AA 2021/2022

PERFORMANCE NEI MODELLI DI CLASSIFICAZIONE PER ICTUS E IPERTENSIONE

Progetto di MACHINE LEARNING

Potertì Daniele 844892, CdLM Data Science
Sanvito Alessio 844785, CdLM Data Science
Sanvito Simone 844794, CdLM Data Science

Sommario

Secondo l'Organizzazione Mondiale della Sanità (OMS) l'ictus è la seconda causa di morte a livello globale, responsabile di circa l'11% dei decessi totali. L'ictus cerebrale è causato dell'improvvisa chiusura o rottura di un vaso cerebrale e dal conseguente danno alle cellule cerebrali dovuto dalla mancanza dell'ossigeno e dei nutrienti portati dal sangue (ischemia) o alla compressione dovuta al sangue uscito dal vaso (emorragia cerebrale). La consapevolezza precoce dei diversi segni premonitori dell'ictus può ridurre al minimo l'ictus. La **prima domanda** a cui abbiamo cercato di dare una risposta è stata: è possibile classificare il verificarsi o meno di un ictus in un soggetto considerato in base ad alcune caratteristiche del soggetto stesso? In seguito abbiamo riproposto la stessa **domanda**, ma su uno dei fattori di rischio per quanto riguarda l'ictus, ovvero l'ipertensione: è possibile classificare il verificarsi o meno di ipertensione in un soggetto considerato in base ad alcune caratteristiche specifiche (bmi, age, avg_glucose_level e smoking_status) del soggetto?
Per rispondere a queste domande sono stati utilizzati diversi modelli classificatori quali Logistic Regression, SVM, Random Forest, Weka J48, Naive Bayes, Decision Tree, XGBoost e Gradient Boosted; in seguito ne sono stati visualizzati i risultati comparandoli tra di loro in modo da vedere quale modello fornisca le performance migliori.

Indice

1. Descrizione delle variabili

https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

- **id**: identificatore univoco;
- **gender**: genere, "Male", "Female" o "Other";
- **age**: età del paziente;
- **hypertension**: 0 se il paziente non soffre di ipertensione, 1 se il paziente soffre di ipertensione;
- **heart_disease**: 0 se il paziente non ha alcuna malattia cardiaca, 1 se il paziente ha una malattia cardiaca;
- **ever_married**: valori "No" o "Yes" che stanno ad indicare se il paziente si è mai sposato;
- **work_type**: tipo di lavoro del paziente. I valori possono essere "children", "Govt_jov", "Never_worked", "Private" or "Self-employed";
- **Residence_type**: tipo di residenza del paziente. Può essere di tipo "Rural" o "Urban";
- **avg_glucose_level**: livello di glucosio medio nel sangue;
- **bmi**: body mass index (indice di massa corporea);
- **smoking_status**: descrive se il paziente fuma o meno. Può assumere 4 valori: "formerly smoked", "never smoked", "smokes" o "Unknown" (quando l'informazione non è disponibile per il paziente);
- **stroke**: 1 se il paziente ha avuto un ictus, 0 se non ce l'ha avuto.

2. Data exploration/preprocessing

2.1. Variabili con missing values

bmi: effettuando un'analisi statistica è emerso che la variabile bmi presenta 201 valori mancanti (4%). Sono dunque stati trasformati i valori N/A in Null, successivamente sono stati eliminati i missing values sostituendoli con valori ottenuti da un simple regression tree basato sulle variabili age e gender (poiché il valore di bmi dipende da sesso, età, altezza e peso, ma non abbiamo altezza e peso nel dataset).

2.2. Visualizzazione delle variabili

Sono stati realizzati diversi grafici per analizzare le variabili del dataset relative a age, hypertension, heart_disease, residence type, work type e smoking status, relativamente alla variabile target stroke.

Ciò che emerge dalle visualizzazioni è che, come ci si può aspettare, le feature di heart disease e hypertension sono correlate al fatto che un paziente abbia avuto un ictus.

Lo smoking status di un paziente non sembra essere particolarmente rilevante. Sorprendentemente, infatti, notiamo che la probabilità di ictus in chi non ha mai fumato è statisticamente non differente tra chi fuma e chi ha fumato in passato.

