban	57.08	22	formerly smoked	0
63884	Female	37	0	0	0 Yes	Private	Rural	162.96	39.4	never smoked	0
37893	Female	37	0	0	0 Yes	Private	Rural	73.5	26.1	formerly smoked	0
67855	Female	40	0	0	0 Yes	Private	Rural	95.04	42.4	never smoked	0
25774	Male	35	0	0	0 No	Private	Rural	85.37	33	never smoked	0
19584	Female	20	0	0	0 No	Private	Urban	84.62	19.7	smokes	0
24447	Female	42	0	0	0 Yes	Private	Rural	82.67	22.5	never smoked	0
49589	Female	44	0	0	0 Yes	Govt_job	Urban	57.33	24.6	smokes	0
17986	Female	79	0	1	1 Yes	Self-employed	Urban	67.84	25.2	smokes	0
29217	Female	65	1	0	0 Yes	Private	Rural	75.7	41.8	Unknown	0
72911	Female	57	1	0	0 Yes	Private	Rural	129.54	60.9	smokes	0

Machine Learning PIPELINE

The chase name of lewing to usser Learning your laon and theprance
hean dotimation prupose's to some , and crormtaneants deering.
teniculig and for mot your cheare speccel fireing.



Pipeline del Progetto

1

Exploratory Data Analysis (EDA)

Analisi dei dati grezzi attraverso plot e report.

2

Pre-processing del Dataset

Pulizia, bilanciamento e selezione delle feature.

3

Addestramento & Validazione

Training, ottimizzazione iperparametri e threshold.

4

Valutazione

Metriche di performance e interpretabilità.

Analisi Esplorativa dei Dati (EDA)

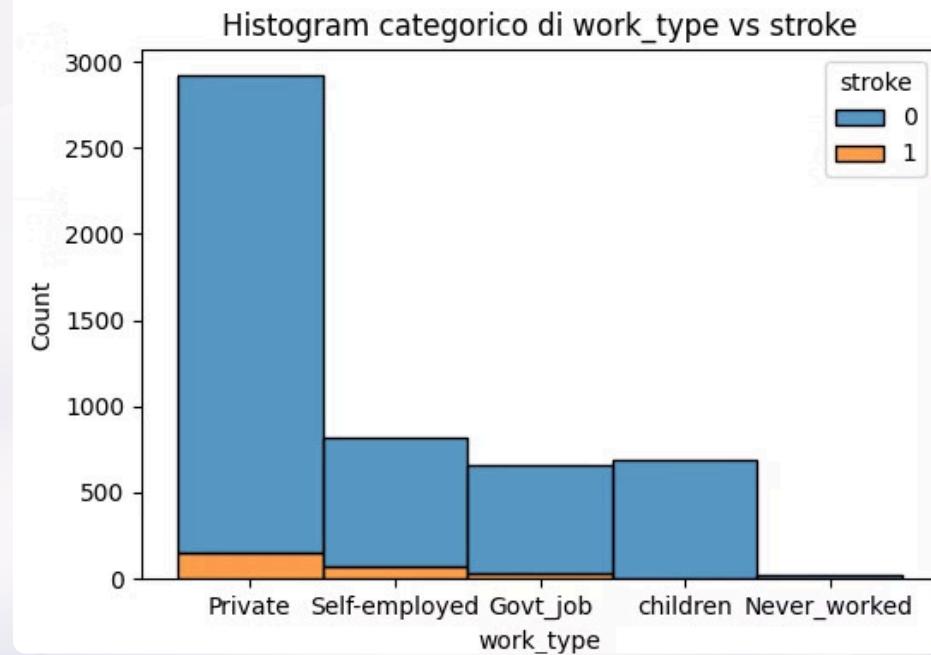
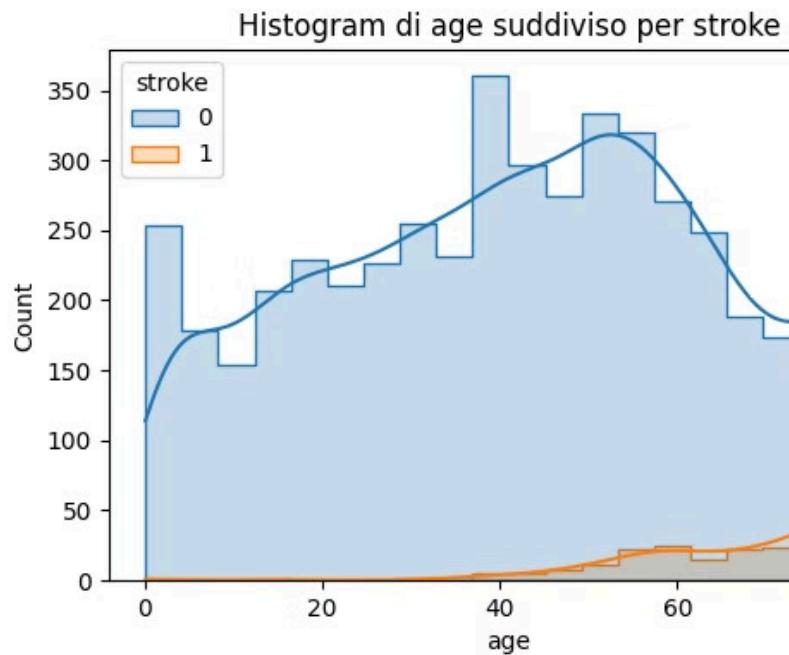
1 Obiettivo

