Analisi classificativa delle patologie tiroidee principali

ABSTRACT

La tiroide è il più grande viscere endocrino presente nell'organismo umano. Situata nel collo, produce e secerne ormoni utili per regolarizzare svariate funzioni corporee. Le patologie legate alla tiroide sono piuttosto frequenti nella popolazione mondiale. Tra queste, l'ipotiroidismo e l'ipertiroidismo sono le più note e comuni.
Attraverso uno studio di classificazione si andrà a valutare se e quali tra i fattori considerati incidano sulla diagnosi di malattie tiroidee. Verrà posta l'attenzione sul confronto tra modelli con target binario, utili per la classificazione e previsione di soggetti ipotiroidei.

Lavoro a cura di:

Gabellini, Luca
Mercandelli, Manuel Luca
Provasi, Matteo
Ronco, Giuliano

INDICE

1. INTRODUZIONE

La tiroide è un organo a forma di farfalla posto al centro del collo; il suo compito principale è quello di secernere ormoni, la tri-iodotironina (o T3) e la tiroxina (T4).
La T4 rappresenta circa il 90% del totale di ormoni prodotti dalla tiroide, ed è una forma inattiva della T3. Gli ormoni tiroidei stimolano i processi cosiddetti anabolici, vale a dire di crescita, sviluppo e movimento dell'organismo. Inoltre, svolgono un'azione di controllo sugli enzimi che presiedono al metabolismo energetico[1].
La tiroide funziona correttamente, garantendo un'adeguata sintesi ormonale, se può disporre di adeguate quantità di iodio, un oligoelemento essenziale, che entra nella costituzione della tiroxina e della tri-iodotironina. Un basso apporto di iodio nella dieta può causare la comparsa del gozzo colloidale, caratterizzato da un aumentato volume della ghiandola tiroidea[2].
Un altro ormone di fondamentale importanza in questo contesto è il TSH, o ormone tireotropo, secreto dall'ipofisi, che controlla e stimola l'organo tiroideo nella produzione di T3 e T4.
Le patologie legate alla tiroide sono piuttosto frequenti nella popolazione mondiale. Tra queste, l'ipotiroidismo e l'ipertiroidismo sono le più note e comuni. L'ipertiroidismo porta ad un aumento dell'azione degli ormoni tiroidei, con conseguente aumento del metabolismo, perdita di peso, aumento dell'appetito, tachicardia e un maggior sviluppo tiroideo. L'ipotiroidismo, invece, porta ad un ridotto metabolismo con conseguente bassa temperatura, aumento di peso, riduzione dell'appetito, bradicardia, ipotensione, ipotonia della muscolatura scheletrica e apatia.

Nei paesi meno sviluppati è più comune la forma di ipotiroidismo causata da un'insufficienza di iodio nella dieta, mentre nei paesi più industrializzati è più frequente un ipotiroidismo in una forma autoimmune. La terapia da seguire, in questi casi, è quasi ed esclusivamente ormonale, relativamente complicata in quanto non è possibile aumentare o ridurre la stimolazione in modo diretto sulla tiroide.
L'obiettivo di questo progetto sta nel classificare e prevedere in maniera soddisfacente i soggetti malati in funzione delle esplicative considerate, cercando di comprendere quali siano i fattori più importanti associati alla diagnosi di malattie tiroidee.
Per queste ragioni il dataset analizzato è caratterizzato da variabili relative a misurazione di livelli ormonali o a particolari condizioni in cui versa il soggetto (es. stato di gravidanza, presenza di gozzo).

2. DATASET E PREPROCESSING

Descrizione Il dataset analizzato proviene da uno studio clinico su un campione di individui di Sydney, Australia volto a valutare il funzionamento della tiroide e l'eventuale presenza di ipo-ipertiroidismo; i dati sono stati raccolti in un arco di tempo che va dal 1984 al 1987[3].
Nel dataset erano presenti più di 9000 record con 29 variabili rilevate e frequenti dati mancanti. Per rendere più efficiente l'analisi sono state fatte delle trasformazioni alle variabili:

