UNIVERSITÀ
DEGLI STUDI
FIRENZE

Università degli Studi di Firenze

SCUOLA DI ECONOMIA E MANAGEMENT
Corso di Laurea in Statistica scienze attuariale e finanziare

Corso di MASL

Relazione finale:
Gli animali e gli alberi

Candidato:
Chiara Camerota
Matricola 6235110

Relatore:
Prof. A.Gottard

Indice dei Contenuti

Capitolo 1

Abstract

Il problema affrontato nelle prossime pagine, è di classificazione multinomiale. In particolare, il dataset contiene caratteristiche di diversi animali che fungono da predittori per la classificazione di questi in sette categorie. Per affrontare il problema sono stati utilizzati: random forest (scelto per la sua flessibilità), stochastic gradient boosting (scelto per la sua restistenza all'overfitting) e il k-nearest neighbors. Il fine del seguente lavoro è confrontare i diversi metodi, sia sul piano predittivo, sia su quello di costo.

Capitolo 2

Gli animali

2.1 Presentazione dei dati

I dati considerati sono la raccolta di informazioni riguardo 101 animali presenti in uno zoo. Il dataset è formato da una variabile di risposta multinomiale, con 7 modalità, e un insieme di predittori per lo più booleani. L'obiettivo è quello di classificare gli animali nelle seguenti classi:

- mammiferi
- uccelli
- rettili
- pesci
- anfibi
- insetti
- invertebrati.

I predittori contengono caratteristiche fisiche e caratteriali degli animali, in particolare sono:

hair *(Booleano)*: indica se l'animale presenta peli;

feathers *(Booleano)* : indica se ha un piumaggio;

eggs *(Booleano)* : indica se è oviparo;

milk *(Booleano)* : indica se allatta;

airborne *(Booleano)* : indica se è un volatile;

aquatic *(Booleano)* : indica se è acquatico;

predator *(Booleano)* : indica se è predatore;

toothed *(Booleano)*: indica se ha i denti;

[backbone] *(Booleano)*: indicasese ha una spina dorsale;

breathes *(Booleano)* : indica se respira;

venomous *(Booleano)*: indicase se è velenoso;

fins *(Booleano)* : indicasese ha le pinne;

legs *(insieme di valori "0,2,4,5,6,8")*: indica il numero di gambe;

tail *(Booleano)* : indica se ha una coda;

domestic *(Booleano)* : indica se è domestico;

catsize *(Booleano)* : indica se ha le dimensioni di un gatto.

.

2.2 Analisi Preliminari

Molto spesso si tende a confondere i problemi di classificazione, con quelli di clustering, quindi è opportuno specificare cosa si intende per classificazione, ovvero: l'insieme di metodologie statistiche che aspirano ad assegnare una classe ad un'osservazione di classe sconosciuta, sulla base di informazioni fornite da un campione di classe invece nota.
Avendo a disposizione il numero di classi e le assegnazioni di ognuna, risulta inopportuno usare metodi di clustering (ovvero: insieme di tecniche che hanno come scopo l'individuazione di gruppi).

Una volta capito il tipo di problema che stiamo affrontando, è importante capire la forma dei dati, un metodo veloce e intuitivo è quello di plottare la densitÃ dei diversi gruppi con le diverse variabili. Nella seguente figura si può notare come la presenza di una classe rispetto ad un'altra cambia a seconda della variabile considerata.