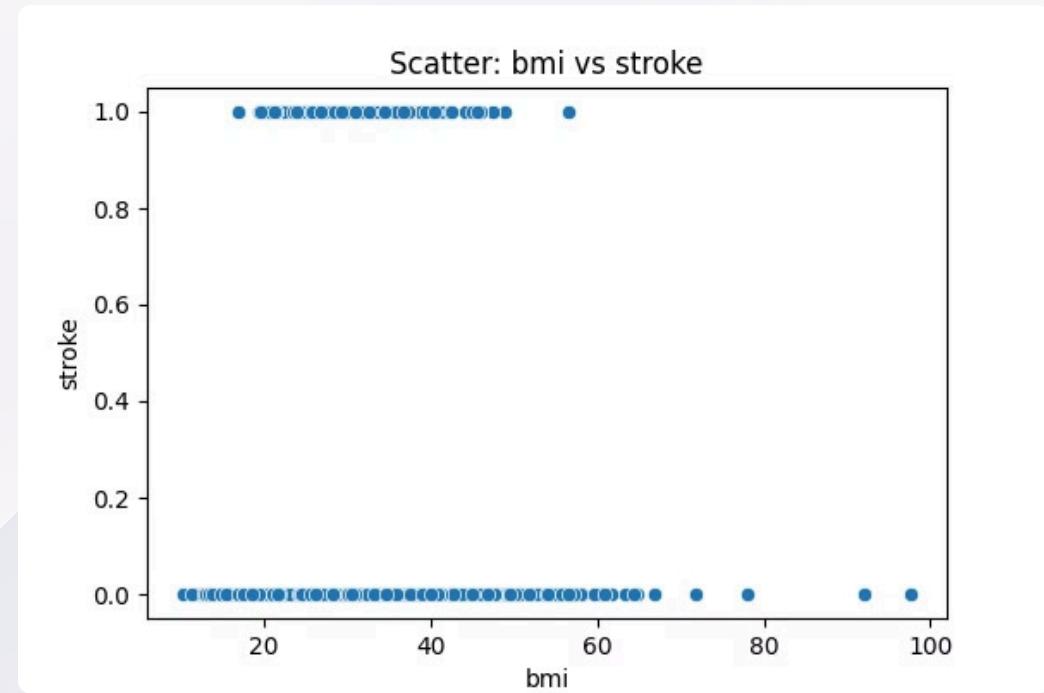
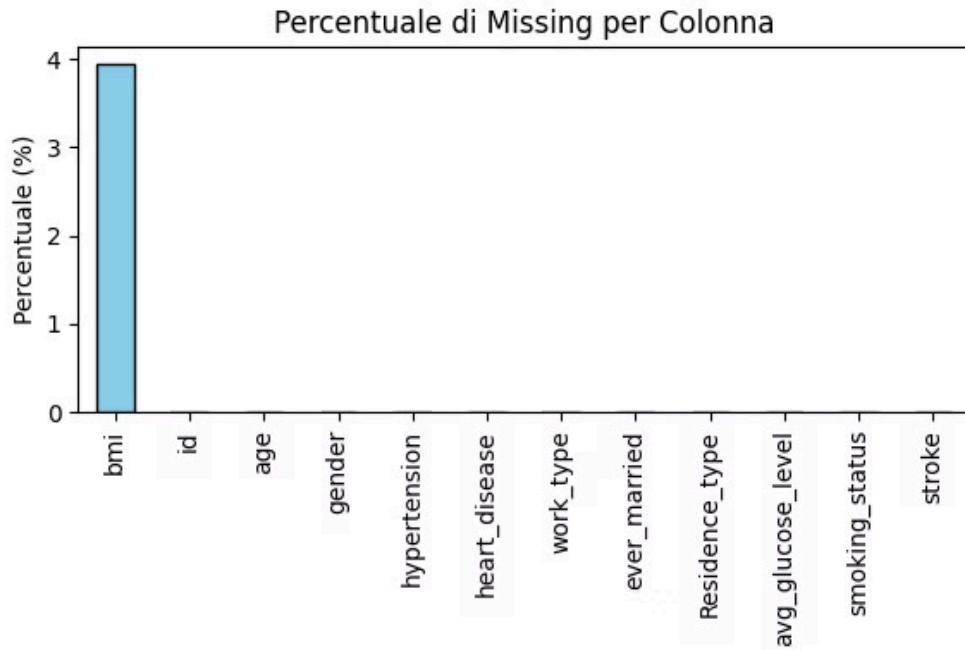
Comprendere la distribuzione dei dati, individuare correlazioni e identificare potenziali problemi.