Per quanto riguarda la feature work_type, chi è nella categoria dei "private" è più soggetto a ictus. La visualizzazione suggerisce anche che tra i lavoratori, chi è nella categoria "govt_job" è meno soggetto a ictus.
I bambini e chi non ha mai lavorato non sono soggetti a ictus.

Analizzando la feature residence_type, si può notare che le persone che hanno un'abitazione in un contesto rurale siano meno soggette a ictus rispetto alle persone che abitano in città.

Per quanto riguarda la variabile avg_glucose_level si ha una distribuzione bimodale della variabile stroke, che registra i valori più alti, i due picchi, per i valori di glucosio pari a 80 e 220 circa.

Si nota che la variabile stroke ha una distribuzione pressoché normale con un forte picco intorno al valore di 30 della variabile bmi. La coda corrispondente alla variabile stroke è più spessa e più lunga rispetto alla variabile no stroke, quindi indica che per pazienti con bmi alto la probabilità di avere un ictus è maggiore. Infine, notiamo una correlazione significativa tra ictus e anzianità; infatti il numero di occorrenze di pazienti che hanno avuto un ictus aumenta all'aumentare dell'età.

Per quanto riguarda la classificazione sulla variabile stroke sono state considerate tutte le variabili, anche se alcune si sono rilevate statisticamente poco significative (si veda il corrplot). Considerando, invece, la classificazione sulla variabile hypertension, sono state tenute in considerazioni 4 colonne: bmi, age, avg_glucose_level e smoking_status. Le prime 3 risultano tra le variabili più correlate con la variabile su cui si è deciso di fare la classificazione; smoking_status, anche se statisticamente non correlata, è stata comunque utilizzata nella classificazione in quanto, secondo ricerche scientifiche, risulta una delle cause che portano un soggetto a soffrire di ipertensione e ictus.

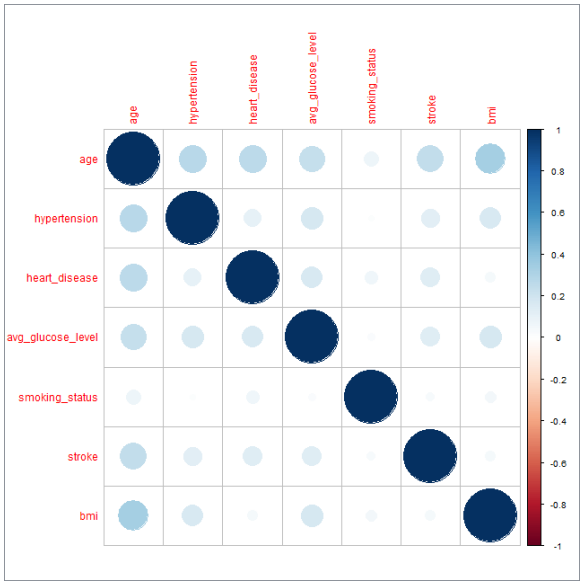

Figura 1: Corrplot tra alcune variabili del dataset

3. Modelli utilizzati

I modelli utilizzati per rispondere alle nostre domande di ricerca sono stati:

3.1. Logistic Regression

La regressione logistica, nonostante il suo nome, è un modello lineare per la classificazione, piuttosto che la regressione. In questo modello, le probabilità che descrivono i possibili esiti di una singola prova sono modellate utilizzando una funzione logistica.

3.2. SVM

Le macchine vettoriali di supporto (SVM) sono un insieme di metodi di apprendimento supervisionato utilizzati per la classificazione, la regressione e il rilevamento di valori anomali.

3.3. Random Forest

Nelle foreste casuali ogni albero decisionale nell'insieme è costruito da un campione estratto con sostituzione (cioè un campione bootstrap) dal set di addestramento.

3.4. Weka J48

È un algoritmo per generare un albero decisionale generato da C4.5 (un'estensione di ID3). È anche noto come classificatore statistico.

3.5. Naive Bayes

I metodi Naive Bayes sono un insieme di algoritmi di apprendimento supervisionato basati sull'applicazione del teorema di Bayes.

