Classificazione delle piu comuni patologie tiroidee

Aldeghi C., Crippa C., Iacoban I., Viganò E.

Sommario

Oggigiorno le patologie legate alla tiroide sono sempre più diffuse e frequenti. Per questo motivo può essere interessante realizzare un modello di classificazione capace di individuare chi, all'interno di una certa popolazione, è affetto da ipotiroidismo, ipertiroidismo o risulta essere classificato come normale. Per raggiungere questo obiettivo, a seguito di accurate analisi descrittive ed esplorative sulle variabili riconducibili alla storia medica dei pazienti, sono state considerate diverse combinazioni di regressori. L'allenamento di differenti tipologie di modelli ha permesso l'identificazione di quelli dotati delle migliori capacità di classificazione in termini di *accuracy*. Ottime *performances* sono state ottenute non solo grazie all'implementazione della tecnica di analisi discriminante quadratica utilizzando, su di un dataset bilanciato e sbilanciato, gli ormoni *FTI*, *T4U* e *T3*, ma anche con un albero di classificazione (in presenza e in assenza di potatura) applicato, anche in questo caso sia al *dataset* bilanciato che a quello sbilanciato.

Introduzione

Il tiroidismo è una patologia che, a causa del cattivo funzionamento della tiroide, provoca una significativa alterazione delle quantità di determinati ormoni presenti nel sangue. La tiroide, in particolare, è una ghiandola endocrina in grado di regolare il corretto funzionamento del nostro metabolismo, lo sviluppo fisico e psichico dell'individuo e l'attività del cuore.
Il tiroidismo si manifesta generalmente in due forme:
- Ipotiroidismo
- Ipertiroidismo

Si parla di ipotiroidismo quando, a causa di una diminuzione dell'attività di secrezione della ghiandola tiroidea, si genera un quantitativo ristretto di ormoni tiroidei nell'organismo. I pazienti che manifestano questo tipo di patologia generalmente soffrono di affaticamento, diminuzione della memoria, rallentamento dell'attività fisica e psichica, aumento di peso, intolleranza al freddo e stipsi. Al contrario, la manifestazione dell'ipertiroidismo è dovuta all'aumento della capacità di secrezione della tiroide che causa la presenza di un eccesso di ormoni tiroidei nell'organismo. I principali sintomi che i pazienti manifestano nel caso di ipertiroidismo sono: disturbi del sonno, intolleranza al calore, nervosismo e irritabilità, affaticamento e debolezza muscolare, tremori alle mani e frequenza cardiaca accelerata. A livello medico, per comprendere se un soggetto è affetto da una di queste patologie, si eseguono appropriati test clinici che valutano il livello degli ormoni presenti nel sangue. Tra le diverse variabili investigate da sempre figura l'ormone *TSH* il quale è considerato essere il marcatore per eccellenza nell'identificare la presenza o l'assenza della malattia. Questo aspetto non preclude però l'importanza degli altri ormoni. È infatti noto che i valori assunti da *T3* e *T4* sono determinanti per il *TSH* ed il complessivo funzionamento della tiroide.

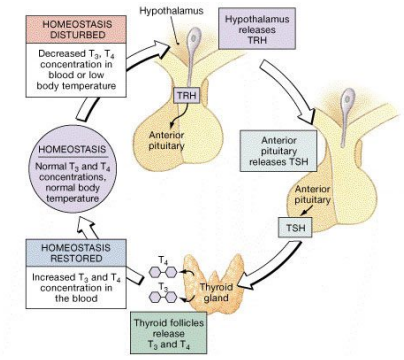

Figura 1: Ciclo TSH

Descrizione del dataset

Il *dataset* è composto da 7200 osservazioni e 22 variabili così definite:

Variabile	Tipologia	Supporto	Descrizione
Class	Categoriale	1,2,3	"Hypothyroidism" (1), "Hyperthyroidism" (2), "Normal" (3)
Age	Continua	[0.01,0.97]	Eta' dei pazienti oggetto dello studio
Sex	Discreta	[0, 1]	Genere dei pazienti oggetto dello studio: donna (0), uomo (1)
On thyroxine	Discreta	[0, 1]	Assunzione di Tiroxina: no (0), si (1)
Query on thyroxine	Discreta	[0, 1]	Il paziente conosce il trattamento con la tiroxina : no (0), si (1)
On antithyroid medication	Discreta	[0, 1]	Assunzione di ormoni che contrastano gli ormoni tiroidei: no (0), si (1)
Sick	Discreta	[0, 1]	Paziente: non malato (0), malato (1)
Pregnant	Discreta	[0, 1]	Paziente donna: non in stato di gravidanza (0), in stato di gravidanza (1)
Thyroid surgery	Discreta	[0, 1]	Intervento chirurgico alla tiroide: non subito (0), subito (1)
I131 treatment	Discreta	[0, 1]	Trattamento con lo iodio: non sostenuto (0), sostenuto (1)
Query hypothyroid	Discreta	[0, 1]	Il paziente conosce la patologia ipotiroidismo? : no (0), si (1)
Query hyperthyroid	Discreta	[0, 1]	Il paziente conosce la patologia ipertiroidismo? : no (0), si (1)
Lithium	Discreta	[0, 1]	Trattamento con il litio: non sostenuto (0), sostenuto (1)
Goitre	Discreta	[0, 1]	Tiroide non ingrossata (0), tiroide ingrossata (1)
Tumor	Discreta	[0, 1]	Assenza di tumore (0), presenza di tumore (1)
Hypopituitary	Discreta	[0, 1]	Malfunzionamento dell'ipofisi che non produce alcuni ormoni fondamentali
Psych	Discreta	[0, 1]	Disturbo psichico: assenza (0), presenza (1)
TSH	Continua	[0.0, 0.53]	Livello dell'ormone tireotropo nel sangue
T3	Continua	[0.0005, 0.18]	Livello dell'ormone Triiodotironina nel sangue
TT4	Continua	[0.0020, 0.6]	Livello dell'ormone Tiroxina Totale nel sangue
T4U	Continua	[0.017, 0.233]	Tasso di utilizzo della Tiroxina
FTI	Continua	[0.0020, 0.642]	Indice di Tiroxina libera