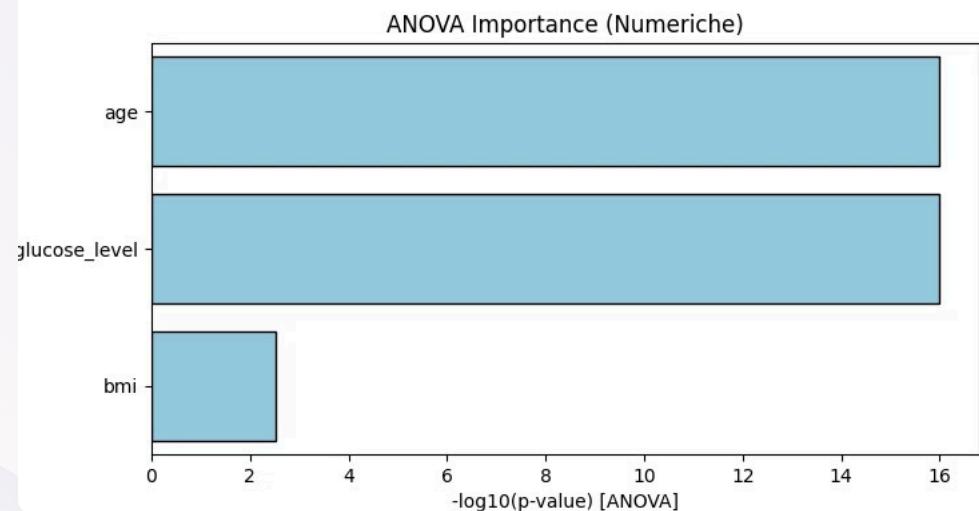
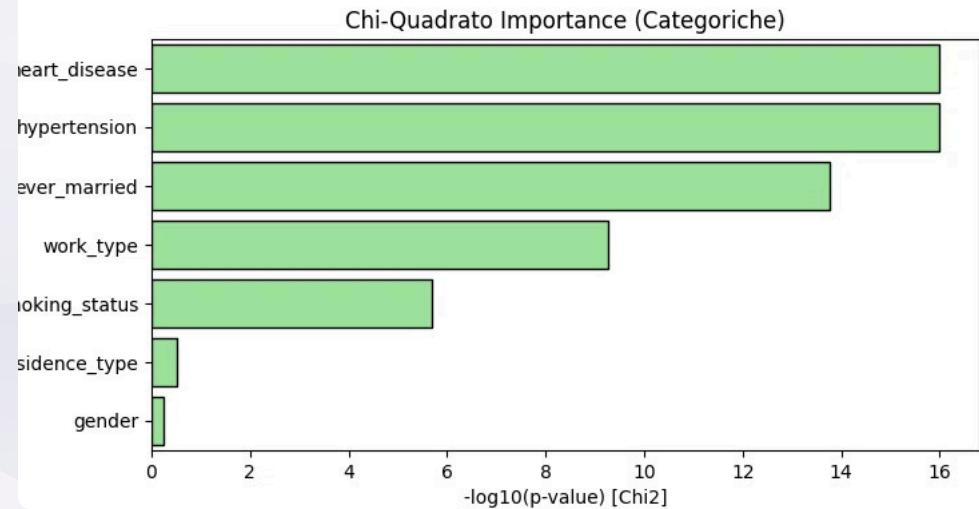
2 Principali analisi effettuate

