Insegnamento di Analisi dei dati (Data mining)
Prova d'esame dell' 11 luglio 2017 - parte pratica

Daniele Cugnigni

2023-02-20

Testo d'esame

Alcuni ricercatori sono interessati alla classificazione di una voce ottenuta da un file audio, come maschile o femminile a seconda di alcune proprietà acustiche. A tale scopo, nel file voce.csv, è disponibile un campione di registrazioni della durata di 2-5 minuti, tratti da 62 file audio provenienti da diversi capitoli di audiolibri. Di questi capitoli, 32 sono letti da lettori di sesso maschile e i restanti 30 da lettrici. Per ciascun file è stato creato uno spettrogramma (rappresentazione grafica dell'intensità di un suono in funzione del tempo e della frequenza) dal quale sono stati estratti i campioni che effettivamente rappresentano un suono, tralasciando il rumore bianco. Si sono quindi isolati 19090 spezzoni dei quali 10313 provenienti da lettori di sesso maschile e 8777 da lettrici. Per ciascun spezzone sono disponibili le seguenti variabili:

- meanfreq: frequenza media (in kHz)
- sd: deviazione standard della frequenza
- median: mediana della frequenza (in kHz)
- Q25: primo quartile (in kHz)
- Q75: terzo quartile (in kHz)
- IQR: scarto interquartile (in kHz)
- skew: misura di asimmetria della distribuzione
- kurt: misura di curtosi della distribuzione
- sp.ent: entropia spettrale
- sfm: piattezza spettrale
- mode: moda della frequenza (in kHz)
- centroid: centroide della frequenza
- peakf: picco di frequenza
- meanfun: media della frequenza fondamentale
- minfun: frequenza fondamentale minima
- maxfun: frequenza fondamentale massima
- meandom: media della frequenza dominante

- mindom: frequenza dominante minima
- maxdom: frequenza dominante massima
- dfrange: range della frequenza dominante
- modindx: indice di modulazione

E' inoltre disponibile la variabile qualitativa "genere" che identifica il gruppo di appartenenza.

Pulizia del dataset

Il file "voce.csv" è composto da 19090 unità statistiche (gli spezzoni dei 62 file audio) sulle quali sono state rilevate complessivamente 23 variabili, con la variabile *genere* che rappresenta la variabile risposta.
Prima di procedere all'analisi del dataset, è opportuno effettuare delle operazioni di pulizia. In primo luogo si nota come la variabile X non è nient'altro che l'indicatore di riga, pertanto viene eliminata. Inoltre si nota come la variabile *sound.files* faccia riferimento a quale dei 62 file audio (o capitoli) appartiene il singolo spezzone, ovvero è una variabile indicatrice dell'unità statistica originale (il singolo file audio), pertanto viene eliminata.
A questo punto, tenendo conto anche del fatto di avere a disposizione un numero relativamente piccolo di variabili esplicative, è opportuno effettuare delle considerazioni meramente statistiche. Prima di tutto si nota che sono state rilevate le variabili *Q25*, *Q75* e *IQR*, e la relazione che lega queste tre variabili è lineare ed è data da *IQR = Q75 - Q25*, pertanto l'informazione contenuta nello scarto interquartile può essere ricavata dalla differenza tra terzo quartile e primo quartile. Alla luce di questa considerazione, viene deciso di eliminare la variabile *IQR*. Con un ragionamento analogo, emerge che sono state rilevate le variabili riguardanti la frequenza dominante massima, *maxdom*, la frequenza dominante minima, *mindom*, e il range della frequenza dominante, *dfrange*, e viene deciso di eliminare quest'ultima. Inoltre vi è la presenza sia della variabile *meanfreq* che della variabile *centroid*: la correlazione tra queste due variabili è pari ad 1, in particolare i valori della variabile *meanfreq* e della variabile *centroid* sono identici per ogni unità statistica, pertanto l'informazione portata dalle due variabili è la medesima e si decide di eliminare la variabile *centroid*.

Inoltre, nel dataset non sono presenti valori mancanti.

