Modelli di classificazione per la predizione dell'ictus

Silvia Grosso[1], Paola Impicciché[1] e Elisa Merelli[1]

[1] *Università degli Studi Milano Bicocca, CdLM Data Science*

Abstract—L'ictus, conosciuto anche come apoplessia o ischemia cerebrale, indica un danno del tessuto cerebrale o la morte di una sua porzione, causati da una scarsa perfusione sanguigna. Secondo l'Organizzazione Mondiale della Sanità (WHO) nel 2013 l'ictus è stata la seconda causa di morte più frequente, responsabile di 6,4 milioni di decessi (circa il 11% del totale) [1]. Nell'articolo "Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010" [2] viene studiato che dal 1990 al 2010 l'incidenza di ictus - standardizzata per età - è diminuita del 12% nei paesi ad alto reddito ed aumentata della stessa percentuale in quelli a basso e medio reddito. Si rendono dunque necessari ulteriori studi per approfondire la relazione che intercorre tra la manifestazione dell'ictus e i principali fattori favorenti. L'elaborato si propone di indagare attraverso modelli di classificazione le possibili caratteristiche legate a questa malattia, al fine di predirne la manifestazione.

Keywords—Ictus, Machine Learning, classificazione

INDICE

I Introduzione 1

II Struttura del dataset 1
- a Dati statisitci . . . 2

III Metodologia 2
- a Preprocessing . . . 2
- b Holdout e K-fold Cross Validation . . . 2
- c Modelli di classificazione . . . 2
- d Feature selection . . . 3
- e Dataset sbilanciato e Cost-Sensitive Learning 3
- f Valutazione delle performance . . . 3

IV Analisi dei risultati 3
- a Classificazione con metodo Holdout e Feature Selection . . . 3
- b Classificazione mediante Matrice dei costi . . 4
- c Classificazione mediante K-fold Cross Validation e Matrice dei costi . . . 4

V Conclusioni e sviluppi futuri 5

Riferimenti bibliografici 5

I. INTRODUZIONE

I sintomi dell'ictus esordiscono per lo più all'improvviso e si differenziano a seconda del tipo di ictus: ischemico o emorragico. Vi sono vari sistemi per il riconoscimento precoce dell'episodio di ictus che, pur non costituendo una diagnosi, possono rivelarsi preziosi in fase acuta per una rapida valutazione. Tuttavia, l'aspetto principale da approfondire sono i fattori di rischio [3]: in entrambe le tipologie di ictus si identificano come elementi favorenti l'età, l'ipertensione arteriosa, il diabete mellito, il tabagismo, l'abuso di alcol, l'alimentazione scorretta, la sedentarietà e l'obesità. Al fine di promuovere le misure preventive per l'ictus è importante identificare i soggetti a rischio partendo dall'analisi di dati oggettivi e facilmente acquisibili attraverso controlli periodici di routine. Questo lavoro nasce con lo scopo di analizzare i principali elementi caratterizzanti di un soggetto colpito da ictus e modellare una classificazione binaria volta a predirne l'episodio. Tale obiettivo verrà perseguito tramite l'uso di vari tipi di classificatori e l'ausilio di diverse tecniche di classificazione.

II. STRUTTURA DEL DATASET

Il set di dati utilizzato per lo sviluppo del progetto è lo "Stroke Prediction Dataset", presente su Kaggle [4], che raccoglie molteplici informazioni potenzialmente utili per la previsione dell'ictus. Il dataset contiene 5110 istanze, ciascuna riferita alle condizioni di un paziente, ed è composto da 11 attributi, di cui 5 categorici, 3 numerici e 3 binari.
Gli attributi presenti nel dataset sono i seguenti:

- *gender*: genere del paziente;
- *age*: età del paziente;
- *hypertension*: indica se un paziente è affetto da ipertensione;
- *heart_disease*: indica se un paziente soffre di problemi cardiaci;
- *ever_married*: stato civile del paziente;
- *work_type*: tipo di occupazione del paziente;
- *Residence_type*: tipologia di dimora in cui vive il paziente;
- *avg_glucose_level*: livello medio di glucosio presente nel sangue del paziente;