- Analisi dei Valori Mancanti (Bar plot) + Little's MCAR Test
- Distribuzione delle Variabili Numeriche e Categoriche (Istogrammi)
- Analisi degli Outlier (IQR e Scatter plot)
- Correlazioni tra Feature (Pearson, Cramér's V, VIF)
- Rilevanza delle Feature sul Target (ANOVA / Chi-Quadro)
- Analisi dello Squilibrio del Target (Pie chart)



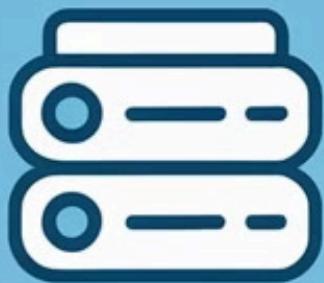








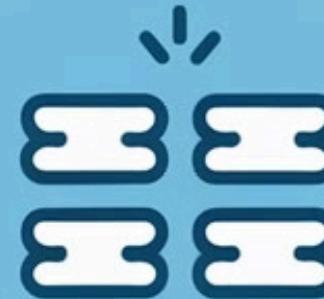
Data cleaning



Data imputation



Data balancing



Train/Validation/Test

Pre-processing del Dataset

Data Cleaning

Rimozione di **dati inconsistenti**.

Data Imputation

Sostituzione valori nulli nel **BMI** con la mediana.

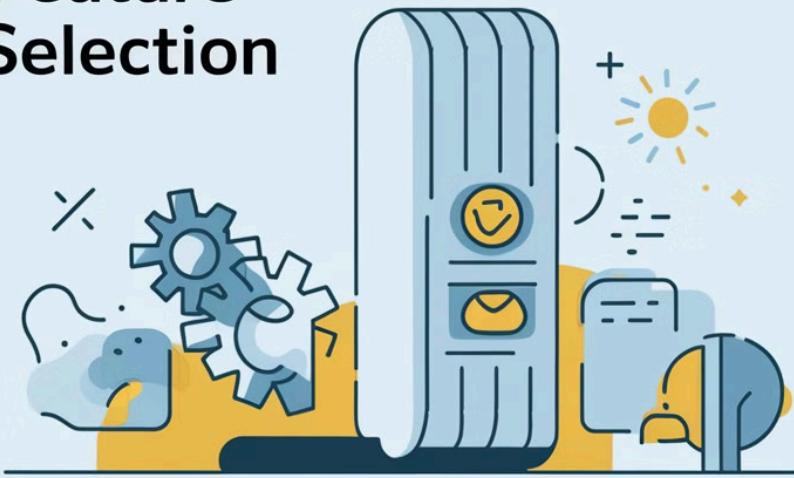
Data Balancing

SMOTE-Borderline per aumentare i casi di ictus.

Train/Validation/Test split

70%-15%-15% con **stratificazione** della classe target.

Feature Selection



Feature Encoding



Operazioni di Feature Engineering

Feature Selection

Rimozione delle variabili **id**, **gender** e **residence_type** in seguito a un'analisi della loro bassa correlazione con la variabile target.

Feature Encoding

- **Binarizzazione** della variabile **ever_married** (conversione in 0/1).
- **Label Encoding** delle **variabili categoriche** per adattarle al modello che richiede input numerici.