In seguito a queste operazioni, il dataset è composto da 19090 unità statistiche e 18 variabili. A questo punto, prima di procedere con la modellazione dei dati:

- si verifica l'assenza di variabili esplicative degeneri (e quindi inutili per l'analisi);
- per tenere in considerazione il compromesso tra varianza e distorsione, si procede con la divisione del dataset in insieme di stima (80%) e insieme di verifica (20%), ottenendo un insieme di stima con 15272 osservazioni ed un insieme di verifica con 3818 osservazioni;
- si verifica che nell'insieme di stima le classi della variabile risposta siano bilanciate. In particolare, il 46% degli spezzoni sono di voce femminile ed il 54% di voce maschile, pertanto, poichè le classi della variabile risposta risultano essere bilanciate, non si ha la necessità di svolgere ulteriori operazioni e si può procedere all'analisi esplorativa nell'insieme di stima.

Analisi esplorativa

Tenendo in considerazione che la variabile risposta è una variabile categoriale con due modalità e le variabili esplicative risultano essere tutte quantitative, un'analisi esplorativa (abbastanza) completa ed adeguata si avrebbe con la discretizzazione di ciascuna delle variabili esplicative e l'analisi della distribuzione della variabile dipendente al variare delle singole variabili indipendenti. Poichè l'obiettivo primario non è quello di effettuare l'analisi esplorativa ma di adattare i modelli, si valuta la distribuzione della risposta solamente