- La variabile relativa al genere dell'individuo è stata unita con quella indicante se il soggetto fosse in fase di gravidanza oppure no. In questo modo si è ottenuta un'unica variabile (status) a tre valori, per indicare se il soggetto fosse maschio, femmina oppure femmina in stato di gravidanza.
- Relativamente ad alcuni ormoni erano presenti due variabili: una con il valore rilevato, ed un'altra dicotomica indicante se quell'ormone fosse stato misurato oppure no. Quest'ultime, poiché ridondanti, e non utili ai fini dell'analisi, sono state eliminate.
- La variabile target originale, diagnosis, era strutturata da un codice alfanumerico con cui si indicava la diagnosi precisa del soggetto. Dato che in questa situazione il target avrebbe presentato un elevato numero di modalità, e delle diagnosi eccessivamente specifiche, come ad esempio differenziare fra ipotiroidismo primario, secondario o subclinico, che andavano oltre gli scopi dell'analisi, attraverso uno script in RStudio si sono riclassificate le diagnosi in tre livelli: soggetto sano, ipotiroideo o ipertiroideo.

Dati mancanti Attraverso un nodo di statistiche sono stati conteggiati i valori mancanti di ogni variabile. A seconda delle situazioni si sono prese differenti decisioni. Le modifiche apportate sono state le seguenti:

- La variabile TBG, relativa all'omonima proteina che permette la circolazione degli ormoni tiroidei nel flusso sanguigno, è stata eliminata in quanto i valori risultavano mancanti nel 90% dei soggetti. Per questo motivo e anche per il fatto che i soggetti in cui la TBG non era stata rilevata presentavano dati missing in altre variabili, si è ritenuto inopportuno effettuare un'accurata sostituzione con qualsiasi metodo.
- Nel paragrafo precedente si è parlato della creazione della variabile status unendo le variabili genere e gravidanza; per quanto riguarda la prima erano presenti dati mancanti, probabilmente legati a motivi di privacy. Non potendo avere una forte sicurezza su un'eventuale sostituzione, soprattutto per i pazienti sani, e considerando il numero esiguo di casi in questa situazione, si è deciso di eliminare queste osservazioni attraverso uno script in RStudio.
- Sempre con RStudio si è calcolato il numero di missing per ogni osservazione e successivamente su KNIME si sono eliminati i record che presentavano valori mancanti in tre o più variabili in quanto sarebbe stato eccessivamente approssimativo tenere delle osservazioni in cui una buona parte dei valori non è stato osservato ma stimato.
- Per le variabili TSH e T3 si è scelta la soluzione dell'imputazione tramite un'interpolazione lineare, in quanto l'altra opzione presa in considerazione, il median replacement, non sarebbe stata una scelta affidabile vista la forte asimmetria che queste variabili presentano. Il valore da sostituire è stato ottenuto mediante un modello di regressione lineare multipla, utilizzando altre variabili del dataset come inputs. Le variabili esplicative sono state scelte principalmente in base alla correlazione tra queste ed il target da stimare. Il modello utilizzato è il seguente:
$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}$$
Dove $\mathbf{x}_i^T$ è il vettore delle variabili esplicative relative alla $\boldsymbol{i}$-esima osservazione, mentre $\boldsymbol{\beta}$ è il vettore dei coefficienti.
- Per le variabili T4U, TT4 e FTI, vista la buona simmetria, è stata utilizzata un'imputazione del valore mediano.

Modifica target Originariamente era stata pensata un'analisi con un target a tre livelli: pazienti sani, ipertiroidei ed ipotiroidei. Tuttavia per utilizzare più strumenti di analisi si è deciso di utilizzare un target binario senza considerare i pazienti ipertiroidei. La scelta di escludere questa categoria piuttosto che gli ipotiroidei si basa su due importanti osservazioni: in primis la numerosità di questa classe è minore ed avremmo avuto un dataset troppo sbilanciato; in secondo luogo la valutazione dell'ipotiroidismo nei paesi del primo mondo, come è l'Australia, assume maggiore rilievo medico in quanto è spesso associata con la tiroide di Hashimoto, malattia autoimmune solitamente sottodiagnosticata[4]. Nonostante ciò, per non sprecare le

informazioni derivanti dagli ipertiroidei, è stato condotto uno studio meno approfondito del primo utilizzando modelli con target a tre livelli.