- *bmi*: indice di massa corporea del paziente;
- *smoking_status*: indica il rapporto del paziente con il fumo;
- *stroke*: indica se il paziente è stato colpito da ictus.

a. Dati statisitci

Nelle tabelle 1 e 2 mostriamo la distribuzione degli attributi discreti presenti nel dataset.

gender	hypertention	heart_disease
Female: 58.6%	0: 90.3%	0: 94.6%
Male: 41.4%	1: 9.7%	1: 5.4%
Other: 0.02%		

TABLE 1

Si osserva che alcune variabili, tra cui la variabile di classe *stroke*, non risultano bilanciate; questo importante aspetto sarà tenuto in considerazione per le modellazioni successive. La variabile *age* risulta ben distribuita in un range che varia tra i primi mesi di vita e 82 anni.
È stata inoltre eseguita un'analisi di correlazione tra le variabili che non ha evidenziato risultati notevoli (1).

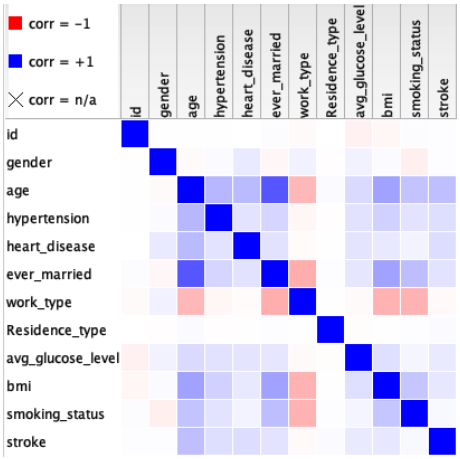

Fig. 1: Matrice di correlazione

III. METODOLOGIA

a. Preprocessing

Dalla fase di data exploration è emersa la presenza di 201 missing values per l'attributo *bmi*. I record che presentavano valori mancanti sono stati quindi trattati con il metodo conditional mean rispetto alla variabile *stroke*.
Inoltre, per soggetti di età inferiore ai 14 anni si è deciso di uniformare i campi di *smoking-type* e *work-type* rispettivamente ai valori "never-smoked" e "children", dato che alcuni record presentavano valori incompatibili con l'età registrata (si assume quindi che non siano stati fatti errori nell'acquisizione dell'età dei soggetti). Infine è stato eliminato l'unico record con il valore "other" in *gender*, rendendo l'attributo dicotomico.

b. Holdout e K-fold Cross Validation

L'insieme dei dati è stato partizionato secondo il metodo Holdout in due sottoinsiemi disgiunti, effettuando lo *stratified sampling* rispetto alla variabile *stroke*:

- training set: corrispondente al 70% del dataset, usato per l'apprendimento;
- test set: corrispondente al 30% del dataset, usato per la validazione.

Tale approccio permette di valutare le performance dei classificatori sugli "unseen record", ovvero i dati non utilizzati nella fase di apprendimento, controllando quindi il fenomeno di overfitting. L'Holdout costituisce il metodo più semplice per testare i modelli, tuttavia lega le misure di performance alle due specifiche partizioni generate. Infatti, dato che il test set è ricavato con una procedura di campionamento, campionamenti differenti possono portare a differenti stime nell'accuratezza dei modelli. Per ovviare a tale problema si effettua una seconda classificazione usando la tecnica più robusta K-fold Cross Validation. In questo secondo approccio il dataset è stato partizionato in 10 sottoinsiemi che costituiscono i K-folds, approssimativamente di uguali dimensioni. A causa del forte sbilanciamento della variabile classe, è stato nuovamente utilizzato il metodo *stratified sampling* per garantire che in ogni fase di apprendimento il partizionamento rispettasse la distribuzione della variabile *stroke* nel dataset iniziale. Ogni modello è stato quindi allenato, tramite più iterazioni, su una combinazione differente di K-1 folds e validato sulla restante parte. In questo modo è possibile effettuare migliori confronti tra le performance ottenute.