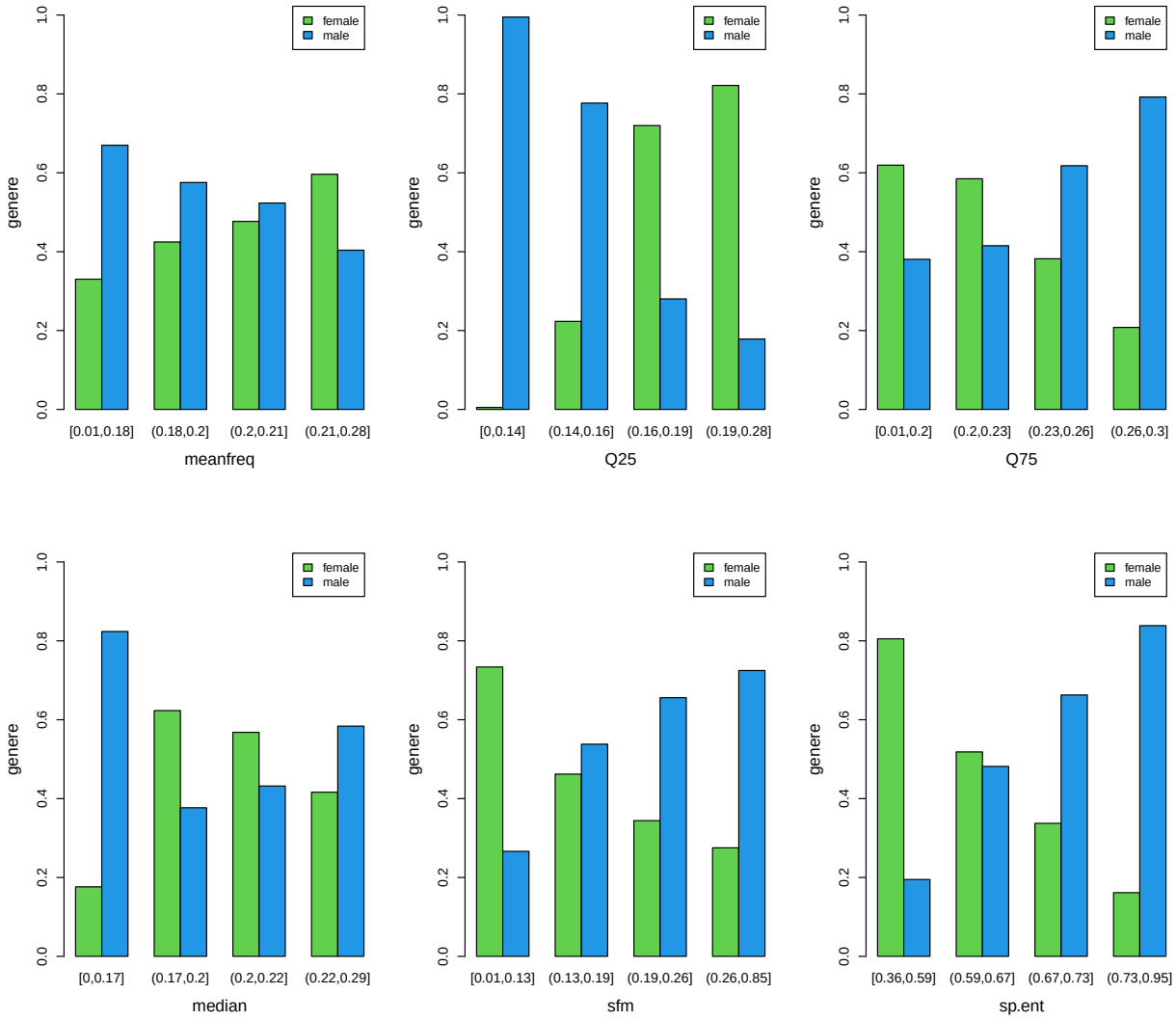

Figure 1: Barplot della variabile risposta rispetto ad alcune variabili esplicative

per alcune variabili esogene.

I barplot in Figura 1 danno indicazione di un possibile effetto molto significativo del primo quartile della frequenza (*Q25*): per valori piccoli del primo quartile la percentuale di spezzoni con voce femminile risulta essere inferiore rispetto alla percentuale di spezzoni con voce maschile, mentre per valori grandi del primo quartile vi è una presenza maggiore di spezzoni con voce femminile. Per quanto riguarda le altre variabili, sembrerebbe esserci un effetto significativo del terzo quartile (*Q75*), dell'entropia spettrale (*sp.ent*) e della piattezza spettrale (*sfm*) in quanto, per tutte e tre, le donne presentano un andamento decrescente all'aumentare del valore della variabile esplicativa mentre gli uomini presentano un andamento crescente. Infine, sembrerebbe non esserci un effetto significativo della frequenza media (*meanfreq*) mentre si nota come per valori piccoli e grandi della mediana (*median*) si ha un numero maggiore di spezzoni con voce femminile, mentre per valori intermedi si hanno più spezzoni con voce maschile.

Conclusa l'analisi esplorativa nell'insieme di stima, si può procedere alla modellazione dei dati.

Modellazione dei dati

Poichè l'interesse dei ricercatori è rivolto alla classificazione della voce, senza focalizzare l'attenzione sul classificare correttamente le voci maschili o le voci femminili, e poichè le classi della variabile risposta nell'insieme di stima risultano essere bilanciate, si utilizzerà una soglia pari a 0.5 e come metrica per il confronto tra i modelli il tasso di errata classificazione.

Modello logistico

Il primo modello che si adatta è il modello di regressione logistica su tutte le variabili esplicative (senza interazione) con funzione di legame la funzione *logit*.

Ad un livello di significatività del 5%, le uniche variabili che risultano avere un effetto statisticamente nullo sulla variabile risposta risultano essere la frequenza dominante minima, la frequenza dominante massima e l'indice di modulazione.
Il tasso di errata classificazione nell'insieme di verifica è pari al **14.75%**.

Modello logistico stepwise

Poichè il modello logistico adattato in precedenza ha messo in luce l'effetto non significativo di alcune variabili esplicative, si ritiene ragionevole adattare un modello di regressione logistica stepwise basato sulla minimizzazione dell'AIC, con ricerca in entrambe le direzioni e a partire dal modello con la sola intercetta.

Nel modello finale sono incluse 15 delle 17 variabili esplicative, ovvero tutte ad eccezione della frequenza dominante massima e dell'indice di modulazione.
Il tasso di errata classificazione nell'insieme di verifica è pari al **14.90%**.