Dati anomali Sempre attraverso il nodo di statistiche si è valutata la distribuzione delle variabili, con riferimento anche ai valori minimi e massimi. Si sono riscontrati dei valori anomali in una ventina di osservazioni, probabilmente relativi a degli errori di codifica, ad esempio valori di 65 mila anni per l'età. Non potendo risalire al reale valore, queste osservazioni sono state eliminate.
Utilizzando dei box plot si sono riscontrati dei valori anomali per le variabili continue, come TSH e FTI (tiroxina libera). Si è deciso di non eliminare queste osservazioni in quanto la condizione clinica del soggetto era concorde con valori alti o bassi degli ormoni e quindi si trattava di casi reali comunque utili per l'analisi, non riconducibili ad errori di imputazione.

Feature selection Prima di procedere con una feature selection si è tenuto conto dei risultati della correlazione svolta in fase di imputazione di valori con modello di interpolazione. Si è notato che le variabili FTI e TT4, rispettivamente indice di tiroxina libera e la tiroxina totale, erano fortemente correlate fra di loro, come si poteva facilmente immaginare. Al fine di evitare il fenomeno della multicollinearità, si è deciso di scartare TT4 e di tenere FTI in quanto più specifica e più considerata negli studi clinici degli ultimi anni rispetto ai valori totali[5].
E'stata quindi proposta una feature selection con metodo wrapper, per ridurre la dimensionalità complessiva e scartare variabili irrilevanti.
Attraverso un Naive Bayes Tree si sono valutate le accuratezze del modello in base a dei subset delle variabili in un dataset di test. La procedura parte con la selezione di una variabile, quella che permette di avere un minor errore di classificazione e prosegue selezionando una variabile per volta, aggiungendo sempre quella che rende minimo l'errore. I problemi di dimensionalità avvengono quando all'aumentare delle variabili selezionate, l'errore di classificazione non diminuisce o aumenta.

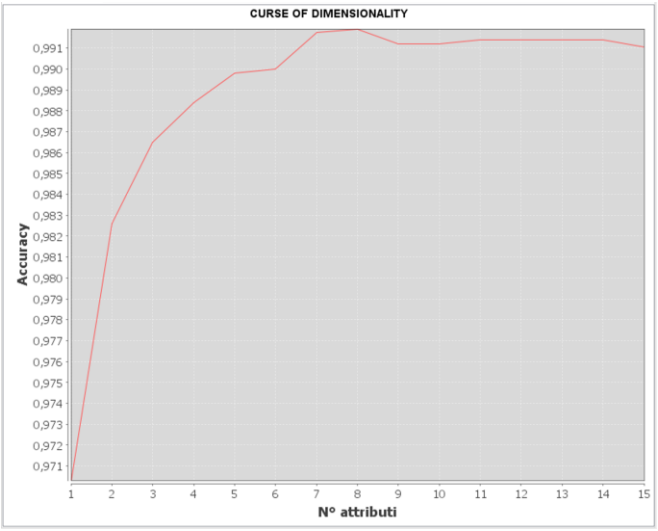

Figura 1: *Feature selection*

L'immagine riporta i risultati della selezione mediante il wrapper: con circa metà delle variabili totali selezionate il modello raggiunge la massima accuratezza, poi cala leggermente e si stabilizza. La scelta ottimale ricadrebbe sull'utilizzo di otto variabili, ma poiché l'incremento di accuratezza ottenuto dal passaggio da sette ad otto attributi è molto contenuto, si è deciso di considerare solamente sette variabili.
Il metodo di wrapper prevede la scelta di un modello di classificazione per la selezione del miglior subset di variabili e quindi i risultati possono essere influenzati dalla scelta del modello. Per avere un confronto, si è deciso di procedere con un ulteriore wrapper utilizzando questa volta un modello logistico. A differenza del Naive Bayes Tree i risultati di accuratezza con questo modello erano peggiori, ma la scelta delle variabili da inserire all'interno del subset seguiva un ordine quasi identico. In base a questo confronto e anche al fatto che l'accuratezza ottenuta con il primo modello risultava ottima si è deciso di mantenere i risultati relativi al primo modello come riferimento. Più in particolare, delle 15 variabili disponibili, la feature selection ha portato all'individuazione del seguente subset:

- **TSH**: quantitativa, livelli di ormone tireotropo nel sangue.
- **Cura tiroxina**: binaria, indica se il soggetto è sotto cura a base di tiroxina.
- **FTI**: quantitativa, indice di tiroxina libera.
- **Operazione tiroide**: binaria, indica se il soggetto ha subito un'operazione alla tiroide.
- **T4U**: quantitativa, livello di assorbimento di tiroxina
- **Cura I131**: binaria, indica se il soggetto è sotto cura a base di iodio I131.
- **T3**: quantitativa, livelli di tri-iodotironina nel sangue.