Analisi descrittive

Il *dataset* analizzato è privo di valori mancanti, presenta 6 variabili quantitative e 15 categoriali.
L'osservazione delle rappresentazioni grafiche ha permesso di evidenziare un'incongruenza nel *dataset*. Infatti, studiando il comportamento delle variabili relative al livello degli ormoni nel sangue è emerso che l'andamento da esse espresso non corrispondeva a quanto riscontrato in letteratura. Per questo motivo si è modificata la nomenclatura della variabile *class* nel seguente modo:
- 1 = ipotiroidismo
- 2 = ipertiroidismo
- 3 = normale.

La concentrazione del numero di individui per ciascuna classe è così strutturata:

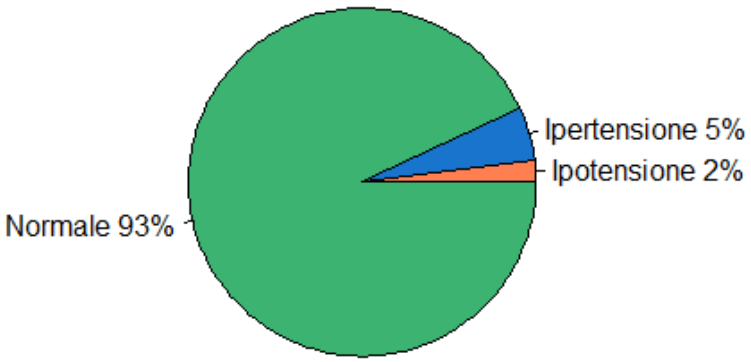

Figura 2: Pie chart della variabile risposta class.

L'analisi univariata delle variabili quantitative mostra:

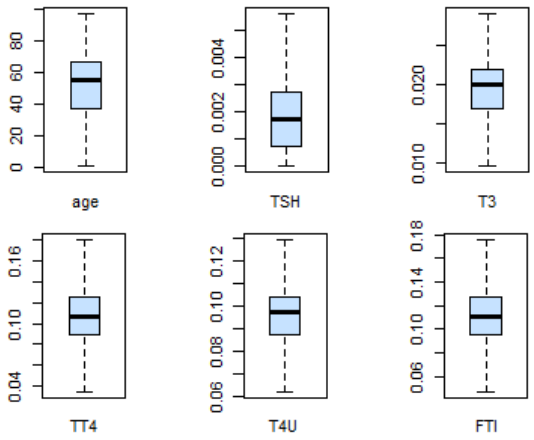

Figura 3: Boxplot univariati.

Dal grafico si evince che le variabili *FTI* e *TT4* risultano essere sufficientemente simmetriche. Le altre variabili (*age*, *TSH*, *T3* e *T4U*), invece, presentano una asimmetria.

E' stata condotta la medesima analisi anche sulle variabili qualitative:

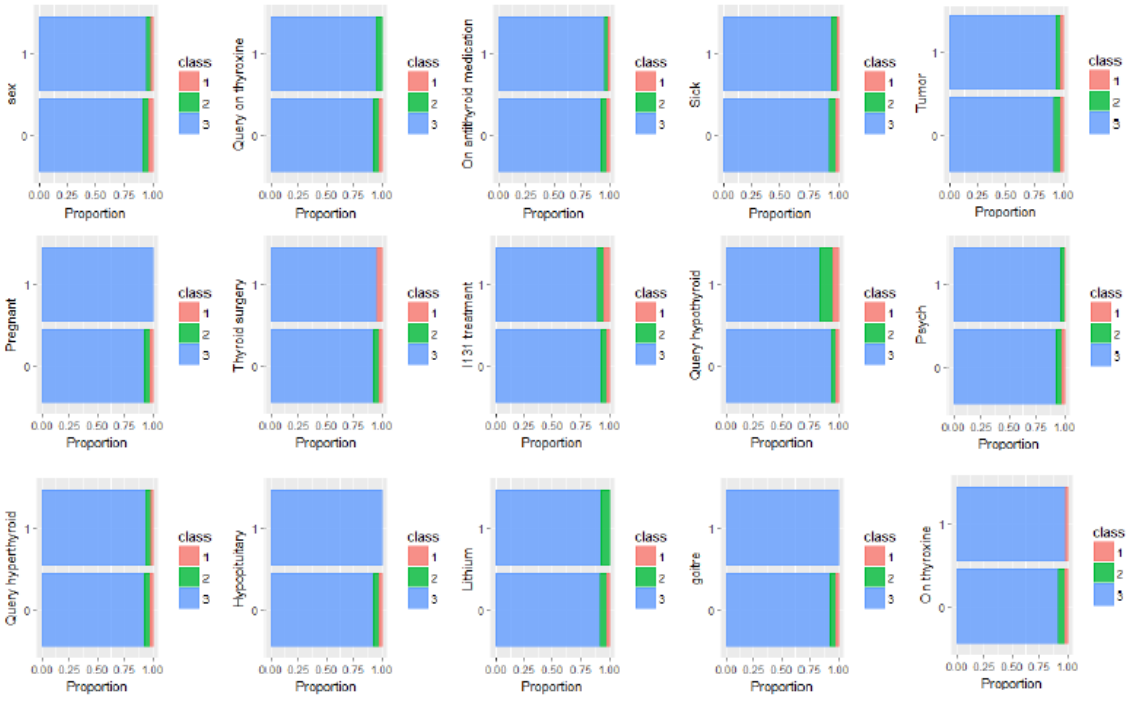

Figura 4: Barplot condizionati alla variabile target.