Albero di classificazione

Si prosegue la fase di modellazione con l'adattamento di un albero di classificazione, con l'entropia come funzione da minimizzare. Poichè questo modello prevede la selezione del numero di foglie ottimale, si divide l'insieme di stima in due sottoinsiemi: un insieme di stima ridotto in cui far crescere l'albero e un insieme di convalida in cui effettuare la fase di potatura. Nella fase di crescita dell'albero viene impostata una numerosità minima di osservazioni per foglia pari a 2 e una diminuzione dell'entropia per consentire uno split pari almeno a 0.000005, in modo da far diventare l'albero il più profondo possibile. Nella fase di

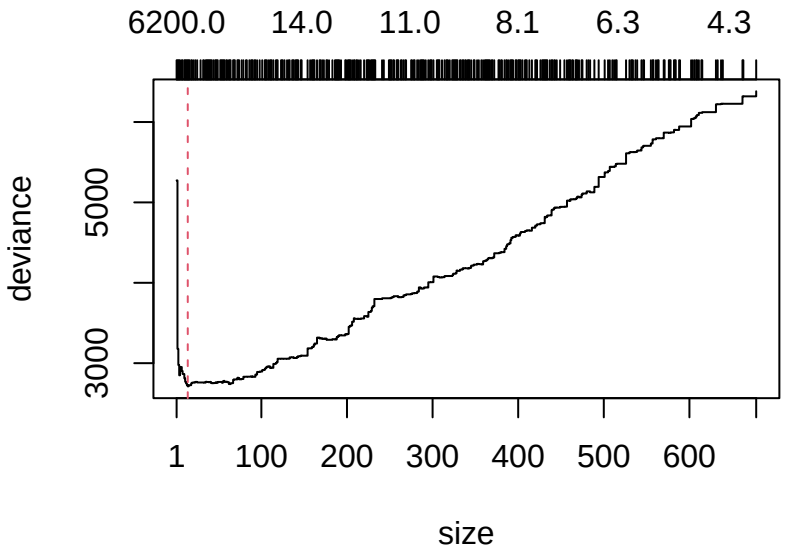

Figure 2: Errore nell'insieme di convalida in funzione del numero di foglie

potatura viene valutata la devianza nell'insieme di convalida al variare del numero di foglie dell'albero. Il grafico in Figura 2 mostra come il minimo si ottenga con un albero con 14 foglie.

Uno dei pregi di questo modello è la facile interpretabilità nel caso in cui l'albero sia poco profondo. A tal riguardo, il grafico in Figura 3 mostra gli split dell'albero selezionato, mettendo in luce che il primo quartile e la deviazione standard della frequenza sono le variabili che entrano in gioco nelle prime suddivisioni dell'albero.

Nell'insieme di verifica il tasso di errata classificazione è pari al **14.41%**.

Modello additivo

Il modello successivamente adattato è il modello additivo generalizzato (GAM). Sono utilizzate le splines di lisciamento con al massimo 3 gradi di libertà equivalenti come lisciatori per le variabili quantitative e una procedura di tipo passo a passo ibrida basata sulla minimizzazione dell'AIC implementata nell'insieme di stima.

Il modello finale include la deviazione standard, la mediana, il primo e il terzo quartile della frequenza, la curtosi e l'asimmetria della distribuzione, la piattezza e l'entropia spettrale, la media e il minimo della frequenza dominante. Nel dettaglio, gli effetti delle variabili appena menzionato sono tutti stimati tramite splines di lisciamento con 3 gradi di libertà equivalenti.
Nell'insieme di verifica si ottiene un tasso di errata classificazione pari al **13.69%**.