Le variabili sono ordinate in base all'ordine di inclusione dell'algoritmo wrapper. E' importante osservare come la variabile TSH, da sola, porti a una significativa riduzione dell'errore atteso.

3. MODELLI CLASSIFICATIVI

Sono stati implementati diversi modelli per assolvere al problema di classificazione di nostro interesse.
Per ottenere stime robuste degli indici di valutazione della bontà classificativa, si è optato per l'utilizzo di una cross-validation. Con questa procedura il dataset viene suddiviso in un numero arbitrario k di parti, nel nostro caso 10, e ad ogni step *k-1* partizioni vanno a costituire il training set, mentre la rimanente fa da test set su cui poi sono previsti i valori della variabile target, che si ipotizza ignota.
Per poter confrontare fra loro modelli diversi, si è selezionata l'opzione di random seed con la quale si crea una partizione pseudo casuale del dataset. In questo modo ogni nodo di cross

validation relativo al modello effettua la suddivisione del dataset nella stessa identica maniera. Questo permette di poter effettuare dei confronti più accurati.
Inoltre si è deciso di optare per un campionamento stratificato: tutte le partizioni, oltre ad avere lo stesso numero di osservazioni, presenteranno anche una proporzione di ipotiroidei pressoché uguale a quella del dataset completo.

Modelli su KNIME I modelli classificativi sono stati raggruppati in due metanodi per facilitare la visualizzazione delle curve ROC. Nel primo metanodo sono stati inseriti alberi classificativi e modelli logistici. È stato implementato un albero classificativo con un nodo di base specificando il gain ratio come funzione di riferimento per lo splitting e un metodo di pruning per evitare overfitting. Si è creato inoltre un albero di classificazione J48 che implementa un algoritmo leggermente diverso volto ad ottimizzare l'information gain normalizzato. Non sono presenti differenze sostanziali nell'algoritmo dei due alberi, il J48 è più versatile in quanto è in grado di eseguire classificazioni utilizzando delle variabili che hanno dei pesi, che però non sono state identificate nel nostro caso, e permette anche la gestione dei dati mancanti, che sono già stati eliminati o imputati secondo una delle tecniche esposte nel capitolo precedente. L'unica sostanziale differenza fra i due metodi è l'implementazione dell'algoritmo di potatura alla fine della procedura e il criterio di splitting. Poiché il nodo J48 utilizza un algoritmo creato più recentemente rispetto agli alberi tradizionali, è logico aspettarsi delle performance classificative migliori.
Per i modelli logistici, anche in questo caso ne sono stati implementati due: un modello di base ed un modello che utilizza un algoritmo sviluppato da Weka.
E' stato poi considerato un nodo per l'implementazione di un algoritmo di random forest. Questo metodo cerca di superare quello che è forse il limite principale degli alberi, ovvero la scarsa robustezza. Con questo metodo vengono creati diversi alberi classificativi e le osservazioni sono assegnate al livello del target con cui sono state classificate la maggior parte delle volte. Per evitare di creare sempre lo stesso albero, le foreste casuali eseguono ogni volta una selezione di un subset casuale di variabili con le quali poi sarà creato l'albero. Nonostante questo approccio possa sembrare poco intuitivo poiché si rischia di prendere un subset senza le variabili migliori per la classificazione, di norma questo metodo non solo garantisce più robustezza ma permette anche di ottenere dei livelli di accuratezza migliori.
Nel secondo metanodo sono stati implementati due modelli appartenenti alla classe dei support vector machine, nello specifico SMO e SPegasos. Data la natura dei modelli che ammettono solo la gestione di variabili quantitative, sono attendibili delle performance peggiori rispetto ad altri classificatori, anche se le variabili continue, quasi esclusivamente gli ormoni e le proteine, risulteranno gli attributi più significativi per la classificazione.
Infine sono stati implementati modelli come Multilayer perceptron e Naive Bayes.
La Multilayer Perceptron è un modello rientrante nella classe delle reti neurali che usa un algoritmo di learning supervisionato. Questi modelli utilizzano dei layer e delle funzioni di attivazione non lineare per effettuare il loro compito di classificazione. Anche se il funzionamento è profondamente diverso, il risultato è comparabile con la suddivisione effettuata dalle SVM, con la particolarità che questi modelli sono più flessibili.
Le reti bayesiane invece sono dei modelli probabilistici grafici che esprimono la dipendenza condizionale di un set di variabili.