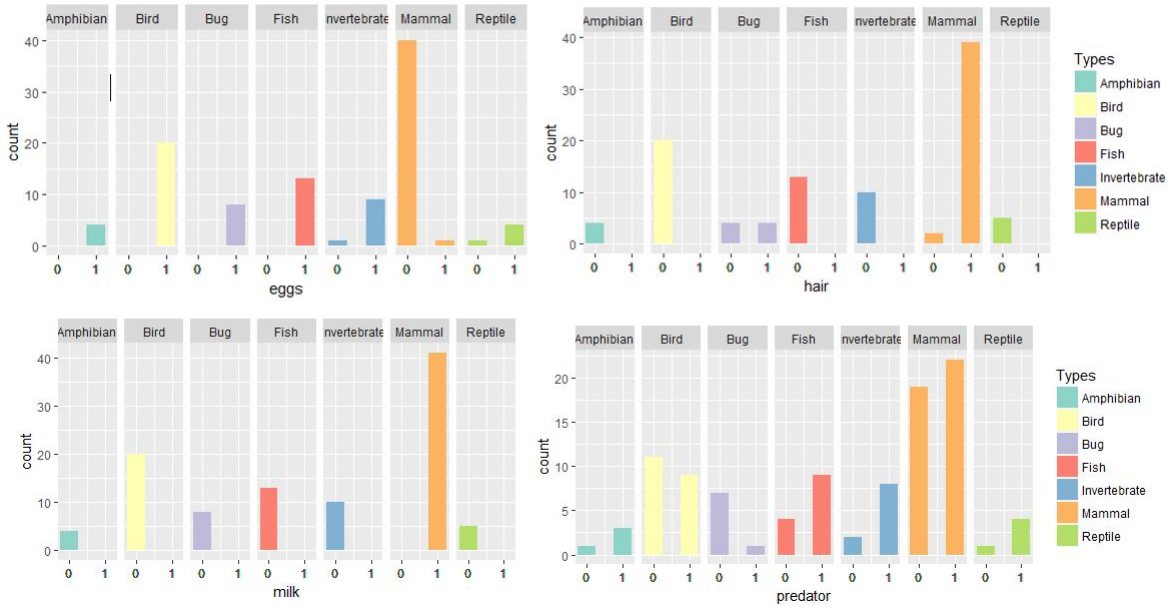


Figura 2.1: Istogrammi di alcuni predittori per ogni classe

Quanto ci mostra l'immagine è facilmente spiegabile, infatti alcune variabili, tendono a caratterizzare completamente la classe degli animali, per esempio: il fatto che un animale sia oviparo, esclude la sua assegnazione alla classe dei mammiferi (fatta eccezzione per i monotremi).
Un'altra importante relazione da esplorare è la correlazione tra variabili, questa ci darà un'idea completa del legame tra variabili. Il seguente grafico rappresenta le correlazioni, calcolate con l'indice di Sperman, sia graficamente che numericamente.

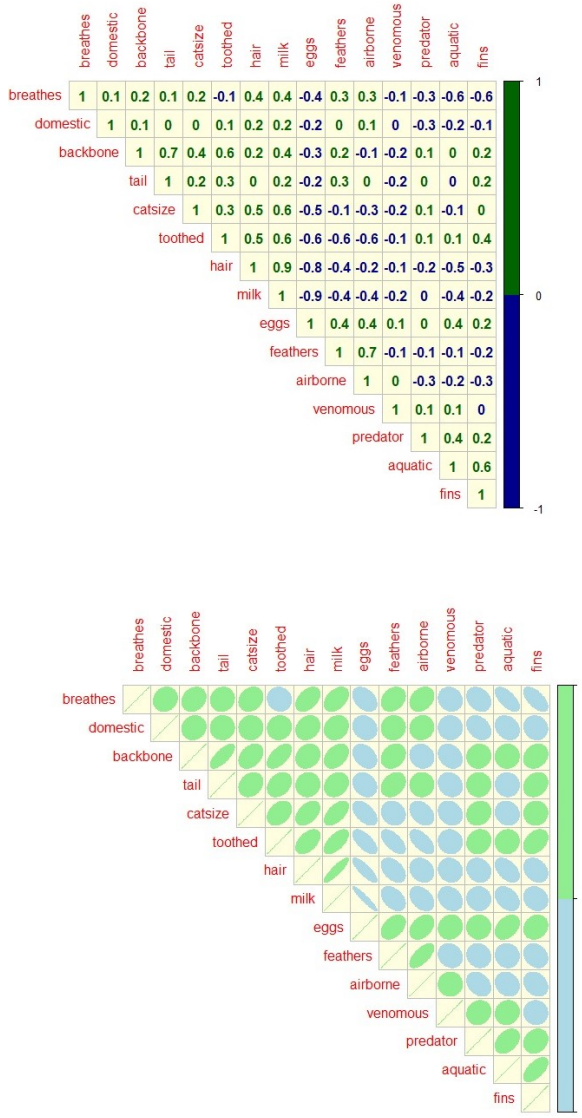


Figura 2.2: Grafici di correlazione tra i predittori

Possiamo notare che i predittori sono ampiamente e differentemente correlati, per lo più negativamente.