Risultati più informativi sono stati ottenuti dall'analisi, effettuata mediante l'uso di *boxplot* (senza *outliers*), delle variabili condizionate al *target* al fine di facilitarne la visualizzazione grafica.

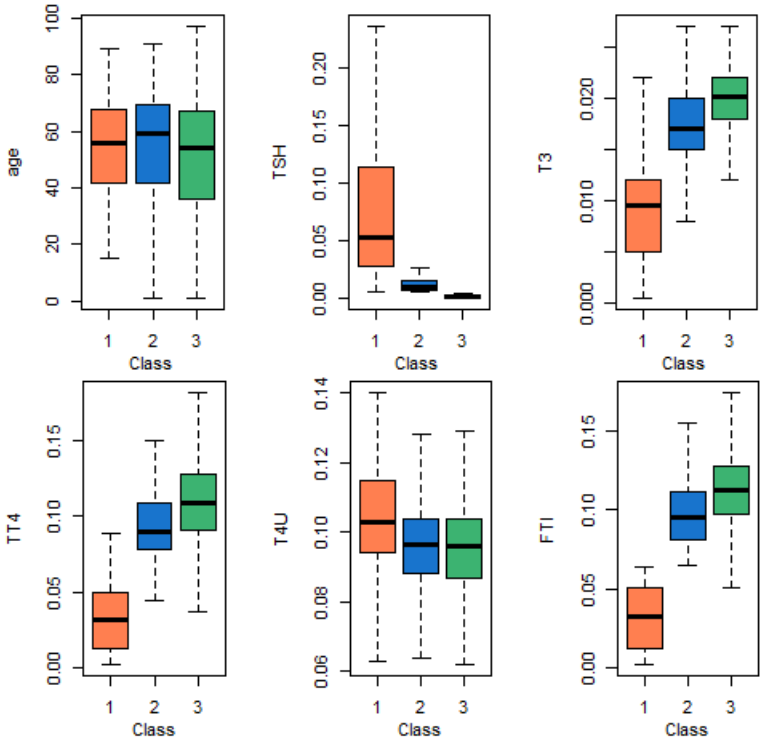

Figura 5: Boxplot condizionati alla variabile target.

Osservando il grafico è emerso che le variabili quantitative più interessanti sono gli ormoni *T3*, *TT4* e *TSH*, perché capaci di discriminare in modo più marcato delle altre i livelli assunti dalla variabile dipendente.

La variabile *TSH* discrimina bene la classe 1 (ipotiroidismo) dalla classe 2 (ipertiroidismo), permettendo quindi di distinguere la patologia cronica, specifica dei soggetti malati. Infatti, valori elevati di questo ormone indicano che la tiroide è poco attiva (ipotiroidismo), al contrario valori bassi possono definire un soggetto che soffre di ipertiroidismo.

Risulta evidente anche che l'ormone *T3* riesce a discriminare rispetto alla variabile *target*. Esso è responsabile di diverse disfunzioni dell'organismo, infatti può inficiare la pressione sanguigna, la temperatura corporea e il regolare battito cardiaco. Chiaro segnale di un paziente affetto di ipotiroidismo è un valore basso di questo ormone. Risulta più difficile, invece, utilizzarlo per comprendere se un individuo soffre di ipertiroidismo, poichè i suoi valori difficilmente si discostano da quelli che si registrano nella classe 3 (normale).

La tiroide agisce in ugual modo sullo stimolo sanguigno della produzione degli ormoni *T3* e *T4*, quindi le stesse considerazioni possono essere estese anche per il *Total Thyroxine*, vale a dire il *TT4*. Un alto livello di *TT4* è indice di una tiroide iperattiva e dunque di un quadro clinico caratterizzato da ipertiroidismo. Al contrario, se il valore assunto da questo ormone è basso si tratta probabilmente di ipotiroidismo, a causa di un'insufficienza tiroidea. Come per il *T3*, si riscontrano maggiori difficoltà nella distinzione tra coloro che soffrono di ipertiroidismo e gli individui della classe 3. Degna d'attenzione risulta essere anche la variabile *FTI*, perchè capace anch'essa di discriminare in maniera piuttosto soddisfacente i livelli del *target*. Le considerazioni sono le medesime di quelle riportate per i precedenti ormoni.

In previsione della costruzione di modelli statistici, si è indagato il possibile legame tra le variabili esplicative mediante opportuni grafici. Si propone il grafico delle correlazioni di *Pearson* per mostrare il tipo e l'intensità delle relazioni che sussistono tra le variabili quantitative. Si individua in questo modo la tendenza che hanno due variabili a variare insieme, cioè a covariare.

Tale coefficiente è standardizzato e può assumere valori nel *range* [-1,1]:
- Correlazione = +1 : indica perfetta correlazione positiva tra le variabili;
- Correlazione = 0 : non persiste alcuna relazione tra le variabili considerate;
- Correlazione = -1 : indica perfetta correlazione negativa tra le variabili.

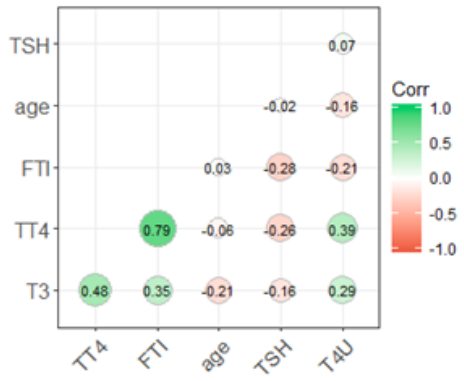

Figura 6: Corrplot delle variabili quantitative.

