Modelli di classificazione applicati ad un dataset riguardante diversi casi di demenza.

Portaluppi Alessandro 816090
Testa Luca 816000

Questo progetto riguarda l'applicazione di diversi modelli di classificazione su un dataset contenente 373 risultati di risonanze magnetiche effettuate su soggetti affetti o non affetti da demenza. L'obiettivo del nostro lavoro è quello di trovare un modello di classificazione valido in grado di predire la demenza in base ad altri attributi presenti all'interno del dataset.

I. Introduzione

La demenza è un termine generico utilizzato per descrivere un declino delle facoltà mentali sufficientemente grave da interferire con la vita quotidiana. La perdita di memoria è un esempio di questo declino. Il morbo di Alzheimer rappresenta la più comune tipologia di demenza. Il morbo di Alzheimer rappresenta il 60-80 per cento dei casi.
La demenza è causata da danni subiti dalle cellule cerebrali. Questo danno interferisce con la capacità delle cellule cerebrali di comunicare tra loro. Quando le cellule cerebrali non possono comunicare normalmente, il pensiero, il comportamento e le sensazioni ne risentono. [4]
Secondo le statistiche più recenti, nel nostro Paese ne sono affette circa un milione di persone: la fascia d'età più colpita è quella degli over 65, con un'incidenza che raggiunge il 5% del totale. Il legame tra vecchiaia e demenza è confermato dall'aumento dei casi in rapporto all'età, che tra gli ultraottantenni oltrepassano il 30%. Purtroppo, questi numeri appaiono destinati a crescere, come conseguenza del graduale aumento dell'aspettativa di vita media. Secondo i dati forniti dall'UE, il numero dei pazienti affetti da demenza crescerà ulteriormente, passando dagli attuali 47 milioni agli oltre 63 milioni stimati per il 2030. [5]
È dunque obiettivo primario di questo progetto riuscire a prevedere se è un soggetto può essere affetto da demenza mantenendo un'elevata accuratezza. Inizialmente presenteremo il dataset con le variabili incluse, dopodichè applicheremo quattro modelli di classificazione distinti per attuare le nostre previsioni ed infine analizzeremo i risultati ottenuti per rispondere alla nostra domanda di ricerca.

II. Descrizione del dataset

Il dataset [3] è composto dai seguenti campi:

- Subject_id: identificatore del soggetto.
- MRI_id: identificatore dell'esame di risonanza.
- Group: variabile di classe che assume valore *Demented, Non Demented o Converted*.
- Visit: numero progressivo di visite del soggetto.
- MR Delay: tempo di esecuzione della risonanza magnetica.
- M/F: genere del soggetto.
- Hand: mano predominante.
- Age: età del soggetto.
- EDUC: anni dedicati allo studio da parte del soggetto.
- SES: indicatore socio - economico.
- MMSE: test di valutazione della demenza.
- CDR: rating di demenza.
- eTIV: volume intracranico.
- nWBV: volume totale del cervello.
- ASF: fattore di scala.

III. Analisi esplorativa

Innanzitutto osserviamo come il dataset contenga variabili oggettivamente ininfluenti come Subject_ID ed MRI_ID, come prima manipolazione sono quindi state eliminate le suddette colonne.

Abbiamo notato inoltre che la variabile CDR era direttamente associata al gruppo di appartenenza tramite un semplice test. Testando i modelli di classificazione con il CDR si otteneva sempre un'accuratezza pari ad 1, quindi per ottenere dei risultati più rilevanti ai fini del lavoro abbiamo deciso di non considerarla durante la fase di train.
In ultima analisi, abbiamo applicato una binarizzazione della variabile *group*: essa infatti assumeva valori nell'insieme *[non demented, demented, converted]* per indicare rispettivamente un soggetto in cui non sia stata rilevata demenza, uno in cui sia stata rilevata ed infine i casi di cambiamenti di diagnosi in visite differenti. Quest'ultimo caso è stato alterato in *demented* o *non demented* a seconda del valore del CDR (attraverso il nodo knime *rule engine*, di cui sotto riporto la configurazione).

```
$Group$ = "Converted" AND $CDR$ > 0 => "Demented"
$Group$ = "Converted" AND $CDR$ = 0 => "Nondemented"
$Group$ = "Demented" => "Demented"
$Group$ = "Nondemented" => "Nondemented"
```

Fig. 1. Binarizzazione del valore group

Analizzando le statistiche ottenute sul dataset, osserviamo che non ci troviamo di fronte ad un problema di sbilanciamento delle classi, in quanto la variabile Group è ripartita non a favore di alcuna classe.