3.6. Decision Tree

Gli alberi decisionali (DT) sono un metodo di apprendimento supervisionato non parametrico utilizzato per la classificazione e la regressione.

3.7. Gradient Boosted

Gradient Tree Boosting o Gradient Boosted Decision Trees (GBDT) è una procedura standard accurata ed efficace che può essere utilizzata sia per problemi di regressione che di classificazione in una varietà di aree.

3.8. XGBoost

XGBoost è un sistema scalabile di machine learning per il potenziamento degli alberi. È un'implementazione efficiente ed effettiva del Gradient Tree Boosting.

4. Prima domanda di ricerca

È possibile classificare il verificarsi o meno di un ictus in un soggetto considerato in base ad alcune caratteristiche del soggetto stesso?

4.1. Holdout e SMOTE

La prima operazione è stata la rimozione della colonna id in quanto poco significativa. Non è

stata effettuata alcuna feature selection in quanto il dominio del dataset è centrato sulla nostra domanda di ricerca: le feature disponibili sono in un numero equilibrato (né troppo piccolo né troppo grande) e rappresentano l'insieme delle cause di ictus che vengono solitamente identificate e indicate dagli esperti di dominio e dalle ricerche scientifiche. Non ci è sembrato necessario, dunque, filtrare e utilizzare solo alcune feature.

È stato creato un partizionamento del dataset in training set (67%, train_set_A) e test set (33%, test_set_A) seguendo la metodologia **holdout**, eseguendo uno stratified sampling sulla variabile stroke.

In seguito, dato che i valori della variabile stroke nel dataset iniziale sono sbilanciati (i valori pari a 1 rappresentano il **4%** dei dati totali per la specifica colonna), è stata utilizzata una tecnica di oversampling nel training set per bilanciare i valori pari a 0 e i valori pari a 1. Per fare ciò si sono dovute trasformare in String alcune variabili (inizialmente numeriche) in modo che venissero trasformate solo in 0 o 1 (lasciandole numeriche avrebbero assunto valori non adatti come 0.1, 0.2, ecc.). Dunque, invece di duplicare i dati, è stata utilizzata la tecnica **SMOTE** (Synthetic Minority Oversampling Technique) per creare dati sintetici per il sovracampionamento.

SMOTE è uno dei metodi di oversampling più popolari; il suo approccio genera osservazioni "sintetiche" a partire dalla classe in minoranza e va ad aggiungere queste osservazioni al set di dati su cui si sta applicando la tecnica. I record "sintetici" sono generati basandosi sulla similarità nello spazio dei predittori: per ogni record della classe di minoranza x_i sono create k osservazioni (basandosi sui K-nearest neighbors) e vengono prese solo quelle più vicine ad x_i.

I modelli usati per rispondere alla prima research question sono stati: Random Forest, J48, Decision Tree, SVM, Naive Bayes, Logistic Regression e XGBoost. Per quanto riguarda i modelli Logistic Regression e SVM sono state trasformate le variabili categoriche che interessavano per creare il modello predittivo (gender, hypertension, heart_disease, ever_married, work_type, Residence_type, smoking_status) in variabili numeriche. Questo poiché questi due modelli non permettono l'utilizzo di variabili categoriche.

4.2. Cross validation

Per ogni modello è stato poi utilizzato il train set A (su cui è stato applicato SMOTE) per stimare l'accuratezza del modello in questione: in particolare, è stata impiegata una tecnica di **cross validation** cambiando, in base al modello di classificazione, il tipo di Learner e di Predictor. In più è stata fatta una cross validation sui dati di train con 5 fold, dividendo dunque in 80% per il train set (train_set_B) e 20% per il test set (test_set_B). Da qui sono poi state calcolate alcune misure di performance per ogni modello: precision, recall e f-measure (per i valori pari a 1 della variabile stroke) e accuracy totale del modello.