Tra i valori ottenuti risalta l'elevata correlazione positiva tra le esplicative *TT4* e *FTI* e, se pur con intensità minore, la correlazione tra *TT4* e *T3*. Le altre variabili presentano correlazione quasi nulla in ogni incrocio combinato. Le stesse considerazioni emergono dal grafico della matrice di diagrammi di dispersione, in cui si è introdotta l'informazione della variabile *class*.

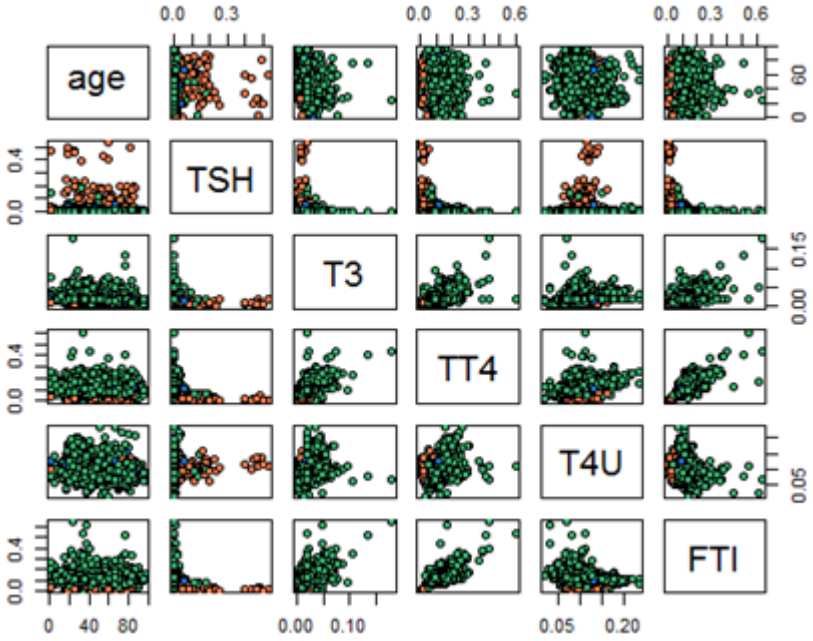

Figura 7: Diagramma di dispersione tra le variabili quantitative.

L'elevata correlazione positiva per la coppia *TT4* e *FTI* come per *TT4* e *T3*, mostra una possibile collinearità tra le variabili che, in vista di un modello, apporterebbero un analogo carico informativo. Nonostante si sia sottolineato la capacità discriminante dell'ormone *TT4* nel distinguere i diversi livelli del *target*, si è considerato più opportuno proseguire le analisi utilizzando l' esplicativa *T4U*, essendo componente della stessa e debolmente correlata con *FTI* e *T3*.

Per valutare il livello di associazione tra variabili qualitative nominali si è deciso di utilizzare l'indice *V di Cramer* definito in un *range* tra 0 e 1:
- V di Cramer = 0 : indica assenza di associazione tra i caratteri studiati;
- V di Cramer = 1 : indica perfetta connessione tra le variabili.

Si riporta la rappresentazione grafica dei risultati ottenuti, di più immediata interpretazione.

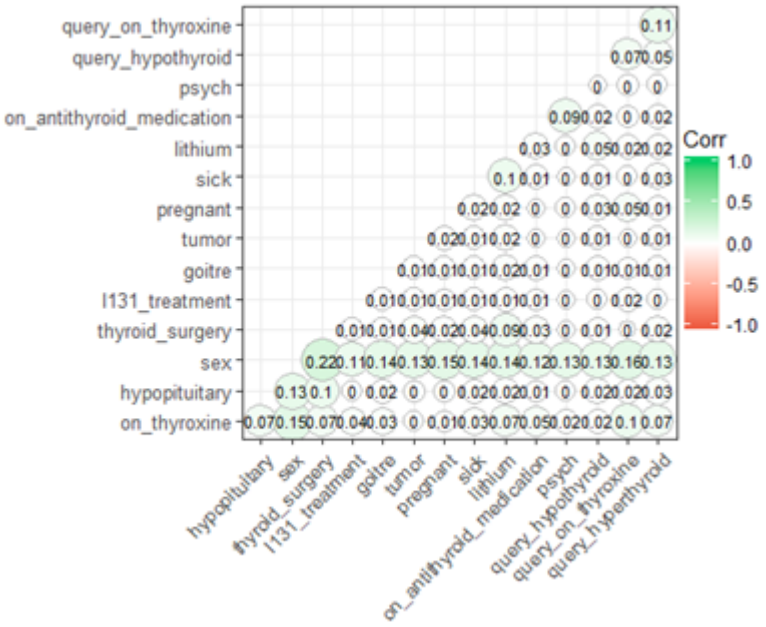

Figura 8: Grafico di associazione tra le variabili qualitative.

Dal grafico risulta evidente l'assenza di connessioni significative tra le variabili categoriali del *set* di dati a nostra disposizione. Si può dedurre che nessuna delle variabili qualitative risulta essere significativamente legata al *target*.

Modelli di classificazione

I risultati conseguiti in sede di analisi descrittiva ed esplorativa hanno consentito di raggiungere una migliore comprensione del fenomeno oggetto di analisi. Si sfrutta questa maggiore conoscenza per applicare delle tecniche supervisionate ai dati, ovvero delle tecniche che permettono di analizzare la relazione esistente tra la variabile risposta Y e le esplicative $X_1, ..., X_i, ..., X_p$ con *i= 1, . ,22*.