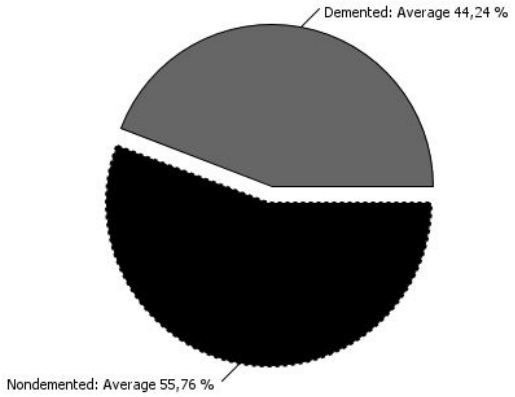

Fig. 2. Suddivisione variabile di classe

IV. Misure di Performance

Nella nostra analisi sono stati utilizzati più criteri in grado di valutare la performance. In particolare, si è scelto di calcolare: Accuracy, Recall, Precision e F1-measure.
L'Accuracy indica la percentuale di osservazioni positive e negative previste correttamente e permette di selezionare l'istanza che garantisce la miglior performance sui record da prevedere:

$$\textbf{\textit{Accuracy}} : \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Dove TP e TN indicano il numero di istanze classificate correttamente come appartenenti rispettivamente alla classe positiva e negativa; FP e FN indicano il numero di istanze positive e negative classificate erroneamente.
L'indicatore Recall rappresenta la porzione di record positivi correttamente classificati dal modello. In particolare:

$$\textbf{\textit{Recall}} : \frac{TP}{TP + FN} \tag{2}$$

L'indicatore Precision descrive la frazione di record che sono effettivamente positivi tra tutti quelli predetti come tali:

$$\textbf{\textit{Precision}} : \frac{TP}{TP + FP} \tag{3}$$

Precision e Recall sono due metriche di valutazione del modello estremamente importanti. Mentre la precision si riferisce alla percentuale dei risultati rilevanti, la recall si riferisce alla percentuale dei risultati rilevanti correttamente classificati dall'algoritmo. Sfortunatamente, non è possibile massimizzare entrambe queste metriche contemporaneamente, poiché una viene a costo di un'altra. [7]
Per semplicità, è disponibile un'altra metrica, chiamata F-Measure, che è la media armonica tra la Precision e la Recall.

$$\textbf{\textit{F-Measure}} : \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

V. Modelli di classificazione convalidati con metodo holdout

In questo studio sono state implementate diverse tecniche di classificazione con lo scopo di individuare la più adatta, sulla base dei dati disponibili. Sono stati utilizzati due modelli euristici basati su alberi decisionali:
Random Forest: classificatore costituito da numerosi alberi di decisione.
J48: consente di classificare anche i dati nominali.
E due modelli probabilistici:
NBTree: genera un albero di decisione attraverso il classificatore Naive Bayes.
Naive Bayes: classificatore basato sul teorema di Bayes.[1]

Per valutare e comparare l'efficienza degli algoritmi di classificazione, è stato utilizzato il metodo *holdout*, che prevede la divisione del dataset in due parti disgiunte (train, test) seguendo la proporzione (70, 30). Viene eseguito quindi il fit dei modelli utilizzando la porzione di train, e successivamente validati con i valori disponibili nella sezione di test.

A. Applicazione senza Feature Selection

In prima analisi, abbiamo deciso di non applicare alcun algoritmo di Feature Selection per poi ottenere una comparazione dei risultati. Sono stati quindi utilizzati tutti gli attributi nella fase di train ad eccezione di Subject_id, MRI_ID e CDR. Per valutare l'efficienza dei classificatori sono state utilizzate le seguenti quattro misure calcolate in base all'accuratezza e ai tassi di errore. Per i valori di Precision, Recall ed F-Measure è stata computata la media dei risultati ottenuti da entrambi i valori della variabile di classe.

Classificatore	Recall	Precision	F - Measure	Accuracy
NBTree	0.722	0.731	0.725	0.732
J48	0.739	0.752	0.742	0.750
NaiveBayes	0.741	0.778	0.744	0.759
RandomForest	0.737	0.755	0.739	0.750

TABLE I
Classificazione senza Feature Selection.

