UN'ANALISI DI CLASSIFICAZIONE BINARIA

Giulia Lumia

Abstract

L'analisi in oggetto ha come scopo principale l'individuazione di un classificatore in grado di determinare la fascia di prezzo di appartenenza dei telefoni cellulare prodotti dall'azienda di Bob.
In questo modo Bob, tenendo conto delle caratteristiche strutturali e funzionali dei propri prodotti, potrà applicare loro prezzi adeguati ed in linea con quelli applicati dalla concorrenza.
Nell'ambito di un'analisi di classificazione statistica, numerose sono le procedure adattabili, pertanto la scelta è stata effettuata valutando e confrontando la bontà dei risultati da esse generati.
È possibile ricondurre l'analisi a cinque fasi:

- ANALISI ESPLORATIVA, finalizzata principalmente a descrivere i dati in esame;
- ANALISI DI CLASSIFICAZIONE TRAMITE SINGOLO DECISION TREE;
- ANALISI DI CLASSIFICAZIONE TRAMITE RANDOM FOREST;
- CONFRONTO DELLE PRESTAZIONI, finalizzato all'individuazione della procedura migliore da utilizzare;
- CONFRONTO DELLE VARIABILI UTILIZZATE.

1. Analisi esplorativa

Durante la fase esplorativa dell'analisi l'attenzione è stata posta su:

- rilevare l'eventuale presenza di dati mancanti, errati o outlier;
- evidenziare il potere discriminante delle variabili esplicative (caratteristiche dei cellulari) rispetto alle modalità della variabile di risposta (fascia di prezzo).

I dati in oggetto fanno riferimento a 2.000 telefoni cellulare, di cui sono state rilevate le seguenti variabili:

- **battery_power**, ossia la capacità della batteria, misurata in mAh;
- **talk_time**, tempo di chiamata massimo con batteria carica;
- **clock_speed**, velocità del microprocessore;
- **fc**, megapixel della fotocamera frontale;
- **pc**, megapixel della fotocamera posteriore;
- **int_memory**, capacità della memoria interna, misurata in Gigabyte;
- **ram**, capacità della RAM, misurata in Megabyte;
- **n_cores**, numero di processori core;
- **px_height**, risoluzione verticale in pixel;
- **px_width**, risoluzione orizzontale in pixel;
- **sc_h**, altezza del telefono cellulare, espressa in cm;
- **sc_w**, larghezza del telefono cellulare, espressa in cm;
- **m_dep**, profondità dello schermo, espressa in cm;
- **mobile_wt**, peso del telefono cellulare;

VARIABILI QUANTITATIVE

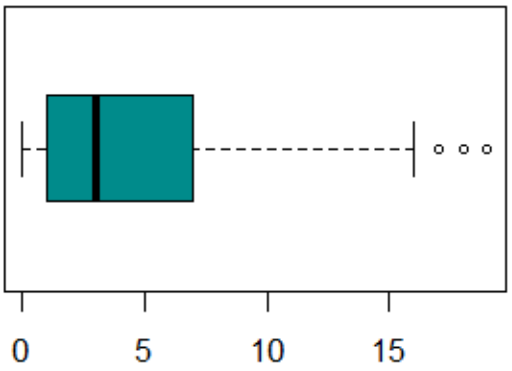
- **three_g**, variabile dicotomica, indicatrice della presenza o meno del 3G;
- **four_g**, variabile dicotomica, indicatrice della presenza o meno del 4G;
- **blue**, variabile dicotomica, indicatrice della presenza o meno del bluetooth;
- **dual_sim**, variabile dicotomica, indicatrice della presenza o meno della doppia sim;
- **touch_screen**, variabile dicotomica, indicatrice della presenza o meno del touch screen;
- **wifi**, variabile dicotomica, indicatrice della presenza o meno del wi-fi;
- **price_range**, codifica della classe di prezzo di appartenenza del cellulare.

VARIABILI QUALITATIVE

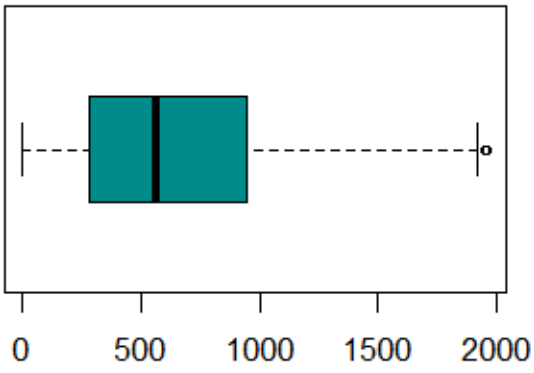
Il dataset in esame non presenta variabili aventi valori mancanti o fuori dal proprio campo di esistenza (es. misure in cm negative). Outlier sono presenti in due sole variabili (GRAFICO 1.1).