Lo scopo perseguito è, quindi, quello di andare a stimare una funzione *f()* sulla base delle osservazioni contenute nel *dataset*. Essa è utilizzata nella costruzione di un classificatore al fine di classificare al meglio nuove osservazioni di cui non si conosce la classe di appartenenza, ma è noto il valore delle esplicative.

Nonostante, la natura fortemente sbilanciata del *dataset*, in un primo momento si è valutata la bontà dei classificatori senza introdurre alcun tipo di manipolazione dei dati. I modelli applicati sono stati due:
- Analisi discriminante quadratica;
- Alberi di classificazione.

Le misure utilizzate per valutare la bontà della *performance* dei modelli implementati sono, per la maggior parte, ottenute a partire dalla matrice di confusione:

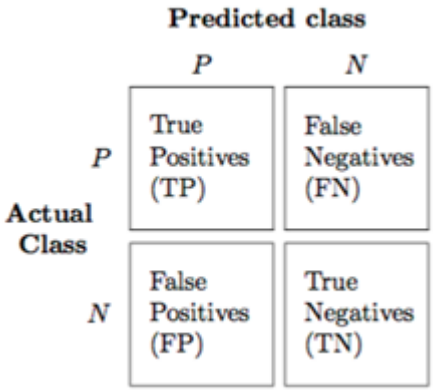

Figura 9: Matrice di confusione.

-Errore di previsione $= \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$
-$Accuracy = 1 -$ Errore di previsione $= \frac{TP+TN}{TP+TN+FP+FN}$
-$Sensitivity = \frac{TP}{TP+TN}$
-$Precision = \frac{TP}{TP+FP}$

dove :
-$I(\ )$ è la funzione indicatrice la quale assume valore pari a 1 se la condizione contenuta nelle parentesi è soddisfatta.
-y_i è la vera classe di appartenenza dell'*i-esima* unità.
-$\hat{y}_l$ è la previsione della classe di appartenenza dell'*i-esima* unità.

Analisi discriminante

L'analisi discriminante è una tecnica di classificazione che ha l'obiettivo di definire una modalità di assegnazione delle unità statistiche alla corrispondente classe della variabile risposta, sulla base del valore assunto dalle variabili esplicative osservate.

Esistono due tipologie di analisi discriminante:
- Lineare
- Quadratica

Entrambi i modelli si basano sull'ipotesi fondamentale per cui la distribuzione degli attributi condizionati alla classe della variabile risposta, deve essere di tipo normale:

$$X|Y = J \sim N(\mu_l, \sigma^2)$$

Dall'osservazione di questa relazione è già possibile individuare una delle principali differenze che permette la distinzione tra questi due classificatori. Infatti, se da un lato in entrambi i casi la media delle distribuzioni condizionate varia al variare della classe considerata, dall'altro lato non sempre la varianza permane costante tra le classi. Nel caso dell'analisi discriminante lineare, infatti, viene fatta l'assunzione che la popolazione sia omoschedastica. Ciò implica che σ^2 rimane sempre costante indipendentemente da quale classe della variabile risposta Y viene utilizzata come elemento di condizionamento delle esplicative.

Nell'analisi discriminante quadratica, invece, è ammessa l'ipotesi per cui σ_l^2 può variare tra le classi. La popolazione in questo caso viene definita eteroschedastica. L'assunzione di base circa la distribuzione delle esplicative condizionate, assume quindi la forma: $X|Y = J \sim N(\mu_l, \sigma_l^2)$.

Al fine di costruire un modello adeguato è stato necessario innanzitutto verificare la validità dell'ipotesi di normalità tramite l'osservazione degli istogrammi delle variabili continue, condizionati al valore assunto dalla variabile dipendente.

Dall'osservazione dei grafici risulta che le variabili *age*, *T3*, *TT4*, *FTI*, *T4U* risultano soddisfare l'ipotesi di normalità. La variabile *TSH*, invece, è difficilmente riconducibile a tale distribuzione. Data la forte capacità discriminante di questo attributo, in un primo momento si è ritenuto inopportuno eliminare questa variabile dalle analisi. Si è reso perciò necessario applicare il metodo della trasformazione di *Box-Cox*, al fine di rendere le distribuzioni marginali normali. La trasformazione di *Box-Cox*, per ipotesi, può essere applicata solo se i valori osservati delle esplicative risultano essere maggiori di 0. *TSH* assume però valori nell'intervallo [0, 0.53], quindi per garantire la possibilità di applicare tale trasformazione, si è aggiunto ad ogni valore osservato una costante (+1), assicurandone l'invarianza. In questo modo *TSH* assume valori nel range [1, 1.53].

La trasformazione *Box-Cox* prevede:

$$X(\lambda) = \frac{x^{\lambda} - 1}{\lambda} \quad \text{se} \quad \lambda \neq 0$$

$$X(\lambda) = log(x) \quad \text{se} \quad \lambda = 0$$

con $\lambda \in R$.

Tuttavia tale trasformazione non ha prodotto risultati soddisfacenti in termini di normalità, per questo motivo si è escluso il marcatore *TSH* dall'analisi discriminante in favore degli ormoni *T3*, *T4U* e *FTI*.