4. ANALISI DEI RISULTATI

In questa sezione verranno messe in luce le performance classificative dei modelli precedentemente presentati, in termini di accuracy, recall e AUC.
E' da sottolineare come tutti i modelli implementati raggiungano un'accuracy molto elevata, vicina all'unità.

Models	*Accuracy*
Simple logistic (weka)	0.955
Decision tree	0.989
J48	0.987
Random forest	0.990
Logistic regression	0.956
Bayes net	0.974
SMO	0.952
Spegasos	0.941
MLP	0.975
NB	0.959

Tabella 1: *Accuratezza nei modelli con target binario*

Bisogna però ricordare che, per quanto riguarda la nostra analisi, è ben più importante classificare correttamente le osservazioni a cui è associata la classe meno frequente del target, ovvero gli ipotiroidei. E' quindi necessario considerare, oltre all'accuracy, anche la recall, che corrisponde alla proporzione di individui ipotiroidei correttamente classificati.

Models	*Accuracy*	*Recall*
Decision tree	0.989	0.975
Random forest	0.990	0.973
J48	0.987	0.956
Bayes net	0.974	0.806
MLP	0.975	0.742
NB	0.959	0.553
Logistic regression	0.956	0.467
SMO	0.952	0.44
Simple logistic (weka)	0.955	0.436
Spegasos	0.941	0.224

Tabella 2: *accuratezza e recall nei modelli con target binario*

Dalla tabella, ordinata in base ai valori della Recall, emergono i modelli migliori: Decision Tree, Random Forest e J48, tutti riconducibili alla stessa famiglia (alberi classificativi). Come da previsione, i modelli SMO e Spegasos risultano tra i peggiori. Questi infatti, come riportato in precedenza, non accettano esplicative categoriali, e sono quindi stati implementati con un numero di inputs minore rispetto agli altri modelli.
Un'efficace visualizzazione grafica che permette di fare confronti tra modelli è data dalle curve ROC. Per semplificare la visualizzazione, verranno presentate le curve relative ai modelli migliori per ogni famiglia di classificatori. Le curve rappresentano un plot della percentuale di veri positivi (asse y) contro falsi positivi (asse x). I modelli migliori sono quelli le cui curve saranno più aderenti all'asse verticale, ovvero quelli con AUC (area under the curve) maggiore.

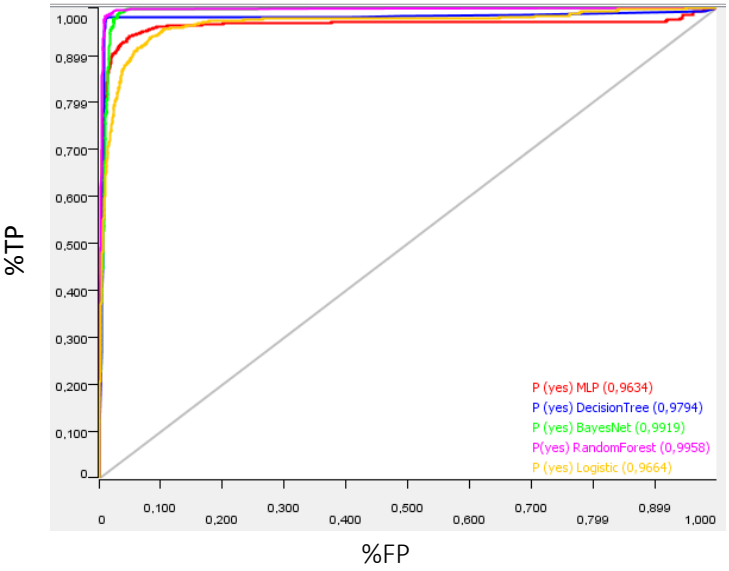

Figura 2: *curve ROC dei modelli migliori*

Nel nostro caso i modelli rappresentati presentano tutti delle curve ampie, con valori di AUC vicini all'unità. Tra i migliori abbiamo il Random Forest (AUC=0.9958) e il Bayes Network (AUC=0.9919).