Per applicare i diversi metodi sui dati, è necessario pre-processare i dati, come prima operazione, dividiamo il nostro dataset in un un train set, contenente il settanta percento delle osservazioni, e un test set, con le restanti osservazioni.

Attraverso una summary possiamo avere un'idea di come è costituito il train set.

```
> summary(data[dta == 1,])
      hair          feathers         eggs            milk          airborne        aquatic
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :1.0000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.4133   Mean   :0.2133   Mean   :0.5867   Mean   :0.3867   Mean   :0.2533   Mean   :0.3467
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.5000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000

    predator         toothed          backbone         breathes         venomous           fins
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:0.0000
 Median :1.0000   Median :1.0000   Median :1.0000   Median :1.0000   Median :0.00000   Median :0.0000
 Mean   :0.5067   Mean   :0.5733   Mean   :0.7867   Mean   :0.7733   Mean   :0.06667   Mean   :0.1467
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000

      legs            tail         domestic         catsize               Types
 Min.   :0.000   Min.   :0.00   Min.   :0.0000   Min.   :0.0000   Amphibian   : 3
 1st Qu.:2.000   1st Qu.:0.00   1st Qu.:0.0000   1st Qu.:0.0000   Bird        :16
 Median :2.000   Median :1.00   Median :0.0000   Median :0.0000   Bug         : 6
 Mean   :2.867   Mean   :0.72   Mean   :0.1067   Mean   :0.4133   Fish        : 9
 3rd Qu.:4.000   3rd Qu.:1.00   3rd Qu.:0.0000   3rd Qu.:1.0000   Invertebrate:10
 Max.   :8.000   Max.   :1.00   Max.   :1.0000   Max.   :1.0000   Mammal      :29
                                                                  Reptile     : 2
```

Figura 2.3: Summary del train set

```
         ytest          hair           feathers          eggs             milk
 Amphibian   : 1   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 Bird        : 5   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Bug         : 4   Median :0.0000   Median :0.0000   Median :1.0000   Median :0.0000
 Fish        : 7   Mean   :0.4545   Mean   :0.1515   Mean   :0.5758   Mean   :0.4242
 Invertebrate: 1   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Mammal      :14   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
 Reptile     : 1
    airborne         aquatic         predator         toothed         backbone
 Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.0000
 Median :0.0000   Median :0.000   Median :1.0000   Median :1.000   Median :1.0000
 Mean   :0.1818   Mean   :0.303   Mean   :0.5152   Mean   :0.697   Mean   :0.8485
 3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.000   Max.   :1.0000

    breathes         venomous           fins             legs            tail
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
 1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.0000
 Median :1.0000   Median :0.0000   Median :0.0000   Median :4.000   Median :1.0000
 Mean   :0.7576   Mean   :0.1212   Mean   :0.2121   Mean   :2.879   Mean   :0.7576
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :6.000   Max.   :1.0000

    domestic         catsize
 Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000
 Mean   :0.0303   Mean   :0.4848
 3rd Qu.:0.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000
```

Figura 2.4: Summary del test set

.

Capitolo 3

Gli alberi

Gli alberi di decisione, o di classificazione, sono analoghi agli alberi di regressione, infatti usano, non solo, le stesse operazioni sugli alberi, ma anche la stessa logica.
I dati originari vengono divisi ricorsivamente in due parti rispetto ad uno, o più generici attributi, la suddivisione produce una gerarchia ad albero, dove i sottoinsiemi (di record) vengono chiamati nodi e, quelli finali (o terminali), foglie. In particolare, i nodi sono etichettati con il nome degli attributi, gli archi sono etichettati con i possibili valori dell'attributo soprastante, mentre le foglie dell'albero sono etichettate con i differenti valori dell'attributo target (valori che descrivono le classi di appartenenza).
Un oggetto è classificato seguendo un percorso lungo l'albero che porti dalla radice ad una foglia. I percorsi rappresentano le regole di classificazione (o regole produttive).
Un'importante differenza tra alberi di classificazione e di regressione è il diverso metodo di valutazione degli split. Nel caso di classificazione non è possibile calcolare l'RSS, per cui si ricorre ad altri metodi, i più comuni sono:

- **Classification error rate:** proporzione di osservazioni, del train set, nella regione, che non appartengono alla casse più comune;

  $$E = 1| - max_k(\hat{p}_{mk})$$

  dove $\hat{p}_{mk}$ indica la proporzione di osservazioni del train set nella regione m-sima che appartiene alla classe k.
  Questo indice non è ottimale se si vuole un albero parsimonioso, infatti non risente della grandezza di questo.