Random forest

Si procede con l'adattamento del *random forest*. Il parametro di regolazione del modello è il numero di covariate da considerare ad ogni suddivisione dell'albero. A tal riguardo, l'insieme di stima viene diviso in un insieme di stima ridotto e uno di convalida e viene adattato il *random forest* con 250 alberi in corrispondenza

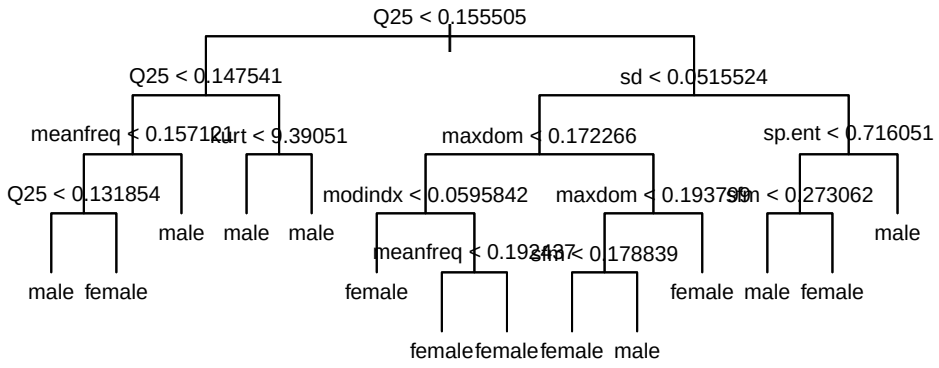

Figure 3: Albero di classificazione selezionato

di ognuno dei possibili valori del numero di covariate considerate. Il numero di covariate selezionato è il valore corrispondente al modello con tasso di errata classificazione minore nell'insieme di convalida. La Figura 4 mostra che, con tale procedura, si sceglie un numero di colonne da campionare in ogni albero pari a 8.

Successivamente il modello selezionato è adattato sull'intero insieme di stima e permette di ottenere un tasso di errata classificazione nell'insieme di verifica pari al **9.71%**.
Questo modello permette di ottenere una misura di importanza delle variabili esplicative, senza però avere indicazione sulla direzione dell'effetto di esse sulla risposta. In queso caso, la Figura 5 mette in luce che le variabili più importanti in termini di diminuzione dell'errore di previsione risultano essere il primo quartile, la piattezza spettrale, la media e il massimo della frequenza fondamentale e la frequenza media.

Bagging

Si adatta un *bagging* con alberi di classificazione. Viene calcolato l'errore OOB per diversi valori del numero di campioni boostrap (e quindi di alberi) utilizzato dal modello, scegliendo il valore per cui l'errore OOB è minore. In questo caso è pari a 150, come si evince dalla Figura 6, in cui si riporta il grafico dell'errore OOB in funzione del numero di campioni bootstrap.
Il modello selezionato ottiene sull'insieme di verifica un tasso di errata classificazione pari al **10.05%**.

Boosting

Si adatta un *boosting* con alberi di classificazione. Per individuare il numero di alberi necessari a stabilizzare l'errore di previsione, si divide l'insieme di stima in un insieme di stima ridotto e uno di convalida. La Figura 7 mostra l'errore di previsione nell'insieme di convalida in funzione del numero di iterazioni dell'algoritmo, facendo notare che l'errore si stabilizza dopo 130 iterazioni.

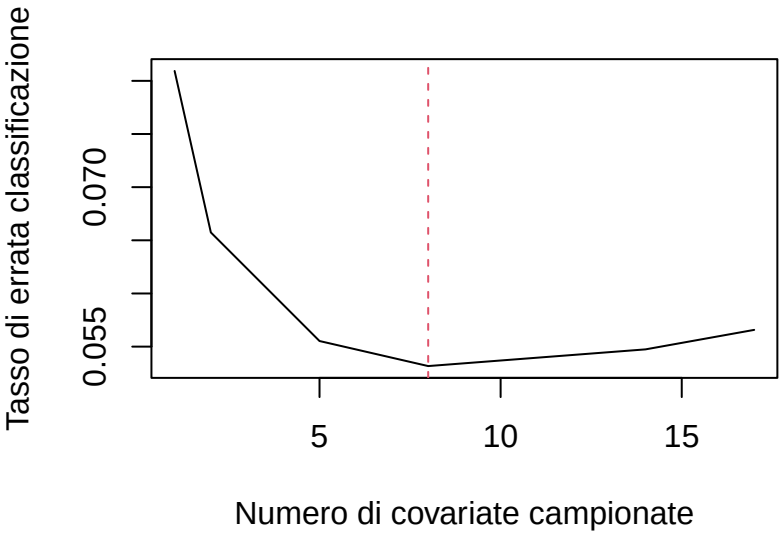

Figure 4: Errore nell'insieme di convalida in funzione del numero di covariate campionate