4.3. Analisi risultati modelli

Modello	*Recall*	*Precision*	*F-measure*	*Accuracy overall*
Logistic Regression	0.842	0.766	0.803	0.793
SVM	0.947	0.84	0.89	0.883
Decision Tree	0.924	0.878	0.9	0.898
Random Forest	0.966	0.88	0.921	0.918
J48	0.94	0.874	0.906	0.902
Naive Bayes	0.895	0.744	0.813	0.794
XGBoost	0.965	0.945	0.955	0.954

Tabella 1: misure di performance per la classificazione di stroke con cross validation

Per tutti i modelli sono stati presi i valori di performance del modello per il valore 1 della variabile, ovvero quando il soggetto ha un ictus.

Successivamente è stato utilizzato il test set A dell'holdout per ottenere delle predizioni da parte del modello su dati nuovi, mai visti. I

risultati ottenuti sono coerenti con quelli derivati dalla cross-validation.

Dai risultati si evince che i modelli con performance migliori sono XGBoost e Random Forest. Per confermare quanto affermato qui sopra sono state analizzate le ROC Curve dei vari modelli utilizzati: queste vanno ad evidenziare che tutti i modelli sono buoni dato che le aree sottese dalle curve dei modelli sono maggiori di quella sottesa dalla no-discrimination line (ovvero quella di un classificatore casuale, che non porta nessun beneficio); in aggiunta XGBoost si conferma come il modello migliore, essendo superiore anche a Random Forest.

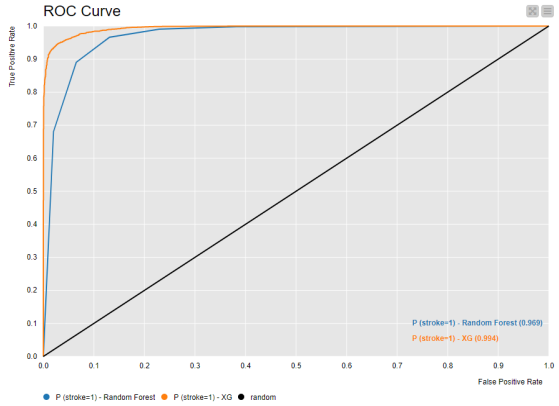

Figura 2: ROC curve per RF - XG per stroke

Successivamente è stata effettuata una comparazione più approfondita di questi due modelli andando a calcolare, usando una distribuzione T di Student, un intervallo di confidenza al 95% per il valore della media della differenza degli errori.

$$\left(\bar{d} - t^{K-1}_{1-\alpha/2} \cdot \hat{\sigma}_{d^{cv}}, \bar{d} + t^{K-1}_{1-\alpha/2} \cdot \hat{\sigma}_{d^{cv}}\right)$$

Figura 3: intervallo di confidenza

I risultati sono stati visualizzati in due boxplot, uno riguardante l'intervallo di confidenza, l'altro riguardante l'errore del modello sul test set; dal primo si evince che l'estremo sinistro dell'intervallo è strettamente maggiore di 0, quindi il modello XGBoost è statisticamente preferibile al modello Random Forest con il 95% di confidenza. Inoltre, l'errore del modello XGBoost è minore di quello del modello Random Forest.

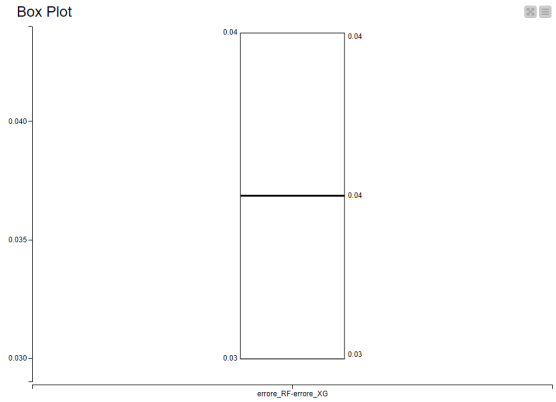

Figura 4: Boxplot intervallo di confidenza XG - RF

Per quanto riguarda i modelli che classificano solo attributi numerici, ovvero Logistic Regression e SVM, il procedimento seguito è stato lo stesso.