- **Indice di Gini:** misura la variabilità all'interno della classe k.

  $$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_m k)$$

Questo indice può indicare anche il grado di purezza di un nodo, infatti se assume bassi valori numerici nella regione di riferimento predomina una classe.

- **Cross-entropy:** metodo alternativo al precedente.

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_m k.$$

Negli ultimi anni, questi metodi sono sempre più utilizzati, infatti sono di semplice interpretazione per gli inesperti, in alcuni casi è possibile visualizzarli e non neccessitano della creazione di variabili dummy in caso di variabili qualitative.

3.1 CART

La tipologia di alberi decisionali più semplice sono i CART (Classification and Regression Tree), di seguito è riportato l'algoritmo.

1. Scegli la radice (primo nodo).

2. Per ogni predittore X trova il sottoinsieme S che minimizza la somma dell'indice di impurità dei due nodi figli. Scegli la coppia che rende X e S minimo.

3. Quando la regola di uscita è soddisfatta, finisce l'algoritmo. Altrimenti continua ad iterare il punto 2. fino a che non finiscono i nodi figli.

3.1.1 Applicazione

Il pacchetto usato per le applicazioni è *Rpart*, questo usa il metodo di cross-validation k-fold con k=10 di default, inoltre si ha la possibilità di scegliere se usare l'indice di Gini o l'error rate come regola di split.
Di seguito sono raffigurati due alberi ottenuti con le due diverse regole per lo split, come si può notare l'albero finale cambia, anche se come vedremo più avanti, la classificazione è pressochè identica.

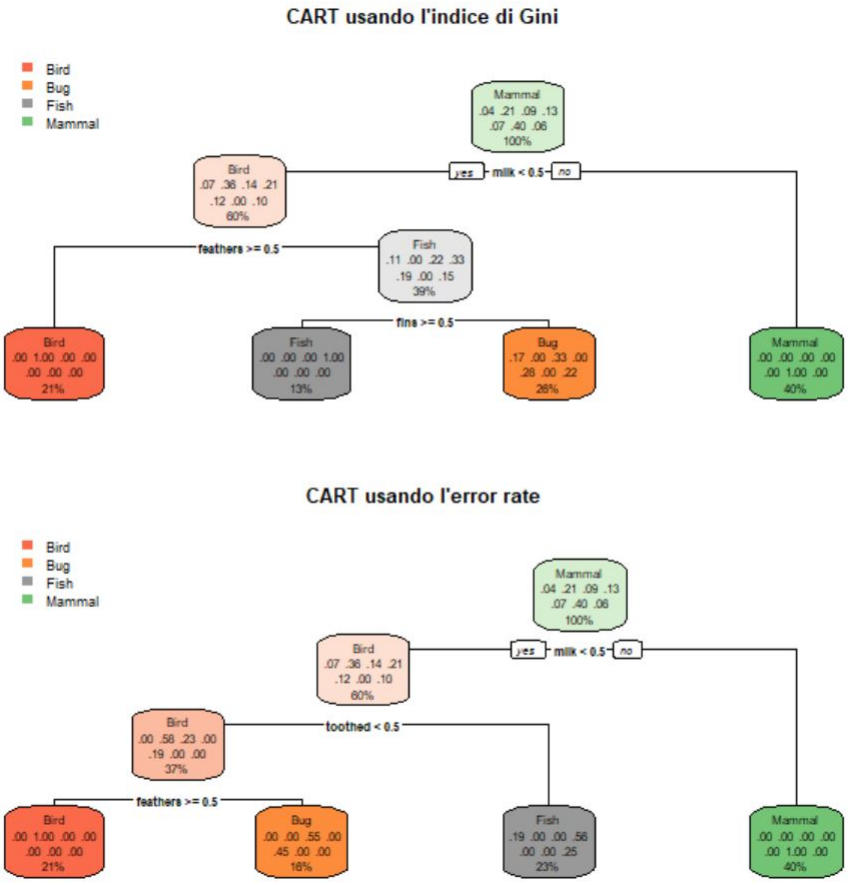


Figura 3.1: I due diversi alberi finali

Il grafico riporta sia le variabili usate per lo split, sia la proporzione di ogni classe presente nel sottoinsieme, sia la percentuale di dati che si trovano in quel gruppo.
Un interessante parametro è quello di complessitá (cp) che indica quanto diminuisce l'errore dopo ogni split, fino ad un minimo fissato.
È possibile visualizzare come il cp vari a seconda della grandezza dell'albero e dell'errore relativo.