5. COSTI DI CLASSIFICAZIONE

Come osservato i modelli sviluppati ottengono delle ottime performance classificative. Nonostante il target sia sbilanciato, non tutti i modelli seguono quella che è la ZeroR Rule, ovvero il fenomeno secondo cui un modello di classificazione ottiene delle eccellenti performance solamente per il fatto che i livelli del target sono molto sbilanciati e quindi classifica correttamente le osservazioni con il livello più comune e viceversa.
Anche se il numero di osservazioni classificate erroneamente è esiguo, è verosimile che in ambito reale esistano dei costi dovuti a errori di classificazione. Non essendo stata fornita nessuna matrice dei costi insieme al dataset, si è deciso di imputare dei valori secondo un ragionamento logico: è più grave classificare pazienti malati come sani e meno grave il contrario.
A differenza dell'analisi precedente si è utilizzato uno specifico nodo che permetteva l'imputazione di una matrice di costi: per i pazienti classificati correttamente non è stato scelto nessun costo, per i pazienti sani classificati come malati si è scelto un costo unitario, mentre per il caso opposto il costo è dieci volte tanto.
I modelli sono stati costruiti secondo un algoritmo che ha lo scopo di minimizzare il costo di classificazione piuttosto che l'accuratezza. Attraverso un nodo di formula è stato calcolato il costo di errata classificazione e sono stati fatti dei confronti sulla base di quest'ultimo.

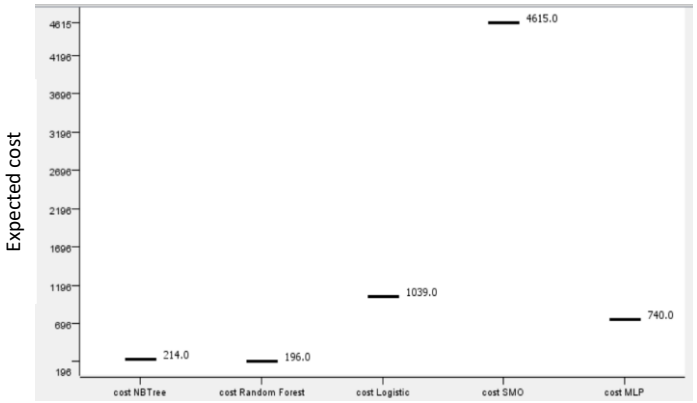

Figura 3: *Costo di classificazione*

Anche in questo caso il modello Random Forest si è dimostrato il migliore, avendo un costo leggermente minore rispetto all'albero classificativo. Abbastanza deludenti gli altri modelli: la Multi Layer Perceptron, ha un costo che è più del triplo rispetto ai primi due; il modello logistico è ancora peggio, mentre al mo dello SMO corrisponde un costo molto più grande in valore assoluto rispetto a tutti gli altri.
E' già stato accennato in precedenza il fatto che gli SVM learners non accettino esplicative categoriali nel modello. Resta comunque sorprendente una performance così scadente dato che le variabili continue erano di gran lunga quelle migliori per la classificazione, quindi le cause sono da ricercare anche all'interno dell'algoritmo stesso usato dal modello.

6. MODELLO CON TARGET NON BINARIO

Dopo l'analisi a target binario, si è condotto uno studio comprendendo anche i pazienti ipertiroidei. Il dataset utilizzato differisce rispetto alla classificazione binaria esclusivamente per il subset di esplicative individuato dalla feature selection: in questo caso il numero di variabili considerate risulta maggiore rispetto a quello della classificazione binaria. Oltre alle 7 variabili elencate precedentemente, sono state utilizzate anche:

- Trattamento antitiroide: binario, indica se il soggetto è sotto cura a base di farmaci antitiroidei
- Gozzo: binario, indica se il soggetto presenta gozzo colloidale
- Status: categoriale, con classi maschio, femmina, femmina in stato di gravidanza
- Tumore: binario, indica se il soggetto ha/ha avuto un tumore (non specifico ma relativo ad un qualsiasi organo corporeo)
- Malattia psichica: binario, indica se il soggetto soffre di una qualunque malattia psichica.

I modelli costruiti sono stati: J48, Multi Layer Perceptron e Random Forest; si è cercato di usare modelli con delle ottime performance per il target binario e differenti fra di loro.
Non potendo utilizzare le curve ROC per il confronto si sono utilizzate due statistiche: l'accuratezza e la recall.