Come possiamo osservare, nessun modello di classificazione raggiunge risultati molto elevati in termini di accuratezza. Questo probabilmente perché non è possibile ottenere un classificatore in grado di prevedere con un tasso di errore relativamente basso se una persona è affetta da demenza o meno. In ogni caso, NaiveBayers è l'algoritmo che ottiene i risultati migliori nel test effettuato senza selezione delle features.

B. Applicazione con Feature Selection

Con l'obiettivo di selezionare le variabili più performanti e ridurre il numero di attributi in input, migliorando l'interpretabilità dei dati, è stata applicata la feature selection con metodo CfsSubsetEval che effettua una valutazione dei singoli attributi prima di sottoporli al classificatore. In questo modo, tramite filtro multivariato, vengono selezionate le variabili che maggiormente influenzano la variabile risposta senza trascurare la correlazione tra le stesse. Come effettuato per la classificazione con metodo Holdout, anche in questo caso, è stato partizionato il dataset iniziale in training set (70% dei record) ed in test set (30% dei record), grazie al procedimento di stratified sampling, selezionando *group* come variabile etichetta. Sono stati utilizzati tutti gli attributi decisi dall'*AttributeSelectedClassifier*: MF, EDUC, MMSE, nWBV. Per valutare l'efficienza dei classificatori sono state utilizzate le quattro misure precedentemente presentate. Anche in questo caso per i valori di Precision, Recall ed F-Measure è stata computata la media dei risultati ottenuti da entrambi i valori della variabile di classe.

Classificatore	Recall	Precision	F - Measure	Accuracy
NBTree	0.767	0.779	0.770	0.777
J48	0.719	0.736	0.721	0.732
NaiveBayes	0.764	0.832	0.766	0.786
RandomForest	0.749	0.747	0.747	0.750

TABLE II
Classificazione con Feature Selection

In questo caso possiamo notare come l'accuratezza migliore è sempre raggiunta dal modello di NaiveBayes. I livelli di accuratezza sono sempre e comunque non elevati.
Il grafico seguente riassume le variazioni del

parametro *accuracy*, utilizzando o meno tecniche di feature selection.

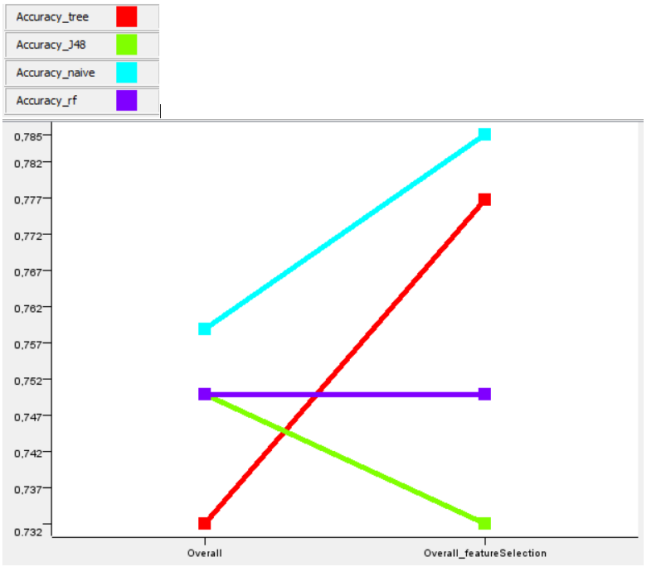

Fig. 3. Comparazione valori accuracy

Come possiamo vedere, nonostante la riduzione della dimensionalità andando a rimuovere attributi considerati non influenti dal classifier, in due casi su quattro si ottengono risultati leggermente migliori in termini di accuratezza.
Peggioramento nel caso dell'algoritmo *J48*, e nessuna differenza per quanto riguarda *RandomForest*.

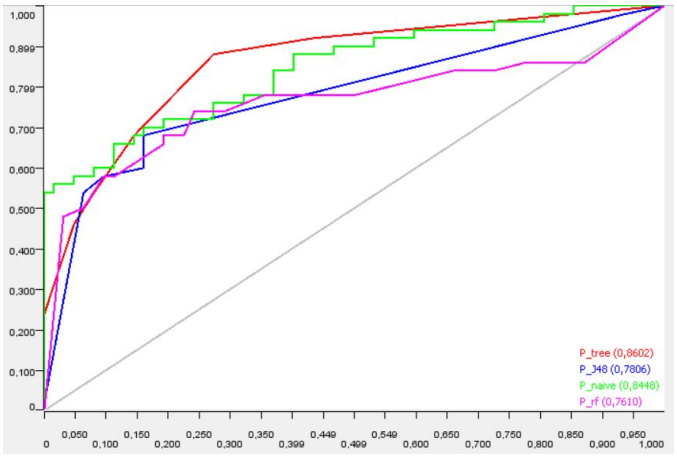

Fig. 4. Comparazione ROC curve

In ultima analisi, viene riportato un grafico [fig. 4] che confronta le ROC curve relative ai quattro modelli di classificazione.