GRAFICO N. 1.1: boxplot delle uniche variabili con outlier

FC: mp fotocamera frontale



PX_HEIGHT: risoluzione verticale in pixel



Alcune variabili si manifestano nel dataset tramite valori inusuali (TABELLA 1.1):

- alcuni cellulari hanno uno spessore di 0,1 cm, ma questo potrebbe essere dovuto a telefoni particolarmente sottili (variabile m_dep);
- alcuni cellulari hanno una risoluzione in altezza pari a 0 (variabile px_height);
- trattandosi di telefoni cellulare, i pixel in altezza dovrebbero essere maggiori di quelli in lunghezza;
- alcuni telefoni hanno un'altezza (sc_h) e una lunghezza (sc_w) particolarmente elevate.

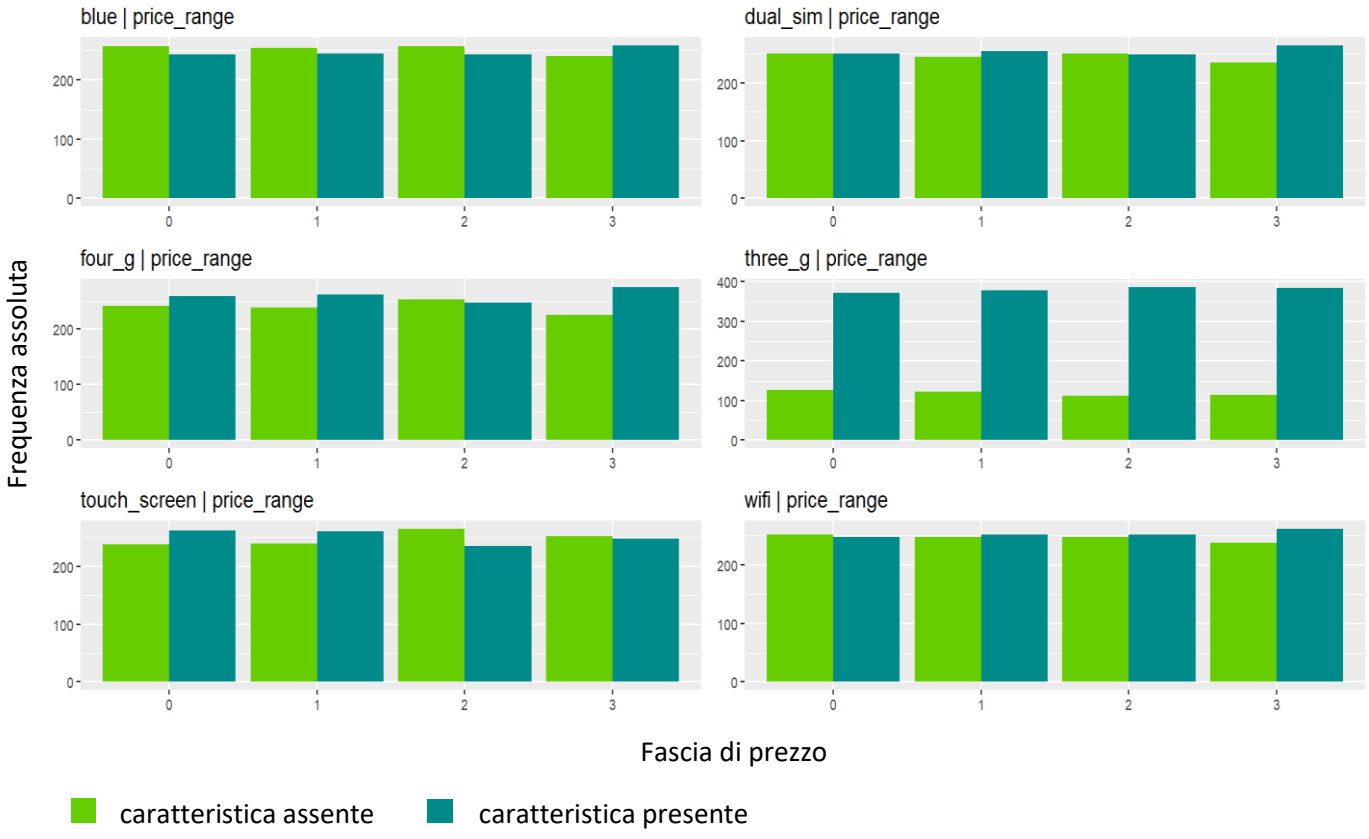
TABELLA N. 1.1: minimo e massimo delle variabili che assumono valori inusuali

VARIABILE	VALORE MINIMO	VALORE MASSIMO
m_dep	0,1 cm	1 cm
px_height	0 pixel	1.960 pixel
px_width	500 pixel	1.998 pixel
sc_h	5 cm	19 cm
sc_w	0 cm	18 cm

Uno degli scopi dell'analisi esplorativa è quello di individuare le variabili esplicative aventi un maggiore **potere discriminante** rispetto alla variabile di risposta. Si tratta di variabili che, noto il valore assunto da un determinato cellulare in oggetto, ne permettono di determinare la fascia di prezzo di appartenenza.
Il potere discriminante è un fattore cruciale nella scelta delle variabili da coinvolgere in una procedura di random forest o di albero di classificazione; per questo motivo, questa fase permette di sviluppare un'idea su quali saranno le variabili oggetto delle procedure di analisi successive.