Addestramento del Modello

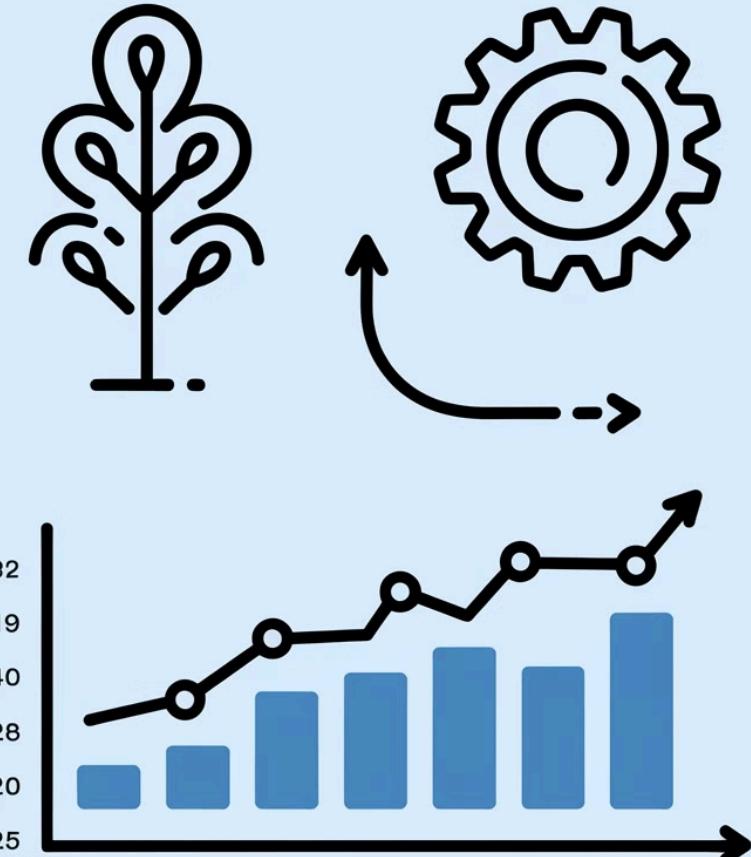
Modello scelto: Random Forest Classifier

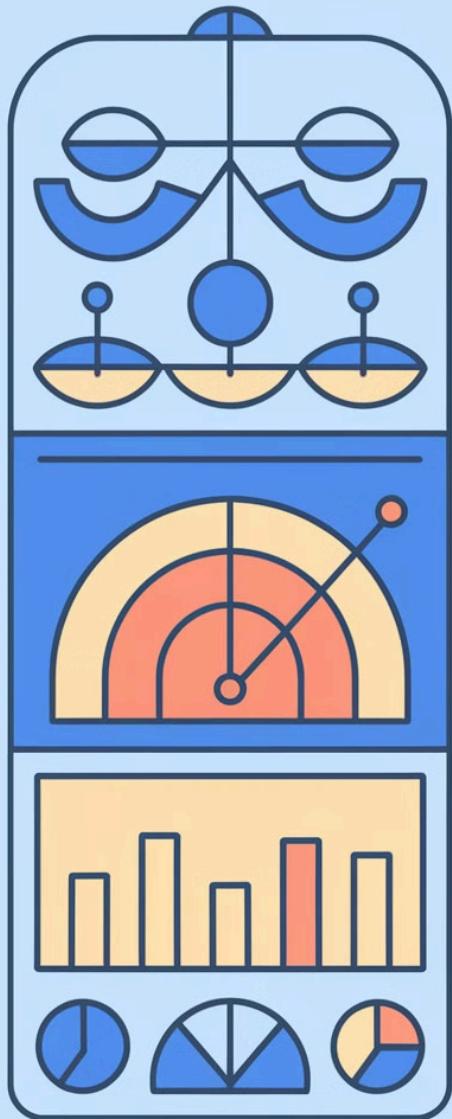
Perché?

- Supporta lo **sbilanciamento** (class_weight=balanced).
- Interpretabile grazie alla **feature importance**.

Ottimizzazione iperparametri

- **RandomizedSearchCV** con **Stratified K-Fold (k=5)**
- **Metriche prioritarie:** F2-score e ROC-AUC
- **Riaddestramento finale** su tutto il training set una volta ottenuti i miglior iperparametri.





Validazione del Modello



Obiettivo

In questa fase, oltre a validare il modello, la **soglia decisionale** viene ottimizzata, evitando il rischio di overfitting e migliorando la capacità predittiva del modello dato lo sbilanciamento del target.



Precision-Recall Curve

Utilizzo della curva per trovare la soglia che massimizza l'**F2-score** per bilanciare precisione e richiamo, con un'enfasi maggiore sulla recall.



Metriche di Validazione

Calcolo di **Precision, Recall, F2-score e ROC-AUC** sul validation set per una valutazione robusta del modello prima della valutazione finale sul test set.