c. Modelli di classificazione

Al fine di individuare il modello più adatto per rispondere alla domanda di ricerca, sono state applicate diverse tecniche di classificazione:

- **modelli euristici**: basati su algoritmi semplici e intuitivi, comprendono la classe degli alberi di classificazione, che sfruttano il principio divide-et-impera per derivare i gruppi di osservazioni che risultano il più possibile omogenee rispettando la variabile classe. In particolare è stato applicato il *RandomForest*;
- **modelli di regressione**: basati sull'estensione della regressione lineare. La regressione logistica è una regressione adattata alla gestione di problemi di classificazione binaria;
- **modelli di separazione**: dividono lo spazio degli attributi in regioni disgiunte, separando le osservazioni secondo la variabile classe. Per questa categoria è stata applicata la *Support Vector Machine* impostando il kernel puk;
- **modelli probabilistici**: basati sull'ipotesi probabilistica riguardante la forma delle probabilità condizionate della variabile classe. In particolare sono stati utilizzati il *NaiveBayes* e *NBTree*.

work_type	residence_type	smoking_status	stroke	ever_married
Private: 57.2%	Urban: 50.8%	never smoked: 94.6%	0: 95.1%	Yes: 65.6%
Self-employed: 16.0%	Rural: 49.2%	Unknown: 5.4%	1: 4.9%	No: 34.4%
children: 13.4%		formerly smoked: 17.3%		
Govt_job: 12.8%		smokes: 15.4%		
Never_worked: 0.4%				

TABLE 2

d. Feature selection

La fase di feature selection è fondamentale per individuare gli attributi rilevanti per la classificazione. La riduzione della dimensionalità del dataset permette inoltre di svolgere analisi più performanti.
L'approccio utilizzato nel caso in esame è quello del multivariate filter (nodo *Weka AttributeSelectedClassifier*, evaluator: *CfsSubseEval*, search: *BestFirst*) . Tale metodo rimuove sia gli attributi irrilevanti che quelli ridondanti, mantenendo solo le variabili significative per la variabile classe e incorrelate tra di loro. Il subset ottimale di attributi si ottiene in modo del tutto indipendente dall'algoritmo di classificazione, attraverso l'ottimizzazione della funzione obiettivo scelta.

e. Dataset sbilanciato e Cost-Sensitive Learning

La variabile classe della domanda di ricerca è caratterizzata da un forte sbilanciamento. I valori 1 dell'attributo *stroke* costituiscono infatti meno del 5% dei valori totali. A causa della distribuzione sbilanciata della variabile risposta, i modelli ottenuti possono risultare distorti, perché il processo di apprendimento tende a focalizzarsi principalmente sulla classe con frequenza più alta e a ignorare gli eventi rari, che nello studio in questione rappresentano pazienti che hanno avuto un ictus.
I due principali approcci per trattare dataset con classi non equilibrate sono le tecniche di campionamento e il Cost-Sensitive Learning (CSL). Si è deciso di utilizzare quest'ultimo, attraverso il nodo *Weka CostSensitiveClassifier*, dal momento che eventuali tecniche di campionamento del dataset avrebbero ridotto notevolmente il numero di istanze. E' stata formulata appositamente una matrice dei costi, sperimentando diverse combinazioni dei valori, nel tentativo di minimizzare i costi di errata classificazione e, in questi termini, ottimizzare le prestazioni dei modelli. Queste combinazioni sono state effettuate consultando la letteratura [5] e tenendo conto della natura del problema di classificazione in esame. Poiché un'errata diagnosi infausta su un paziente sano porterebbe a sottoporre quest'ultimo a ulteriori esami clinici, tale errore sarebbe di gran lunga meno grave della dimissione di un paziente malato che, al contrario, necessiterebbe di cure.