Models	*Accuracy*
J48	0.973
Multi Layer Perceptron	0.952
Random Forest	0.975

Tabella 3: *Accuratezza nel modello a tre livelli*

Anche nel modello a tre livelli l'accuratezza raggiunge valori sorprendenti: nonostante il task classificativo sia più complesso, le 5 variabili supplementari considerate hanno permesso di raggiungere una performance soddisfacente.
Anche in questo caso è meglio ricorrere alla recall per la comparazione dei modelli.

Models' recall stratified per diagnosis			
	J48	*Multi Layer Perceptron*	*Random Forest*
Normal	0.981	0.984	0.987
Hypothyroidism	0.947	0.762	0.964
Hyperthyroidism	0.741	0.359	0.595

Tabella 4: *Recall stratificata per i tre livelli del target*

In questo caso viene calcolata una recall stratificata. I pazienti sani non sono di grande interesse in questo caso e tutte le recall relative hanno valori altissimi. Per gli altri due livelli del target troviamo la Multi Layer Perceptron con delle performance scadenti in confronto agli altri modelli, significa che questo modello fatica a classificare in modo adeguato una buona percentuale di ipo/ipertiroidei. Confrontando i due modelli rimanenti, il J48 risulta quello con performance migliori.
È interessante notare come la Random Forest che si era mostrata il modello migliore nella classificazione binaria sia con che senza costi non sia ancora una volta il modello vincente. Nonostante sia il modello migliore in termini di accuratezza e di recall per i pazienti sani ed ipotiroidei, negli ipertiroidei la recall non è stata all'altezza delle aspettative.

7. CONCLUSIONI

Le performance dei modelli implementati sono risultate decisamente soddisfacenti in termini di accuracy. Solo alcuni dei modelli analizzati, come random forest e Bayes net, presentano dei buoni risultati anche in termini di recall. Le variabili scelte in fase di feature selection sono risultate appropriate per l'analisi. Dallo studio emergono delle relazioni significative tra target ed attributi in input. Alcune confermano le ipotesi teoriche di base: ad esempio, alti valori di TSH, e bassi valori di FTI e T3 sono positivamente associati con l'ipotiroidismo. Un individuo ipotiroideo avrà quindi (in media) alti livelli di TSH nel sangue: ciò porta a stimolare la produzione di T3 e T4, presenti in quantità insufficiente. Al contrario, alti valori di FTI e T3 e bassi valori di TSH sono positivamente associati con una diagnosi di ipertiroidismo.
Altri risultati sono in contrasto con la realtà: solitamente le malattie tiroidee sono più frequenti in individui di sesso femminile, ma nella classificazione binaria la variabile relativa al genere (status) non è risultata significativa.

Il dataset scelto per l'analisi presentava alcune limitazioni:

- I livelli del target erano fortemente sbilanciati; nonostante ciò, viste le performances dei modelli implementati, non è stato necessario ricorrere ad un ricampionamento.
- I modelli classificavano eccessivamente bene, con delle accuracies prossime al 100%. In altri contesti e analisi le cifre di performance raggiunte sono solitamente più basse. La ragione di ciò risiede principalmente nelle variabili continue considerate: il TSH da solo era già in grado di compiere un egregio lavoro classificativo, un fatto in linea con l'eziologia della malattia.
- Dalle analisi del boxplot nella fase di preprocessing è emerso un range di valori di TSH molto più ampio rispetto alle altre variabili. Una possibile soluzione per far fronte a questo problema sta nel trasformare la variabile TSH applicando una funzione logaritmica. Ciò avrebbe portato a dei risultati ancora migliori, ma si sarebbe persa parte dell'interpretabilità legata al fenomeno.
- La conservazione delle categorie iniziali della variabile diagnosi (circa 16) avrebbe permesso di fare un'analisi più specifica e affine alla realtà; ma, come già detto, la classificazione di particolari varianti di ipo-ipertiroidismo avrebbe richiesto una competenza medica e metodologica avanzata per poter analizzare correttamente i risultati.

8. BIBLIOGRAFIA

1 Wikipedia, Tiroide.
2 Venturi, Sebastiano: Evolutionary Significance of Iodine, 2011.
3 UCI Machine Learning Repository, Thyroid Disease Dataset.
4 NIDDIK, Hashimoto's Disease, 2017.
5 Wikipedia, Thyroid Function Tests.