Risultati ottenuti sul Test set

86.18%

Accuracy

21.19%

Precision

65.79%

Recall

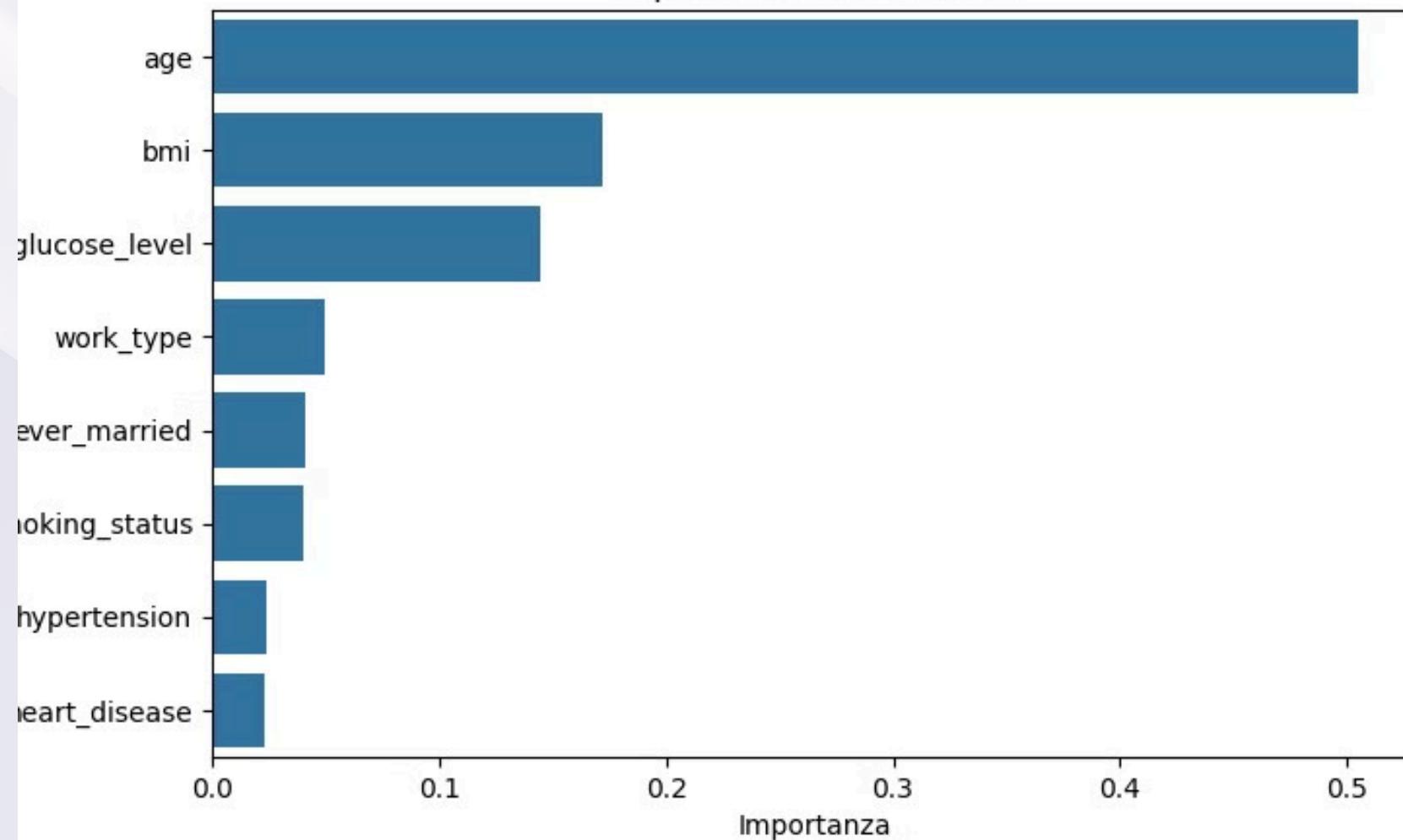
32.05%

F1-score

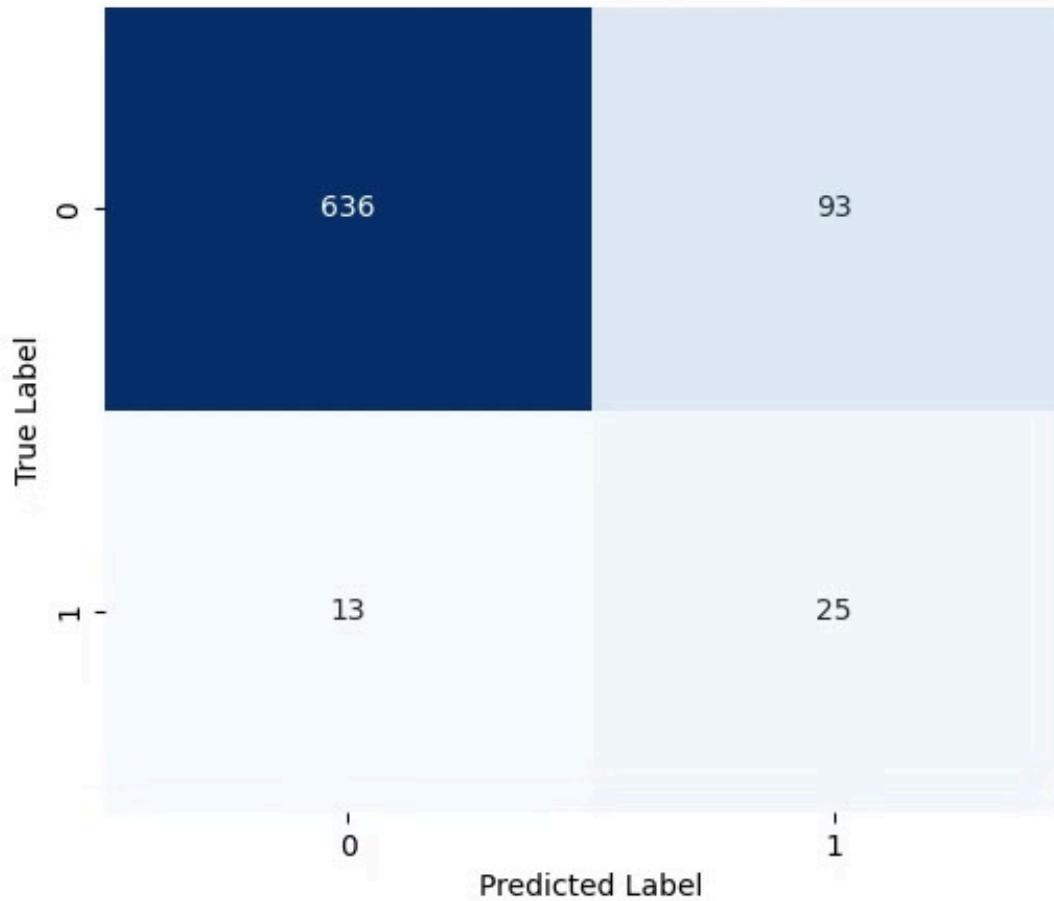
0.82

ROC-AUC

Importanza delle Feature



Matrice di Confusione



Considerazioni e Ottimizzazioni Future

Osservazioni chiave

- **Recall moderata** → efficace per identificare casi di ictus.
- **Precisione bassa** → necessità di più dati reali.

Limiti identificati

- **SMOTE** può causare molto facilmente **overfitting**, per questo abbiamo preferito regolare il peso delle classi tramite l'iperparametro.
- La **precision** lascia ancora desiderare, la **recall** anche se >50% potrebbe essere ulteriormente migliorata.

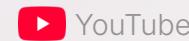
Eventuali miglioramenti

- Acquisire **più dati reali** per migliorare la generalizzazione.
- Esplorare **modelli più avanzati** (es. ensemble con reti neurali).
- Valutare **alternative al Label Encoding** per variabili categoriche (es. Target Encoding o Feature Hashing).

Demo Interfaccia Utente

0
Heart Disease:
0
Ever Married:
0
Work Type:
Never_worked
Average Glucose Level:
110.00
BMI:
23.00
Smoking Status:
never smoked
Predict Stroke Risk

Deploy ⋮



[ML] Stroke Prediction - Video Demo

▶ 01:21

Conclusioni

Il modello ha evidenziato **l'applicabilità del Machine Learning** nella **diagnosi** precoce di ictus.

Nonostante una precision bassa, possiamo **ritenerci soddisfatti** della recall, un aspetto cruciale nei contesti clinici.

L'intelligenza artificiale sta trasformando la medicina, rendendo le diagnosi sempre più rapide e accurate.

Attraverso questo lavoro, abbiamo sperimentato il potenziale del Machine Learning in un contesto accademico, esplorando le sfide e le opportunità di applicarlo alla predizione dell'ictus.

Grazie per l'attenzione!