Come possiamo vedere, tutti i classificatori sono al di sopra della retta relativa al modello di classificazione casuale, il che significa che tutti i modelli, pur avendo un'accuratezza non troppo elevata, sono comunque migliori di un classificatore casuale.
Il grafico conferma inoltre che i classificatori bayesiani ottengono un valore AUC (Area Under Curve) più elevato rispetto ai modelli ad albero.

VI. Modelli di classificazione convalidati con metodo cross validation

Per ottenere una ulteriore riprova sull'efficienza dei modelli di classificazione, abbiamo deciso di attuare un aggiuntivo metodo per la convalida dei risultati. In particolare, il metodo utilizzato è la *10-fold cross validation*. Per ogni classificatore, il dataset è stato quindi diviso in 10 sottoinsiemi disgiunti, applicando sempre un campionamento stratificato sulla variabile *group*.
Ogni partizione è stata poi adoperata come validation set, utilizzando le restanti folds come training set, combinando poi i risultati e analizzando le distribuzioni. L'accuratezza finale era calcolata effettuando la media dei valori di accuracy ottenuti per ognuna delle dieci folds. In particolare, per ogni classificatore, sono stati ottenuti i seguenti risultati:

Classificatore	**Accuracy**
NBTree	0.777
J48	0.775
NaiveBayes	0.791
RandomForest	0.797

TABLE III
Media dei valori di accuracy

Per avere una migliore percezione delle differenze tra i valori, viene proposto diagramma di boxplot[Fig. 5] che mostra la distribuzione dei risultati ottenuti tramite il test cross-validation.

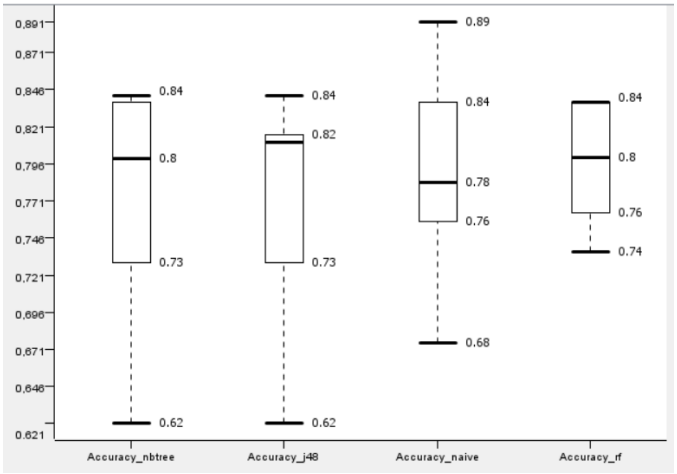

Fig. 5. Comparazione distribuzioni accuracy

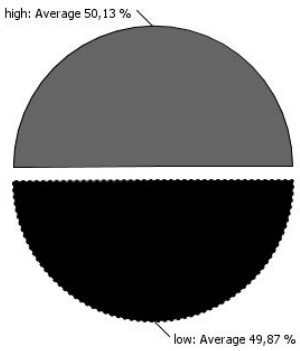

Fig. 6. Bilanciamento della classe EDUC binarizzata

In questo il test sembra confermare che *RandomForest* e *NaiveBayes* risultano essere gli algoritmi leggermente migliori in termini di accuracy. É stato comunque deciso inoltre di effettuare un test statistico *ANOVA* per studiare la significatività delle differenze rilevate. [2] Il test si basa infatti sull'ipotesi nulla H0: *le medie sono uguali, e le differenze sono dovute al caso*, impostando un livello di significatività pari a 0.05.
Con un valore critico F = 0.27043 abbiamo ottenuto un p-value pari a **0.846292**, che è maggiore del livello di significatività, non possiamo quindi rifiutare l'ipotesi nulla H0, ovvero che le differenze non sono statisticamente significative.