f. Valutazione delle performance

Nelle analisi dei risultati sono stati utilizzati i seguenti criteri per valutare le performance dei modelli:

- Accuratezza: rapporto tra il numero di stime corrette e il numero totale di record utilizzati per computare la matrice di confusione.

$$Accuracy^1 = \frac{TP+TN}{TP+TN+FP+FN}$$

- Recall: frazione dei record positivi correttamente classificati dal modello.

$$Recall = \frac{TP}{TP+FN}$$

- Precision: frazione di record effettivamente positivi tra quelli che il modello classifica come appartenenti alla classe positiva.

$$Precision = \frac{TP}{TP+FP}$$

- F_1 - measure: è la metrica più importante per trattare un problema di classe sbilanciata, basata sulla media armonica di precision e recall. Un modello ottiene valori alti di F_1-measure solo se precision e recall sono entrambe elevate.

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

- ROC Curve (*Receiver Operating Characteristic*): permettono di osservare il trade-off tra il tasso di veri positivi e il tasso di falsi positivi.

- AUC (*Area Under the Curve*): misura dell'area sottostante la curva ROC. I valori variano tra 0 e 1, tanto più la misura si avvicina a 1, tanto più il modello ha una buona performance. La baseline del parametro AUC per un modello binario è 0.5, corrispondente al valore di un ipotetico modello che prevede in modo casuale una risposta tra 0 e 1.

- Cumulative Gains Chart: permette di valutare per varie dimensioni del subset la percentuale di veri positivi.

IV. ANALISI DEI RISULTATI

In questa sezione riportiamo e analizziamo i risultati ottenuti nei diversi approcci impiegati nella classificazione.

a. Classificazione con metodo Holdout e Feature Selection

Nella fase di feature selection, mediante l'utilizzo del *CFS multivariate filter*, è stato individuato il seguente subset ottimale di attributi:

[1] TP = True Positive, TN = True Negative, FP= False Positive, FN= False Negative

- *age*;
- *hypertension*;
- *heart_disease*;
- *avg_glucose_level*.

Le misure di performance dei modelli ottenute tramite il metodo Stratified Sampling e Holdout sono mostrate in tabella 3[2].

Modello	P	R	F_1	A
Logistic	?	0	?	0.957
SMOPuk	?	0	?	0.957
NBTree	?	0	?	0.957
RandomForest	0.375	0.048	0.085	0.957
NaiveBayes	0.229	0.302	0.26	0.957

TABLE 3: COMPARAZIONE MODELLI - FEATURE SELECTION

Poiché la misura di Accuracy è determinata dalla *Loss function*, che attribuisce a tutti gli errori il medesimo valore in termini di costi, si osserva che tutti i classificatori forniscono un'elevata misura di Accuracy a causa della natura sbilanciata della variabile classe. É sufficiente osservare le matrici di confusione presenti nei nodi *Scorer* di *Logistic*, *SMOPuk* e *NBTree* nel WorkFlow Knime per verificare che i modelli classificano tutte le istanze negativamente come il modello *Zero Rule*. In generale tutti i modelli in esame non sono adatti a interpretare correttamente il fenomeno considerato, come si evince dall'analisi delle metriche Precision, Recall e dalla F_1-measure[5], in quanto tendono a classificare tutte le istanze come appartenenti alla classe più probabile.

b. Classificazione mediante Matrice dei costi

In questa seconda fase sono stati utilizzati esclusivamente gli attributi risultati significativi dalla Feature selection. I valori delle metriche ottenute tramite metodo Holdout e l'analisi dei costi sono mostrati in tabella 5.