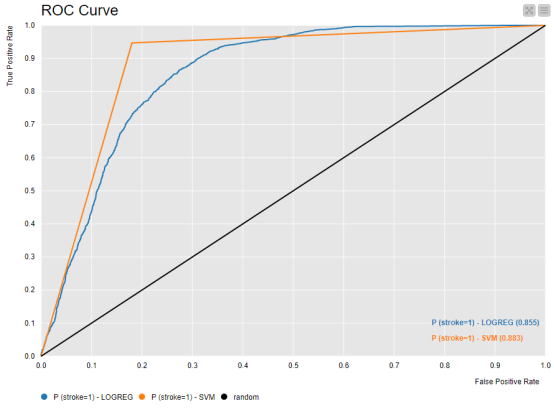

Figura 5: ROC Curve LR-SVM per stroke

Da queste due ROC Curve si può notare che il modello migliore sia SVM fino al valore 0.48 circa del False Positive Rate; dopo questa soglia il modello di regressione logistica performa meglio, ma in generale resta peggiore rispetto a SVM.

Per quanto concerne il boxplot che rappresenta l'intervallo di confidenza si evince che l'estremo sinistro dell'intervallo è strettamente maggiore di 0, quindi il modello SVM è statisticamente preferibile al modello Logistic Regression; l'altro boxplot evidenzia che l'errore di SVM sul test set è minore di quello del modello di regressione logistica,

sottolineando il fatto che SVM sia migliore anche in questo caso.

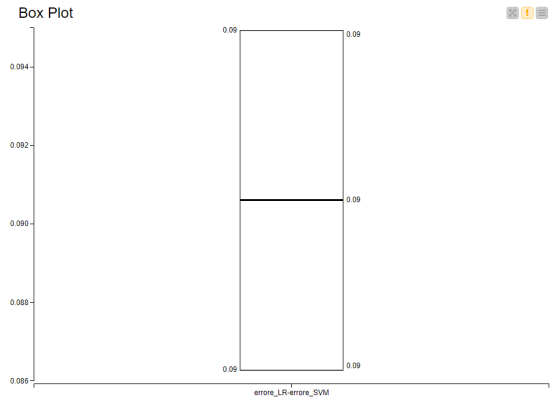

Figura 6: Boxplot intervallo di confidenza LR- SVM

5. Seconda domanda di ricerca

È possibile classificare il verificarsi o meno di ipertensione in un soggetto considerato in base ad alcune caratteristiche specifiche (bmi, age, avg_glucose_level e smoking_status) del soggetto?

5.1. Holdout, ROSE e SMOTE

La prima operazione consiste in una fase di **feature selection:** sono stati eliminati gli attributi non di nostro interesse, in questo caso id, stroke, heart_disease, gender, ever married, work type e residence type. Sono state tenute così solo le colonne che soddisfacessero i parametri di classificazione prestabiliti, ovvero quelli maggiormente influenti a livello scientifico: bmi, age, avg_glucose_level e smoking_status.

A seguire vi è stata la creazione di un partizionamento del dataset in training set (67%, train_set_A) e test set (33%, test_set_A) seguendo la metodologia **Holdout**, eseguendo uno stratified sampling sulla variabile hypertension.

In seguito, dato che i valori della variabile hypertension nel dataset iniziale sono sbilanciati (i valori pari a 1 rappresentano il **10%** dei dati totali per la specifica colonna), è stata utilizzata una tecnica di oversampling nel training set per bilanciare i valori pari a 0 e i valori pari a 1. Per fare ciò si sono dovute trasformare in String alcune variabili (inizialmente numeriche), come fatto anche per la classificazione della prima domanda di ricerca. Dunque, invece di duplicare i dati, è stata utilizzata la tecnica **ROSE** (Random Over Sampling Examples). ROSE è un algoritmo di ricampionamento il cui scopo è quello di generare nuove osservazioni sintetiche attraverso la stima di una funzione di densità kernel.

Inoltre, parallelamente a ROSE, è stato applicato anche **SMOTE** al training set sbilanciato (train_set_A); in questo modo si è potuta confrontare la performance dei modelli bilanciati con ROSE con quella dei modelli bilanciati con SMOTE.
I modelli utilizzati per rispondere a questa domanda di ricerca sono stati XGBoost, Random Forest, J48 (i 3 migliori modelli per la prima domanda di ricerca) e Gradient Boosted.