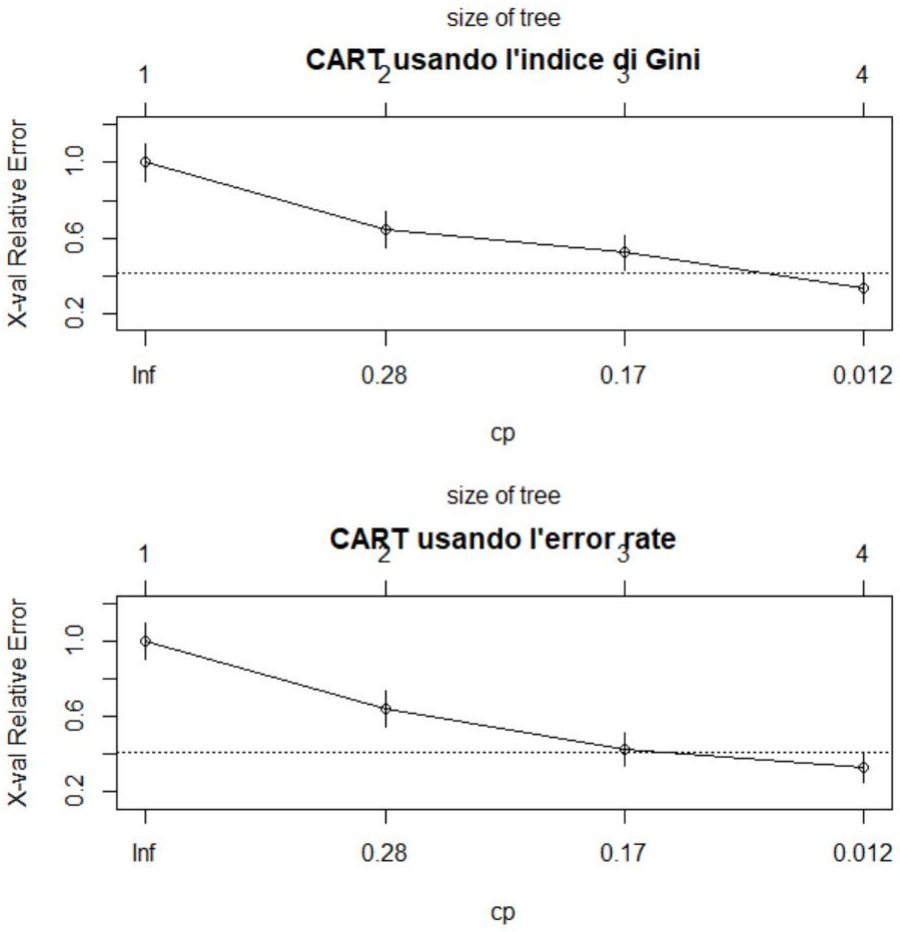


Figura 3.2: L'andamento del cp

In questo caso la curva dell'albero che usa l'error rate è più rapida dell'altra, questo perchè il terzo split Ã¨ fatto su diverse variabili nei due alberi.
Provando i due modelli sul test set si nota come entrambi misclassificano le unità che appartengono a classi poco numerose, questo perchè usano solo quattro variabili per la costruzione, ignorando variabili come *backbone* che identifica un'intero gruppo.

	ytest						
cpredg	Amphibian	Bird	Bug	Fish	Invertebrate	Mammal	Reptile
Amphibian	0	0	0	0	0	0	0
Bird	0	5	0	0	0	0	0
Bug	1	0	2	0	5	0	1
Fish	0	0	0	4	0	0	0
Invertebrate	0	0	0	0	0	0	0
Mammal	0	0	0	0	0	13	0
Reptile	0	0	0	0	0	0	0

Figura 3.3: Matrice di confondimento, usando l'indice di Gini.

	ytest						
cpredi	Amphibian	Bird	Bug	Fish	Invertebrate	Mammal	Reptile
Amphibian	0	0	0	0	0	0	0
Bird	0	5	0	0	0	0	0
Bug	0	0	2	0	5	0	1
Fish	1	0	0	4	0	0	0
Invertebrate	0	0	0	0	0	0	0
Mammal	0	0	0	0	0	13	0
Reptile	0	0	0	0	0	0	0

Figura 3.4: Matrice di confondimento, usando l'error rate.