Modello	P	R	F_1	A	AUC
Logistic	0.072	0.984	0.134	0.454	0.812
SMOPuk	0.957	1	0.978	0.957	0.5
NBTree	0.068	0.984	0.128	0.426	0.473
RandomForest	0.098	0.381	0.155	0.823	0.637
NaiveBayes	0.086	0.873	0.157	0.598	0.567

TABLE 4: COMPARAZIONE MODELLI - COUNTING THE COST

Rispetto alla fase precedente i valori di accuratezza dei modelli sono, come previsto, considerevolmente peggiorati ed è ora possibile osservare differenti valori di F_1-measure, Precision e Recall. Come si può notare dal confronto tra i valori di Accuracy riportati in tabella 5 e il costo dei modelli in figura 2, modelli con una misura di accuratezza migliore sono associati a costi crescenti, a sottolineare il fatto che tale metrica non può essere considerata rilevante quando ci si pone l'obiettivo di minimizzare i costi. Nel complesso si osserva che i modelli *Logistic* e *NBTree* sono preferibili in termini di costi. A causa della complessità di interpretazione

[2]P = Precision; R = Recall; F_1 = F_1-measure; A = Accuracy; AUC = Area Under Curve.

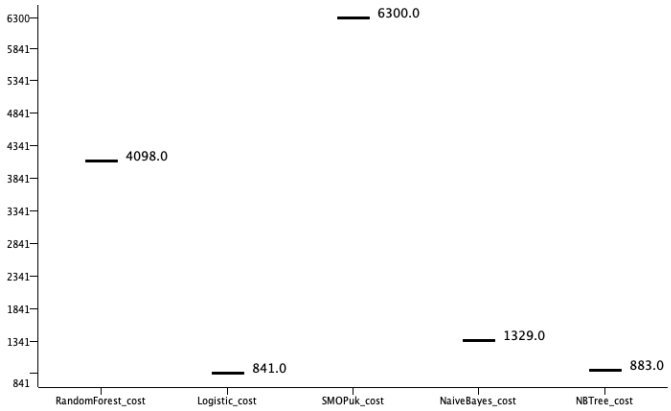

Fig. 2: Comparazione dei costi

della matrice dei costi, per un'ulteriore valutazione della performance dei classificatori sono state analizzate e confrontate la curva relativa all'andamento del Cumulative Gains (esempio del modello *Random Forest* in figura 3) e la ROC Curve di ogni modello.

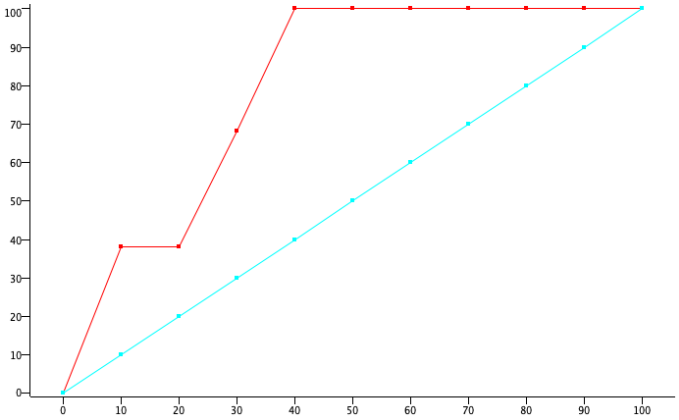

Fig. 3: Cumulative Gains - Modello Random Forest

Dal grafico Cumulative Gains del modello *Random Forest* si evince che selezionando il 30% del dataset il classificatore stima correttamente circa il 70% delle istanze positive. Dall'analisi dell'andamento delle *ROC Curve* in figura 4 si evince che il modello *SMOPuk* non è informativo in quanto la curva corrisponde alla retta rappresentante il *Random Model*. Per l'interpretazione dell'indice AUC proposta da Swets [6], i modelli *NaiveBayes* e *RandomForest* sono poco accurati mentre il modello Logistic risulta moderatamente accurato. Infine il valore di AUC associato al modello *NBTree*, inferiore al valore 0.5, indica che il classificatore performa peggio di un modello casuale, possibile segnale di inefficacia dell'algoritmo di training.