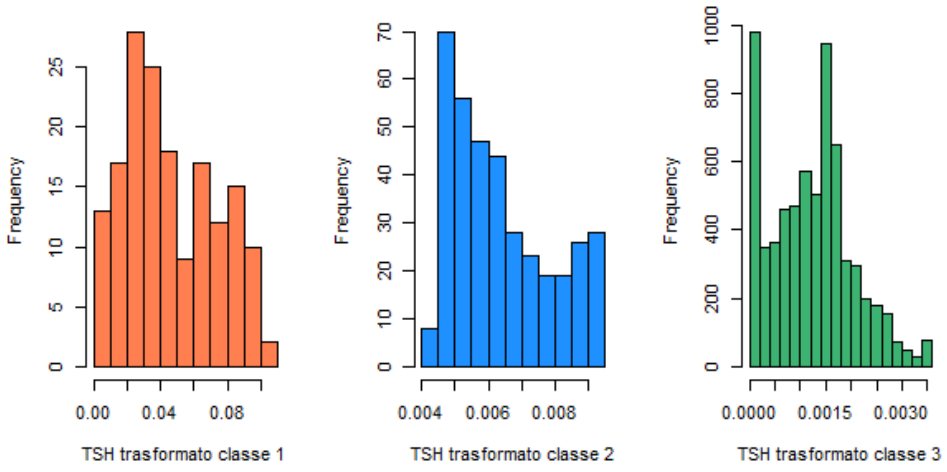

Figura 10: Istogramma della variabile TSH trasformata per classi.

Successivamente è stato implementato il *test* di *Bartlett* per verificare l'omoschedasticità della popolazione. Tale *test* si basa sulle seguenti ipotesi:

$$H_0 : \Sigma_k = \Sigma$$

$$H_1 : \Sigma_k \neq \Sigma$$

La statistica *test* sotto H_0 si distribuisce asintoticamente come una χ^2_{m-1} ovvero come una *Chi-Quadrato* con *m-1* gradi di libertà. Considerato un livello di significatività α=0.01, il valore del *p-value* ottenuto è: *p-value* $<$ *2.2e-16*. Ciò induce a rigettare H_0 ovvero l'ipotesi di omoschedasticità e ad accettare l'ipotesi alternativa.

Il modello scelto, quindi, è l'analisi discriminante quadratica.

Il suddetto classificatore si basa su una regola di classificazione legata al teorema di *Bayes*, ovvero essa assegna ciascuna osservazione alla classe j se:

$$\max_{l \in y} P(Y = l|X = x_0) = P(Y = j|X = x_0) = \frac{\pi_l f_l(x_0)}{\sum_{u=1}^{K} \pi_u f_u(x_0)}$$

Dove π_l è la probabilità a priori, ovvero la probabilità che un'unità statistica ha di appartenere alla generica classe l della popolazione. $f_l(x_0)$ invece, è la probabilità di osservare x_0 nella classe l.

Implementazione

L'analisi discriminante quadratica è stata condotta tramite la metodologia *Leave-one-out cross-validation*. Attraverso l'uso di questa tecnica si vogliono superare le problematiche generate dall'arbitrarietà con cui avviene la scelta della numerosità dell'insieme di stima e di verifica, garantendo una maggiore robustezza dei risultati ottenuti. Il metodo consiste nella suddivisione del dataset in $K = n$ parti, dove n è il numero totale delle osservazioni rilevate ($n = 7200$). Ad ogni passo l'algoritmo considera una delle n osservazioni come *test*, mentre le restanti *n-1* osservazioni vengono trattate come insieme di stima. Successivamente viene implementato il classificatore, si calcola una sua stima e si valuta il valore dell'errore di previsione. Questo processo viene ripetuto n volte, determinando una stima finale dell'errore di previsione come media di tutti gli errori di previsione ottenuti ad ogni passo dell'algoritmo.

Oltre a questo, si vuole anche puntualizzare che il modello è stato allenato unicamente attraverso l'uso di tre variabili selezionate: *T3*, *T4U* e *FTI*.

Analisi risultati QDA

Sebbene il *dataset* sia sbilanciato e contenga poche osservazioni relative alle unità statistiche appartenenti alle classi di interesse, l'*accuracy* ottenuta dall'implementazione del modello QDA è pari a 0,935. Nonostante l'elevato valore assunto, tale risultato non risulta ottimo perchè il modello non classifica adeguatamente le tre classi. In particolare, nessuna unità statistica della classe 2 viene prevista correttamente e molte unità della classe 1 vengono confuse.

Alberi di classificazione

Gli alberi di classificazione vengono costruiti a partire dalla scelta delle variabili esplicative e dei punti di *split* che permettono di classificare le unità in gruppi il più possibile omogenei al loro interno e differenziati tra di loro. Il criterio di *split* che viene utilizzato deve consentire l'individuazione della miglior partizione binaria fra tutte quelle possibili. In generale, data una regione arbitraria R, la divisione in due nuove regioni $R1 = \{X|X_j < c\}$ $R2 = \{X|X_j \geq c\}$ rispetto alla variabile X_j (j=1,2,.,21) e al punto di divisione c, viene valutata massimizzando il guadagno di informazione:

$$i(R) - p_1 i(R_1(j, c)) - p_2 i(R_2(j, c))$$

dove $i(R_m)$ è l'impurità della regione R_m, mentre p_m è la proporzione di osservazioni nelle rispettive regioni m. In particolare, nella suddetta analisi, si è utilizzata come misura di impurità l'indice di Gini, ovvero:

$$i(R_m) = \sum_{l \in Y} \hat{p}_{ml}(1 - \hat{p}_{ml}) \quad \text{e} \quad \hat{p}_{ml} = \frac{1}{card(R_m)} \sum_{t: x_t \in R_m} I(y_t = l)$$

La ripartizione binaria si arresta dunque quando i nodi terminali hanno un livello di purezza elevato e ulteriori divisioni non apporterebbero nessun miglioramento in termini di accuratezza.