3.2 Random Forest

Il *bagging* o *boostrap aggregation* è una procedura volta a ridurre la variabilità presente nei dati, l'idea è semplice: si divide il data set iniziali in più parti, con il metodo boostrap, su ognuna delle quali si costruisce un classificatore, infine si combinano tutti questi.
Un problema di questo metodo è che i classificatori sono altamente correlati tra loro, per evitare questo, nasce la Random Forest: costruisce un certo numero di alberi, basati su campioni boostrap, considerando un sottoinsieme di predittori. Per quanto riguarda le operazioni sugli alberi: rimangono invariate.
Una possibilità che si ha in più del metodo precedente è che possiamo stilare una graduatoria dei predittori in base alla loro importanza,basata sull'indice di Gini nel nodo corrispondente.
Infine, la Random Forest, al crescere dei campioni boostrap riduce la probabilità di overfitting.

3.2.1 Applicazione

Il pacchetto usato è **RandomForest**.
In primo luogo è stata applicata una random forest con 150 alberi, usando due variabili a confronto per ogni split, e l'indice di Gini.
Di seguito è riportato il grafico dell'anadamento dell'errore al variare del numero degli alberi.

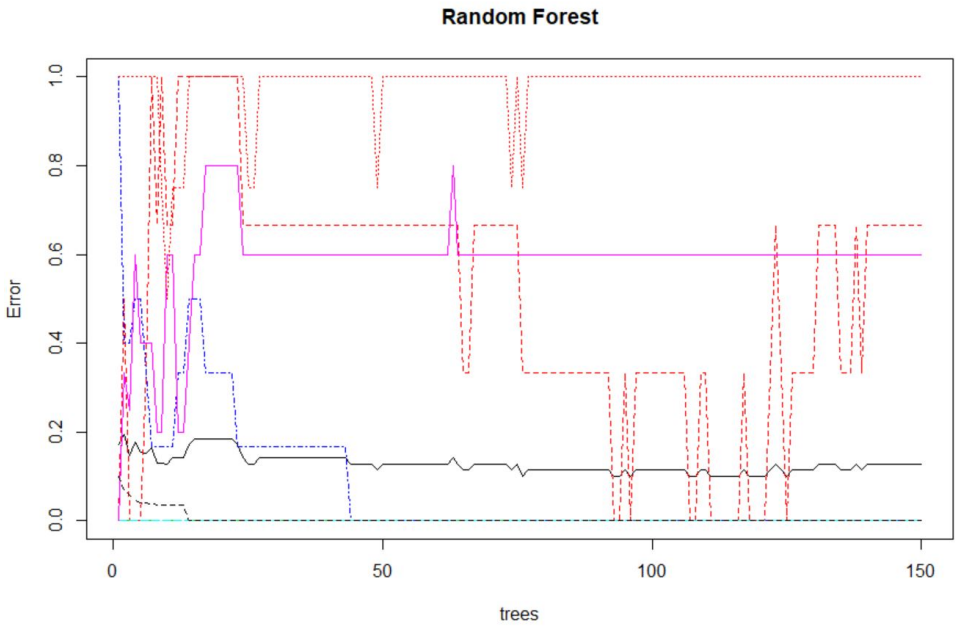


Figura 3.5: Andamento dell'errore

Avendo un numero parsimonioso di osservazioni, è interessante vedere le variabili più importanti, così da poter provare la perfomance dello stesso modello su un sottoinsieme di predittori.