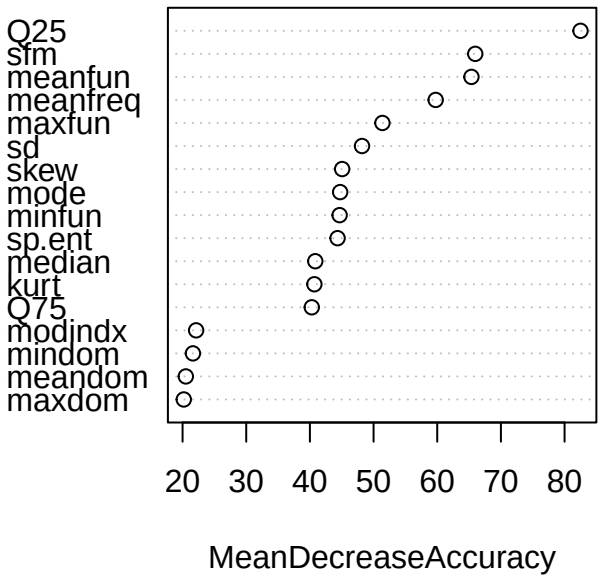

Figure 5: Importanza delle variabili nel random forest

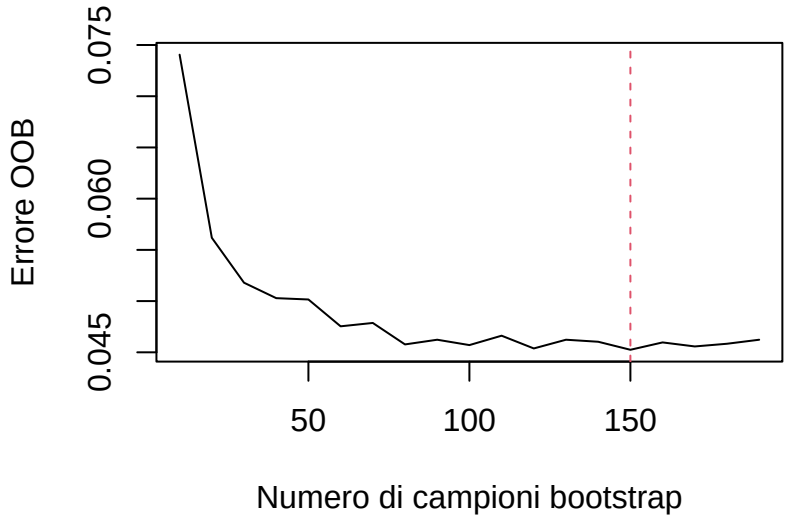

Figure 6: Errore OOB (Out-Of-Bag) nell'insieme di stima in funzione del numero di campioni bootstrap