Dopo aver fatto crescere l'albero si procede con la potatura, ovvero una tecnica che riduce la dimensione dell'albero rimuovendo i rami che hanno poco potere discriminante. Questa metodologia si serve di un parametro reale α che controlla la dimensione dell'albero e l'adattamento di quest'ultimo ai dati. Per scegliere infine il miglior sotto albero, tra tutti quelli possibili, si procede con una *K-fold cross validation*. Questa procedura ha il vantaggio di ridurre la complessità del classificatore finale e di prevenire il problema dell'*overfitting*. Nell'implementazione di tale modello si è provveduto innanzitutto alla divisione del *dataset* in: *training-set* (70%) e *test-set* (30%), in modo tale da verificare la *performance* del modello su dati "nuovi". La suddivisione tra *training-set* e *test-set* è stata effettuata tramite estrazione casuale e settando un seme attraverso l'utilizzo della funzione *set.seed* (123), in modo tale da garantire la riproducibilità dei risultati. Si è deciso inoltre di considerare tutte le variabili presenti nel *dataset*, a differenza di quanto fatto nel modello precedente (QDA), in quanto l'albero è in grado di selezionare autonomamente le variabili ritenute più importanti per la costruzione del modello. In questo modo vengono considerate anche le variabili dicotome.

Risultati

L'*accuracy* ottenuta dalla crescita massima dell'albero è pari a 0.995. In particolar modo si osserva dalla matrice di confusione che sia la classe 1 che la classe 2 vengono classificate molto bene, infatti solo 5 osservazioni della classe 2 vengono classificate erroneamente. Questo suggerisce che l'albero riesce a discriminare bene tra le classi anche se queste sono molto sbilanciate. L'*accuracy* ottenuta con il metodo della potatura, invece, è pari a 0.989. L'albero iniziale presentava una dimensione pari a 12 e in seguito a una potatura è stato ridotto a una dimensione pari a 4. Ciò è dovuto ai termini di penalizzazione introdotti nel modello: *mincut*=20, *minsize*=40. Anche questa volta si nota che la classe 1 e la classe 2 vengono classificate perfettamente e solo 23 osservazioni della classe 3 vengono classificate erroneamente. Le variabili utilizzate per la costruzione del modello sono: *TSH*, *FTI* e *on_thyroxine*.

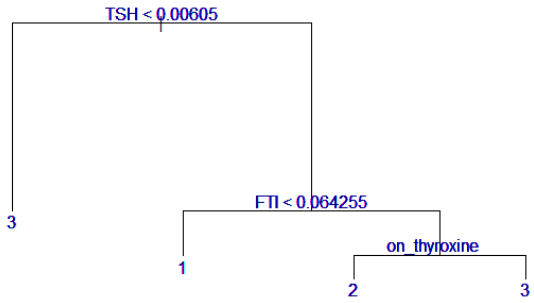

Figura 11: Pruned tree.

Undersampling e oversamplig

Nonostante l'evidente bontà dei risultati raggiunti, si è deciso di trattare il problema dello sbilanciamento del *dataset* e quindi della possibile distorsione nel processo di apprendimento del classificatore causata dal rischio che il modello tenda a focalizzarsi sulla classe predominante (*class* 3), ignorando le classi rare (*class* 1 e 2).

Per cercare di risolvere questo problema possono essere utilizzate due diverse metodologie:
- *Undersampling*: tecnica per cui si effettua un campionamento senza ripetizione tra le osservazioni appartenenti alla classe maggioritaria.

- *Oversampling*: metodo che mira a bilanciare la distribuzione di classe attraverso la replicazione casuale di unità statistiche appartenenti alla classe minoritaria.

L'applicazione di entrambi i metodi richiede, tuttavia, particolare attenzione, in quanto se da un lato applicando l'*oversampling* si potrebbe incorrere nel problema dell'*overfitting*, che consiste in una forte specializzazione del modello sui dati, dall'altro lato l'uso della metodologia dell'*undersampling* potrebbe portare all'esclusione dall'insieme di stima di alcuni dati potenzialmente utili all'analisi.

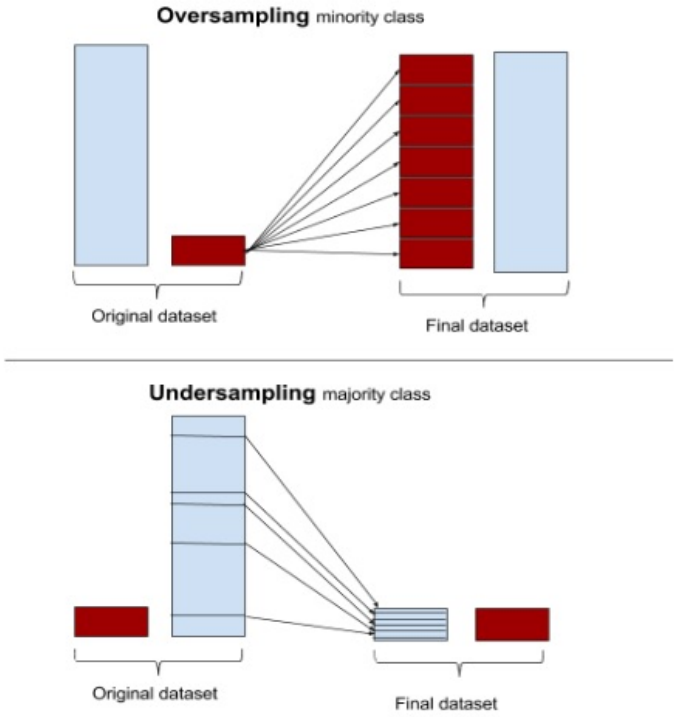

Figura 12: Bilanciamento con under e oversampling.

Entrambi i metodi alterano la dimensione dei dati originali. Al fine di valutare quale sia la metodologia più performante tra le due, si è deciso di implementare sia *undersampling* che *oversampling* sui dati a disposizione e di analizzare i risultati delle previsioni ottenute attraverso la stima dei modelli di QDA e albero sui *dataset* bilanciati.