Applicando la Random Forest solo sulle suddette variabili, otteniamo un errore identico al precente, ma osservando le matrici di confondimento si può notare come gli errori di misclassificazione differiscono, sebbene l'errore totale rimane identico.

	ytest						
predf	Amphibian	Bird	Bug	Fish	Invertebrate	Mammal	Reptile
Amphibian	1	0	0	0	0	0	0
Bird	0	5	0	0	0	0	1
Bug	0	0	2	0	2	0	0
Fish	0	0	0	4	0	0	0
Invertebrate	0	0	0	0	3	0	0
Mammal	0	0	0	0	0	13	0
Reptile	0	0	0	0	0	0	0

Figura 3.6: Matrice di confondimento Random Forest

	ytest						
predfvi	Amphibian	Bird	Bug	Fish	Invertebrate	Mammal	Reptile
Amphibian	1	0	0	0	0	0	0
Bird	0	5	0	0	0	0	1
Bug	0	0	2	0	3	0	0
Fish	0	0	0	4	0	0	0
Invertebrate	0	0	0	0	2	0	0
Mammal	0	0	0	0	0	13	0
Reptile	0	0	0	0	0	0	0

Figura 3.7: Matrice di confondimento Random Forest con selezione di variabili

3.3 Stochastic Gradient Boosting

Il metodo gradient boosting è una tecnica di machine learning per problemi di regressione e classificazione, questo produce un modello predittivo come combinazione di modelli predittivi deboli, in questo caso alberi. Costruisce un modello in maniera simile ai metodi di boosting, e li generalizza permettendo l'ottimizzazione di una funzione di perdita differenziabile arbitraria. In particolare, con questa tecnica gli alberi crescono sequenzialmente: ogni albero cresce sfruttando le informazioni di quello precedente.
Una variante molto interessante è la versione stocastica, introdotta da Friedam, motivata dal metodo di bagging di Breiman.
Per ogni iterazione dell'algoritmo, un base learner viene applicato ad un campione casuale senza reintroduzione del train set; in questo modo l'accuratezza del metodo cresce notevolmente.
Il parametro di campionamento f è una frazione del train set, se è pari ad 1, l'algoritmo è deterministico, bassi valori introducono casualità e prevengono l'overfitting, viene usato anche come parametro di regolarizzazione. Tipicamente è impostato a f=0.5.

3.3.1 Applicazione

Il pacchetto usato in questo caso è *caret*.
L'output descrive eusastivamente il modello.

```
Stochastic Gradient Boosting

68 samples
16 predictors
 7 classes: 'Amphibian', 'Bird', 'Bug', 'Fish', 'Invertebrate', 'Mammal', 'Reptile'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 5 times)
Summary of sample sizes: 54, 54, 54, 54, 56, 53, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  Accuracy   Kappa
  1                   50      0.8756410  0.8350528
  1                  100      0.8990989  0.8673443
  1                  150      0.9050330  0.8749537
  2                   50      0.8659267  0.8217592
  2                  100      0.8904322  0.8560520
  2                  150      0.8871648  0.8515514
  3                   50      0.8839560  0.8460873
  3                  100      0.8959267  0.8623446
  3                  150      0.8990696  0.8673871

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning
 parameter 'n.minobsinnode' was held constant at a value of 10
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 150, interaction.depth =
 1, shrinkage = 0.1 and n.minobsinnode = 10.
```

Figura 3.8: Descrizione del modello

È da notare che la cross validation ripetuta aiuta l'accuratezza della classificazione, come si può vedere dal seguente grafico.