Nel complesso non si può individuare un classificatore migliore in assoluto in quanto non è presente una curva dominante rispetto alle altre. Per una percentuale di record falsi positivi tollerabili inferiore a circa il 15%, è preferibile il modello *RandomForest*, altrimenti è preferibile il modello *Logistic*.

c. Classificazione mediante K-fold Cross Validation e Matrice dei costi

Nel complesso attraverso il Cost Sensitive Learning non sono stati ottenuti risultati soddisfacenti. È possibile che non

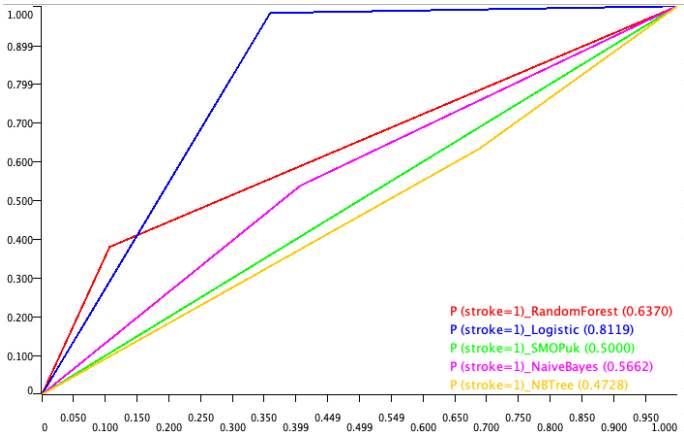

Fig. 4: ROC Curve dei differenti modelli

sia stata identificata la matrice di costo ottimale per il fenomeno in esame o che il bias di cui la tecnica di partizionamento Holdout è affetta abbia inficiato i risultati. In accordo con quanto descritto nella fase di metodologia, si applica il metodo del K-fold per rendere più affidabili i risultati ottenuti. I valori delle metriche sono riportati nella tabella . Nel complesso non si notano variazioni evidenti rispetto alla classificazione precedente.

Modello	P	R	F_1	A	AUC
Logistic	0.071	0.986	0.132	0.448	0.706
SMOPuk	0.957	1	0.978	0.957	0.5
NBTree	0.074	0.957	0.137	0.487	0.526
RandomForest	0.114	0.483	0.184	0.818	0.656
NaiveBayes	0.088	0.928	0.161	0.589	0.543

TABLE 5: COMPARAZIONE MODELLI - K-FOLD CROSS VALIDATION

Sono inoltre riportate in figura 5 le relative curve ROC. Si osserva che, fatta eccezione per il modello *SMOPuk* che è ancora paragonabile al modello casuale non informativo, tutti gli altri modelli in esame hanno raggiunto un valore di AUC superiore a 0.5. Per una percentuale di record falsi positivi tollerabili inferiore a circa il 35%, è nuovamente preferibile il modello *RandomForest*.