VII. Seconda domanda di ricerca

Come seconda domanda di ricerca, ci siamo posti il quesito se l'essere soggetti a patologie relative alla demenza, potesse essere correlato al grado di educazione posseduto dalla persona.
In fase di preparazione, attraverso il nodo *rule engine* abbiamo applicato la binarizzazione della variabile EDUC, in due macrocategorie: *low* se risultava essere minore o uguale a 14, *high* altrimenti. Così facendo abbiamo ottenuto un elevato bilanciamento delle variabili del dataset.

L'obiettivo era quello di testare i modelli di classificazione presentati in precedenza, scegliendo come variabili esplicative solamente quelle correlate allo stato mentale della persona, ovvero *CDR, MMSE e Group*, con lo scopo di prevedere il livello di educazione (high, low).
In altre parole: *se un soggetto soffre di demenza, è anche dovuto al fatto che ha studiato poco?*
Come nella ricerca precedente, abbiamo applicato il metodo holdout, con un partizionamento 70-30, e comparato i risultati dei modelli di classificazione: i risultati sono i seguenti.

Classificatore	Recall	Precision	F - Measure	Accuracy
NBTree	0.536	0.538	0.528	0.536
J48	0.535	0.536	0.532	0.536
NaiveBayes	0.536	0.538	0.528	0.536
RandomForest	0.535	0.541	0.520	0.536

TABLE IV
Classificazione seconda domanda di ricerca

Come possiamo osservare, i classificatori raggiungono il medesimo livello di accuratezza, che si attesta a 0.536, portandoci a pensare che attraverso il dataset fornito, non si è in grado di predire il livello di educazione, dato lo stato mentale del paziente.

VIII. Conclusioni e Sviluppi Futuri

In questo lavoro si è voluto identificare la tecnica di Machine Learning migliore per la classificazione dell'occorrenza sulla presenza di demenza in un soggetto al fine di prevedere se, tramite determinati valori assunti dalle variabili presenti nel dataset, il soggetto potesse avere o meno tale patologia.
Data la fragilità dell'argomento ci si aspettava di trovare un livello di accuratezza non elevato, in quanto è sempre difficile fare una previsione su

argomenti sanitari. Per verificare i livelli di accuratezza si è deciso di utilizzare il metodo *HoldOut*, utilizzando o meno *FeatureSelection*, al fine di verificare un aumento di tale livello. Abbiamo osservato come, i modelli probabilistici *NBTree e NaiveBayes*, abbiano ottenuto un leggero miglioramento, e abbiano dimostrato di classificare con migliore accuratezza la domanda del nostro problema.
Come secondo metodo di valutazione, si è applicata la *CrossValidation*, essa ha portato ad ottenere un miglior valor medio di accuratezza sempre grazie al classificatore probabilistico NaiveBayes ma senza un aumento generale dei livelli di accuratezza.
Per la seconda domanda di ricerca si è invece deciso di verificare una corrispondenza tra il livello di educazione di un dato soggetto e la presenza di demenza o meno. In questo caso abbiamo riscontrato la difficoltà di previsione del livello di educazione in base ai valori clinici sulla demenza, e non siamo arrivati ad una risposta certa.
Il lavoro affrontato può suggerire alcuni sviluppi e migliorie future.
Ad esempio, tecniche di Machine Learning come la Cluster Analysis possono rivelarsi utili ad individuare raggruppamenti di soggetti con caratteristiche simili. Un dataset con più informazioni fisiologiche, come ad esempio: misura della glicemia, analisi delle urine, test tossicologici, analisi del liquido cerebrospinale, misura degli ormoni tiroidei[6], avrebbe probabilmente portato a risultati ancora più accurati.

REFERENCES

[1] URL: https://nodepit.com/category/analytics/mining/weka/weka3.7/classifiers.

[2] URL: https://www.socscistatistics.com/tests/anova/default2.aspx.

[3] Jacob Boysen. "MRI and alzheimer". In: (2018). URL: https://www.kaggle.com/jboysen/mri-and-alzheimers.

[4] "Che cos'è la demenza?" In: (). URL: https://www.alz.org/it/cosa-e-la-demenza.asp.

[5] epiCura. "I sintomi della demenza senile". In: (2019). URL: https://www.epicuramed.it/blog/iii-eta/i-sintomi-della-demenza-senile/.

[6] Antonio Griguolo. "Demenza Senile". In: (2019). URL: https://www.my-personaltrainer.it/salute-benessere/demenza-senile.html.

[7] Shruti Saxena. "Precision vs Recall". In: (2018). URL: https://towardsdatascience.com/precision-vs-recall-386cf9f89488.