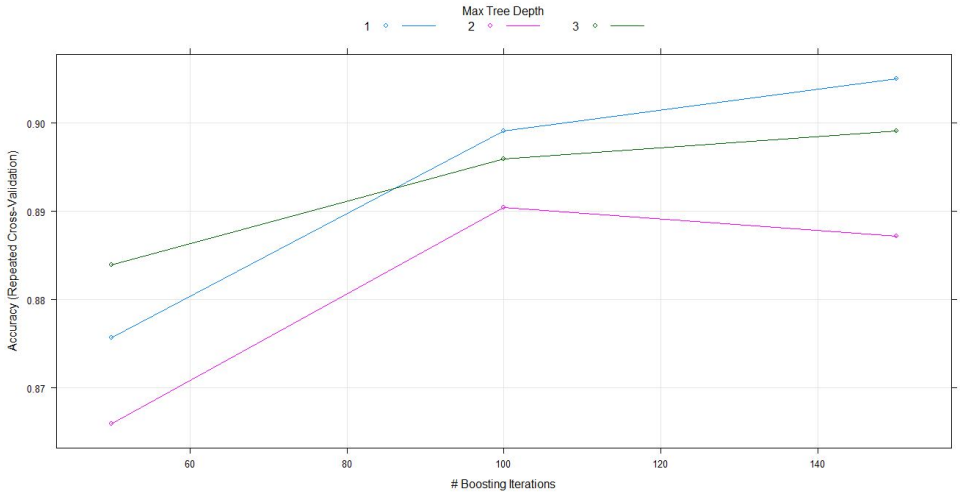
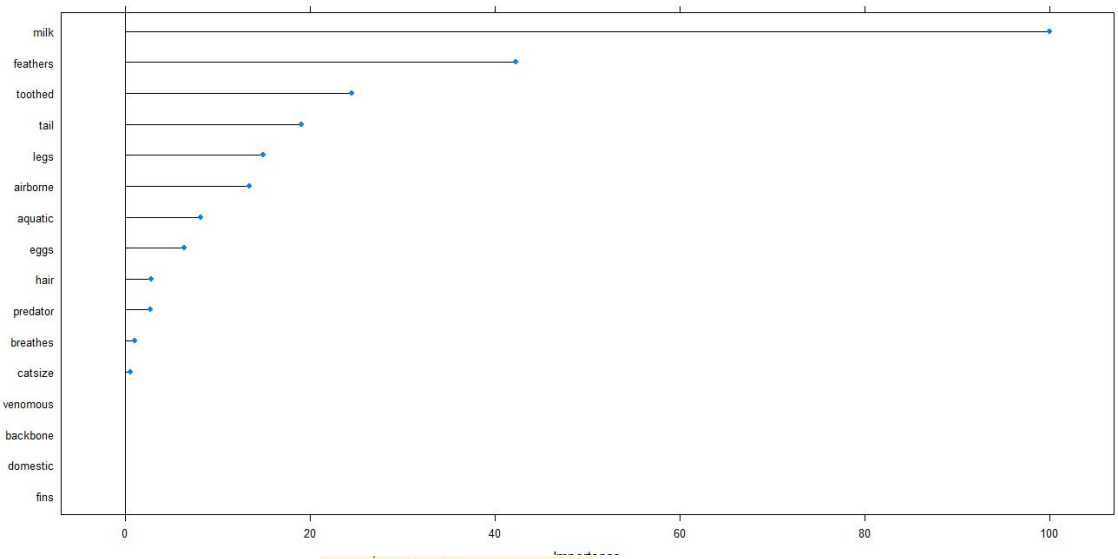


Figura 3.9: Accuratezza del modello

Il metodo permette anche di vedere l'importanza delle variabili, ma riducendo i predittori la performance non cambia.



gbm variable importance

	Overall
milk	100.0000
feathers	42.2873
toothed	24.4825
tail	19.0606
legs	14.9374
airborne	13.3908
aquatic	8.1808
eggs	6.3594
hair	2.8504
predator	2.7492
breathes	0.9745
catsize	0.5562
fins	0.0000
domestic	0.0000
venomous	0.0000
backbone	0.0000

Figura 3.10: Variabili e relativa importanza

La matrice di confondimento di entrambi i modelli è la seguente.

	Reference						
Prediction	Amphibian	Bird	Bug	Fish	Invertebrate	Mammal	Reptile
Amphibian	1	0	0	0	0	0	0
Bird	0	5	0	0	0	0	0
Bug	0	0	2	0	0	0	0
Fish	0	0	0	7	0	0	0
Invertebrate	0	0	2	0	1	0	0
Mammal	0	0	0	0	0	14	0
Reptile	0	0	0	0	0	0	1

Figura 3.11: Matrice di confondimento

Capitolo 4

Conclusioni

Questo ultimo capitolo riassume le prestazioni attraverso l'errore di missclasificazione sul test set.

Classificatore	Accuratezza
CART (entrambi)	0.81
Random Forest (senza selezione di variabili)	0.90
Random Forest (con selezione di variabili)	0.93
Stochastic GB (entrambi)	0.93

Sulla base dei test, il modello migliore è la random forest con la selezione di variabili, in quanto meno costoso in termini di informazioni, ma più laborioso da ottenere. Anche lo stochastic gradient boosting da ottimi risultati, e non prevede analisi preliminari.
In conclusione i due metodi in questo caso si equivalgono in quanto a prestazioni, anche se l'ultimo metodo è il più accurato e robusto.