5.2. Cross validation

Come avvenuto già per la prima research question, anche questa volta è stata applicata una tecnica di cross validation con 5 folds per misurare la performance dei modelli, sia per quelli bilanciati con ROSE che per quelli bilanciati con SMOTE. La cross validation prende in input proprio questi dati di train bilanciati e li divide in 80% train set (train_set_B) e 20% test set (test_set_B).

5.3. Analisi risultati modelli

Per tutti i modelli sono stati presi i valori di performance del modello per il valore 1 della variabile, ovvero quando il soggetto soffre di ipertensione. Per primi vengono visualizzati i risultati delle performance dei modelli che utilizzano i dati di train (train_set_A) bilanciati con la tecnica ROSE.

Successivamente è stato utilizzato il test set dell'holdout (quindi senza ROSE, test_set_A) per ottenere delle predizioni da parte del modello su dati nuovi, mai visti. Dai risultati si evince che i modelli con performance migliori sono XGBoost e Gradient Boosted.

È stata effettuata una comparazione più approfondita di questi modelli andando a calcolare, usando una distribuzione T di Student, un intervallo di confidenza al 95% per il valore della media della differenza degli errori. Anche in questo caso sono state analizzate la ROC curve, il boxplot per l'intervallo di confidenza e il boxplot per l'errore sul test set. I risultati sono stati i seguenti:

- ROC curve: Gradient Boosted si rivela il miglior modello di classificazione (ma anche le altre curve sottendono un'area maggiore rispetto a quella sottesa dalla no discrimination line);
- boxplot per intervallo di confidenza: zero è compreso nell'intervallo, quindi i due modelli classificatori (XGBoost e Gradient Boosted) sono indistinguibili a livello di errore, la differenza tra i due non è statisticamente significativa con confidenza al 95%;
- boxplot per l'errore sul test set: è leggermente migliore Gradient Boosted, ma la differenza non è significativa, infatti gli errori dei due modelli differiscono di 0,01.

In seguito, per confrontare ROSE e SMOTE, sono state calcolate anche le misure di performance dei modelli che utilizzano SMOTE per bilanciare il train_set_A:

Modello	*Recall*	*Precision*	*F-measure*	*Accuracy overall*
Random Forest	0.885	0.822	0.852	0.847
J48	0.898	0.743	0.813	0.794
Gradient Boosted	0.913	0.798	0.851	0.841
XGBoost	0.917	0.895	0.906	0.905

Tabella 2: misure di performance per la classificazione di hypertension con cross validation e SMOTE

Allo stesso modo sono state calcolate queste misure anche sui dati nuovi usando come input il test set A dell'holdout.

Dai risultati si evince, in primis, che si ottengono migliori performance da parte dei modelli che utilizzano SMOTE come tecnica di oversampling piuttosto che ROSE; inoltre, si osserva che i modelli con performance migliori sono XGBoost e Random Forest.

È stata poi effettuata una comparazione più approfondita di questi modelli andando a calcolare, usando una distribuzione T di Student, un intervallo di confidenza al 95% per il valore della media della differenza degli errori.

Analizzando le diverse performance dei modelli tramite la ROC Curve ne risulta che tutti i modelli sono migliori del classificatore casuale; in particolare, il modello migliore è, ancora una volta, XGBoost che ha un valore di AUC (Area Under Curve) pari a 0,972.