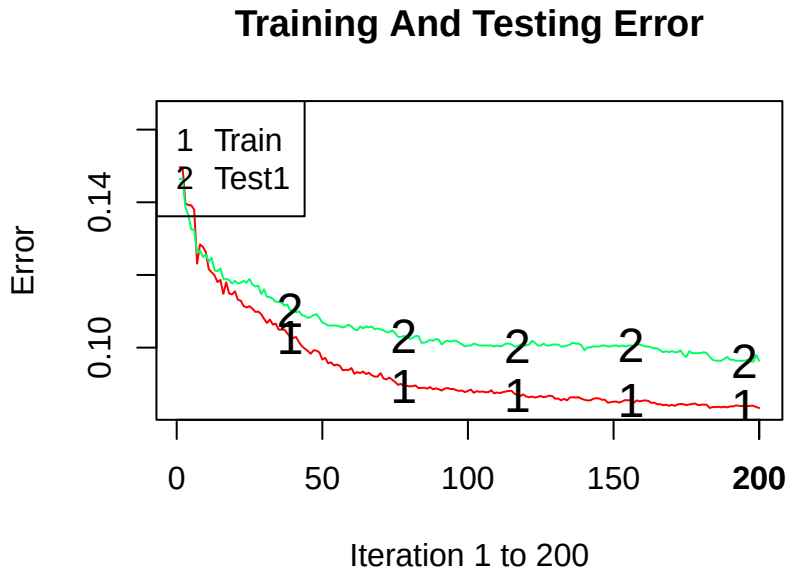

Figure 7: Errore di previsione nell'insieme di convalida in funzione del numero di iterazioni

Il modello selezionato è riadattato sull'intero insieme di stima e ottiene un tasso di errata classificazione nell'insieme di verifica pari al **10.51%**.
Anche questo modello ha il pregio di portare informazione sull'importanza delle variabili esplicative. La Figura 8 permette di far notare che le variabili maggiormente presenti negli stumps risultano essere la mediana, la moda, il primo e il terzo quartile, la media e il massimo della frequenza dominante.

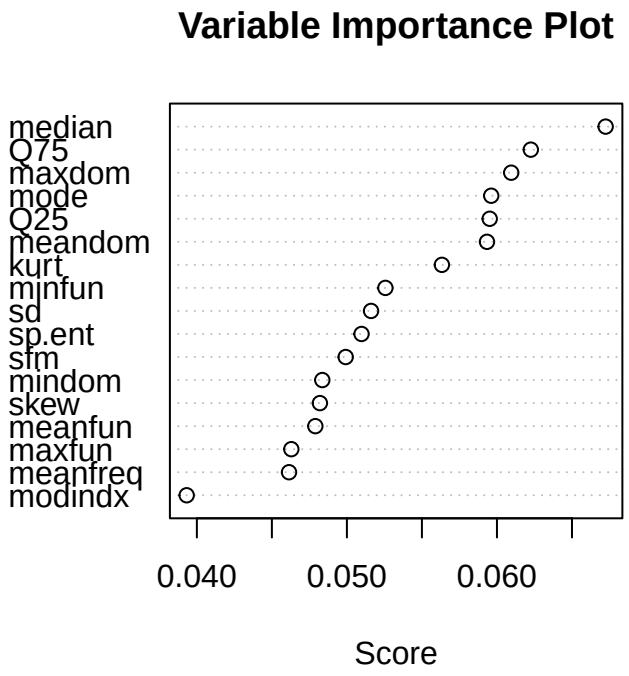

Figure 8: Importanza delle variabili nel boosting

Risultati

Nella Tabella 1 si riportano i risultati ottenuti coi diversi modelli adattati in termini di accuratezza (ovvero il complemento ad 1 del tasso di errata classificazione).

Table 1: Tasso di accuratezza dei modelli adattati

	Accuratezza
Random Forest	0.9030
Bagging	0.8995
Boosting	0.8949
Gam stepwise	0.8631
Albero	0.8559
Logistico	0.8525
Logistico stepwise	0.8510

Si nota come il modello che permette di riconoscere in maniera migliore la voce dei file audio risulta essere il *random forest*, in quanto ha un'accuratezza del 90.29%, seguito dal *bagging* e dal *boosting*, i quali hanno un

tasso di accuratezza praticamente identico a quello del modello migliore e pari rispettivamente all'89.95% e all'89.49%. Una capacità predittiva peggiore si ha con il modello additivo generalizzato (86.31%), l'albero di classificazione (85.59%), il modello logistico (85.25%) e il modello logistico stepwise (85.10%).
Focalizzando l'attenzione sul *random forest*, come già è stato detto in precedenza, questo modello permette di avere una misura di importanza delle variabili esplicative, senza però avere indicazione sulla direzione dell'effetto di queste variabili sulla risposta. In questo caso, le variabili maggiormente importanti risultano essere il primo quartile, la piattezza spettrale, la media e il massimo della frequenza fondamentale e la frequenza media.