L'algoritmo utilizzato per implementare sia l'*undersampling* che l'*oversampling* seleziona, ad ogni passo, solo una delle n=7200 osservazioni facenti parte di ognuno dei due *dataset* iniziali e le inserisce in due *test set* distinti che al termine del processo verranno utilizzati per verificare la bontà delle previsioni ottenute con entrambi i modelli. Contemporaneamente le restanti *n-1* osservazioni di ognuno dei due *dataset* originari vengono inserite in due nuovi *dataset*. A partire da questi ultimi viene individuata la proporzione di osservazioni appartenenti a ciascuna delle tre modalità della variabile risposta. A questo punto a seconda che venga implementato l'*undersampling* o l'*oversampling* si sceglie la dimensione dei nuovi *training set* su cui viene valutata la stima dei modelli. Nel caso dell'*undersampling* si è creato un nuovo *datatset* di 1650 osservazioni in cui si mantengono tutte le osservazioni delle classi minoritarie (1,2) e si estrae in modo casuale una parte limitata delle osservazioni (1155) facenti parte della classe 3, originariamente costituita da 6666 individui.

Le proporzioni utilizzate per la realizzazione dei due differenti insiemi di stima sono state:
- 10% per la classe 1;

- 20% per la classe 2;
- 70% per la classe 3.

Nel caso dell'*oversamplig*, invece, si è deciso di creare due *training set* di 3000 osservazioni. A tale scopo si è ricampionato più volte dagli insiemi contenenti le istanze legate alle classi di dimensionalità ridotta (1 e 2) per garantire una maggiore presenza delle unità statistiche manifestanti le patologie di ipotiroidismo e ipertiroidismo. Si è, inoltre, provveduto al ridimensionamento del numero di individui appartenenti alla classe 3.

Ciò è stato fatto tenendo conto delle seguenti proporzioni:
- 20% per la classe 1;
- 30% per la classe 2;
- 50% per la classe 3.

Su entrambi i *training set* bilanciati prima con l'*undersampling* e poi con l'*oversampling* sono stati poi implementati i modelli sopra menzionati (albero su tutte le variabili e QDA sulle variabili selezionate: $T3$, $T4U$ e FTI). Questo procedimento è stato ripetuto n volte.

Si riportano i risultati ottenuti:

Tecnica	Modello	Accuracy
Senza Bilanciamento	QDA	0.9355556
	Tree	0.9949074
	Tree pruned	0.9893519
Undersampling	QDA	0.9108333
	Tree	0.9927778
Oversampling	QDA	0.6315278
	Tree	0.9916667

Tabella 1: Tabella risultati.

Per entrambe le tecniche applicate, l'*accuracy* dei modelli considerati raggiunge livelli molto elevati, questo perché la maggior parte delle osservazioni viene correttamente classificata. Nonostante gli alti livelli di *accuracy* raggiunti, si predilige l'*undersampling*, come indicato in letteratura.

Confronto a coppie

Reputando interessante il confronto con la classe 3 della variabile *target*, si sono realizzati opportuni confronti a coppie. L'analisi compiuta ripercorre la metodologia precedentemente adottata, con la possibilità di cogliere meglio la clasificazione.

Tecnica	Modello	Accuracy	Recall	Precision	AUC
Senza Bilanciamento	QDA13	0.9859485	0.4879518	0.8804348	0.7431508
	Tree13	0.9970732	0.9111111	0.9534884	0.955
	Tree pruned13	0.9970732	0.9555556	0.9148936	0.977
Undersampling	QDA13	0.9732143	0.9578313	0.4746269	0.966
	Tree13	0.9831674	1	0.5907473	0.991

Tabella 2: Tabella risultati per dataset con variabile risposta contenente solo la prima e la seconda classe.

Tecnica	Modello	Accuracy	Recall	Precision	AUC
Senza Bilanciamento	QDA23	0.9476827	0	0/0	0.5
	Tree23	0.9976315	0.974359	0.9827586	0.986678
	Tree pruned23	0.9976315	0.974359	0.9827586	0.987
Undersampling	QDA23	0.4792437	0.826087	0.0778888	0.643
	Tree23	0.9862098	1	0.791397	0.993

Tabella 3: Tabella risultati per dataset con variabile risposta contenente solo la seconda e la terza classe.

Dai risultati emerge che la *performance* classificativa del modello QDA è sicuramente da ritenersi non tanto buona quanto quella dei diversi modelli di albero implementati.

Conclusioni

I risultati prodotti dai diversi alberi sono particolarmente performanti e confermati anche in caso di bilanciamento con tecniche di *undersampling* e *oversampling*. Le motivazioni che giustificano tali conclusioni possono essere la presenza del marcatore *TSH* a guida dell'analisi oltre al fatto che il modello considera importanti non solo variabili relative agli ormoni, ma anche legate alla storia clinica del paziente stesso. L'analisi qui condotta può senza dubbio essere ampliata. In studi futuri potrebbe essere interessante analizzare le implicazioni derivanti da un più accurato trattamento degli *outliers* e approfondire il confronto tra le due malattie.

Riferimenti bibliografici

[1] Agnieszka Wosiak e Sylwia Karbowiak. *Preprocessing compensation techniques for improved classification of imbalanced medical datasets*. Settembre 2017.

[2] *https://www.everydayhealth.com/hs/healthy-living-with-hypothyroidism/understanding-test-results/*

[3] https://www.dietvsdisease.org/normal-tsh-levels/

[4] https://www.my-personaltrainer.it/salute/gozzo-tiroideo1.html

[5] https://pbrainmd.wordpress.com/2016/06/14/euthyroid-sick-syndrome/

[6] Figura 1: https://www.dietvsdisease.org/normal-tsh-levels/

[7] Figura 9: https://rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix/

[8] Figura 12: http://www.svds.com/wp-content/uploads/2016/08/ImbalancedClasses_fig5.jpg