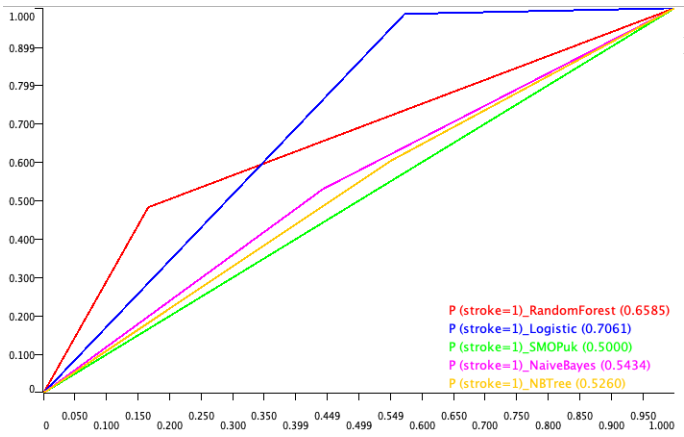

Fig. 5: ROC Curve dei differenti modelli - K-fold Cross Validation

V. CONCLUSIONI E SVILUPPI FUTURI

Il lavoro proposto in questo elaborato era volto all'identificazione del migliore modello di classificazione per la previsione dell'ictus. Per il raggiungimento di tale obiettivo sono state considerate le principali tecniche di classificazione.
Il dataset considerato per lo sviluppo di tale modello ha rivelato diversi attributi fortemente sbilanciati, tra cui la variabile classe *stroke*, oltre che un numero di istanze limitato per questo tipo di analisi.
Il principale problema affrontato in questa sede è stato dunque quello del Class Imbalance. Per prima cosa sono stati confrontate le performance di classificatori diversi tramite Holdout: i risultati hanno rivelato un alto livello di accuratezza che, contestualizzato alle altre metriche, mostrava un'interpretazione sbagliata del fenomeno.
Attraverso la tecnica di feature selection è stata ridotta la dimensionalità del dataset identificando gli attributi maggiormente legati alla manifestazione dell'ictus.
La seconda fase di analisi è stata svolta con l'introduzione della matrice dei costi volta ad ottimizzare le prestazioni dei modelli. I risultati ottenuti tramite Holdout, confrontati sulla base delle metriche, della curva relativa all'andamento del fattore di Lift e la ROC Curve, hanno mostrato che il modello *Logistic*, basato sulla regressione logistica, risulta moderatamente accurato, al contrario del modello probabilistico *NBTree*.
Per la gestione delle classi fortemente sbilanciate si è infine deciso di testare il metodo Cost-Sensitive Learning, partizionando il dataset con la tecnica del K-fold Cross Validation. In questo ultimo caso non si evincono variazioni rilevanti dei risultati rispetto a quanto ottenuto precedentemente.
Gli esiti raggiunti sono certamente influenzati dai valori della matrice dei costi: trattandosi di un ambito sanitario, è opportuno condurre studi approfonditi circa l'individuazione dei valori ottimali, al fine di massimizzare i risultati. Tale obiettivo può essere perseguito tramite l'uso di tecniche algoritmiche specifiche per il problema che si sta trattando. Inoltre, vi è la possibilità di arricchire i risultati tramite l'introduzione e il confronto di nuovi modelli di classificazione.
Le performance mostrate in questo elaborato non permettono l'usabilità in ambito applicativo, in quanto prossime al modello casuale. Tuttavia, si ritiene che gli studi per la previsione di episodi di ictus possano largamente avvalersi degli strumenti forniti dal Machine Learning, al fine di conoscere meglio questa condizione e sviluppare campagne di prevenzione più efficaci.

RIFERIMENTI BIBLIOGRAFICI

[1] GBD 2013 Mortality and Causes of Death Collaborators. "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013." Lancet (London, England) vol. 385,9963 (2015): 117-71. doi:10.1016/S0140-6736(14)61682-2

[2] Feigin VL, Forouzanfar MH, Krishnamurthi R, Mensah GA, Connor M, Bennett DA, Moran AE, Sacco RL, Anderson L, Truelsen T, O'Donnell M, Venketasubramanian N, Barker-Collo S, Lawes CM, Wang W,

Shinohara Y, Witt E, Ezzati M, Naghavi M, Murray C. Global Burden of Diseases, Injuries, and Risk Factors Study 2010 (GBD 2010) and the GBD Stroke Experts Group. “Global and regional burden of stroke during 1990-2010: findings from the Global Burden of Disease Study 2010.” Lancet. 2014 Jan 18;383(9913):245-54. doi: 10.1016/s0140-6736(13)61953-4. Erratum in: Lancet. 2014 Jan 18;383(9913):218. PMID: 24449944; PMCID: PMC4181600.

[3] https://www.salute.gov.it/portale/salute/p1_5.jsp?area=Malattie_cardiovascolariid=28lingua=italiano

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

[5] https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification/

[6] Swets JA. Measuring the accuracy of diagnostic systems. Science 1998; 240: 1285-93