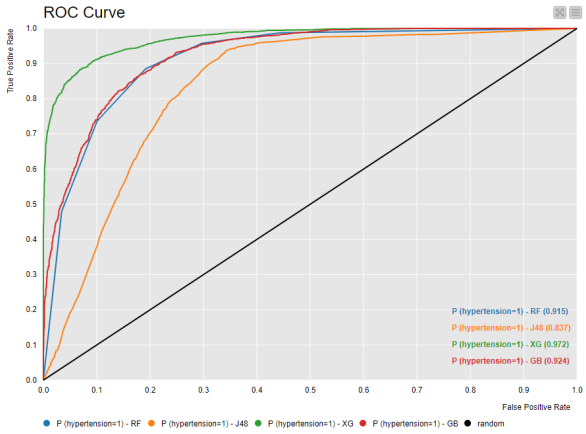

Figura 7: ROC Curve per i modelli per hypertension + SMOTE

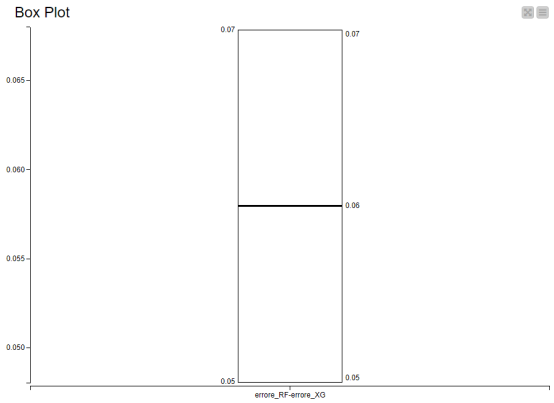

Figura 8: Boxplot intervallo di confidenza RF- XG per hypertension + SMOTE

Confrontando XGBoost con il secondo modello migliore, ovvero Random Forest, si ottiene che XGBoost è statisticamente preferibile in quanto l'estremo sinistro assume valore strettamente maggiore di zero e in quanto l'errore registrato da questo modello è inferiore rispetto a quello del suo diretto "concorrente".

Conclusioni e sviluppi futuri

Concludendo, sono state effettuate le classificazioni definite nelle due domande di ricerca: per rispondere alla prima domanda sono state utilizzate tutte le feature presenti nel dataset, ad eccezione dell'id.
La tecnica di oversampling che ha portato ad un'accuratezza massima è stata SMOTE.
Il modello che è risultato come modello migliore è stato XGBoost, come si è potuto ben notare dai valori delle performance ottenute e dalle ROC curve disegnate.

Per quanto concerne la seconda domanda di ricerca sono state utilizzate solo le feature più rilevanti per la nostra domanda: bmi, avg_glucose_level, age e smoking_status. I risultati ottenuti sono simili a quelli analizzati precedentemente per la prima domanda di ricerca: i risultati migliori, infatti, sono stati trovati usando SMOTE come tecnica di oversampling (a discapito di ROSE) e il modello migliore si è confermato, anche questa volta, XGBoost.

Possibili sviluppi futuri a questo progetto potrebbero essere portati in diversi modi: innanzitutto tramite l'introduzione di ulteriori caratteristiche per gli utenti contenuti nel dataset, come il livello di attività fisica praticata o specifiche sulla dieta seguita dagli utenti, in modo da poter portare una classificazione più mirata su alcuni attributi rilevanti piuttosto che su altri. Inoltre, si potrebbero richiedere informazioni aggiuntive ai pazienti colpiti da ictus e/o che soffrono di ipertensione (questo approccio, però, potrebbe trovare delle frizioni legali): in questo modo si potrebbe avere un numero maggiore di dati specifici su cui allenare i modelli e potenzialmente avere una predizione maggiormente accurata. Infine, si potrebbe migliorare il dataset inserendo nuovi dati per cercare di bilanciare le variabili di classe non eque oppure provare altre tecniche di oversampling da affiancare a SMOTE e ROSE.

Riferimenti

1. Dataset: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
2. Calcolatore dei valori di bmi : https://www.calculator.net/bmi-calculator.html
3. Cosa è lo SMOTE: https://datascience.eu/it/programmazione/smote/
4. SMOTE e ROSE: http://www.fedoa.unina.it/9890/2/Filomena%20Mauriello%20Tesi%20di%20dottorato.pdf#page26
5. Descrizioni dei modelli utilizzati: https://scikit-learn.org/stable/ https://www.softwaretestinghelp.com/weka-datasets/
6. Riferimenti su ipertensione e ictus: https://www.humanitas.it/malattie/ipertensione/ https://www.humanitas.it/malattie/ictus-cerebrale/
7. XGBoost: Tianqi Chen, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System"
8. Performance Analysis of Machine Learning Approaches in Stroke Prediction: https://ieeexplore.ieee.org/document/9297525
9. Ictus cerebrale, il fumo ne raddoppia il rischio: https://www.osservatoriomalattierare.it/malattie-croniche/14677-ictus-cerebrale-il-fumo-ne-raddoppia-il-rischio