I GRAFICI 1.2 e 1.3 rappresentano le distribuzioni delle frequenze assolute condizionate di ciascuna variabile esplicativa, qualitativa e quantitativa, rispetto alle modalità della variabile di risposta price_range.

GRAFICO N. 1.2: distribuzioni condizionate delle variabili esplicative qualitative rispetto alla fascia di prezzo



L'ispezione del GRAFICO N. 1.2, riferito alle caratteristiche di tipo qualitativo, suggerisce che:

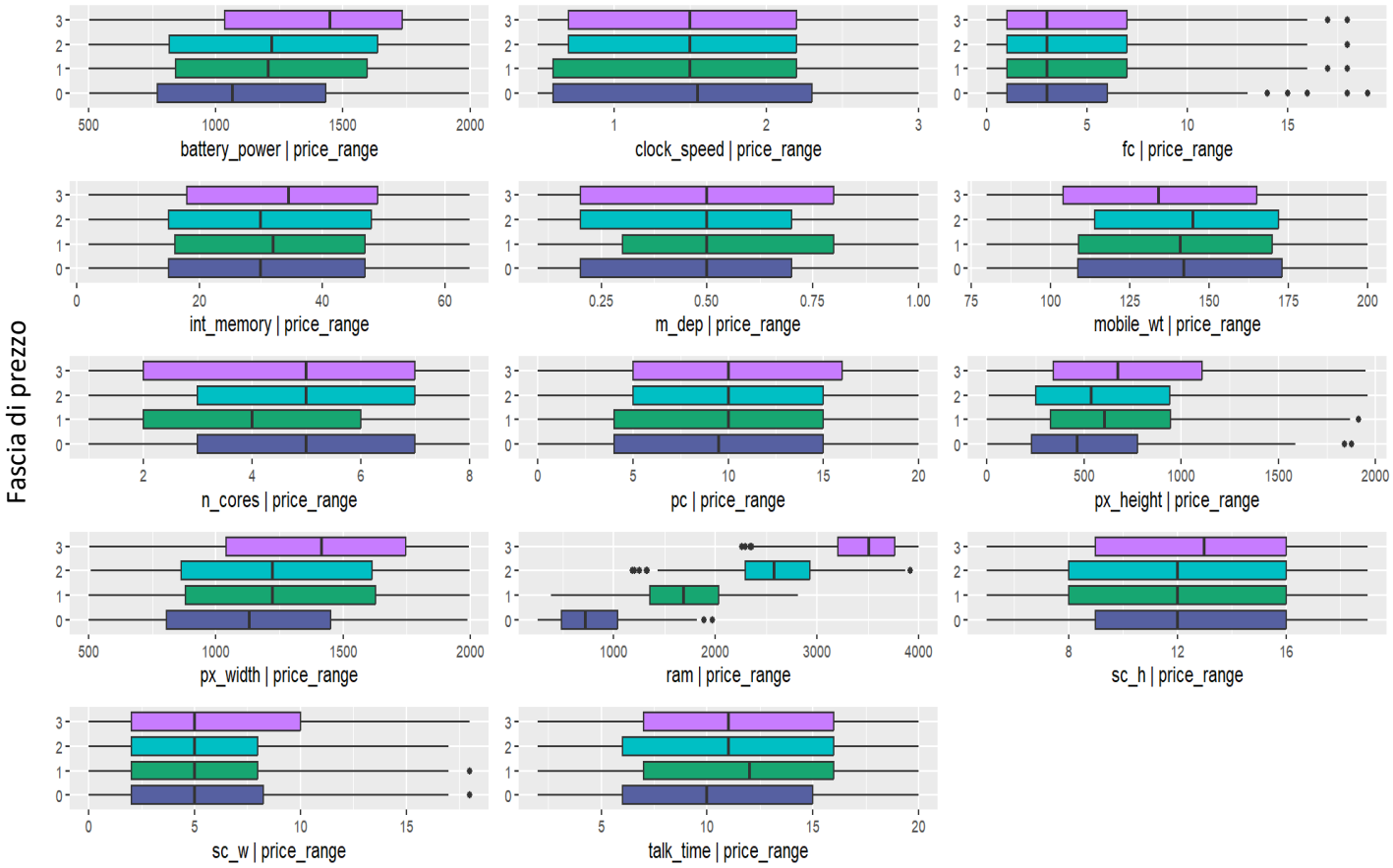
- condizionatamente alla fascia di prezzo le distribuzioni risultano piuttosto bilanciate: ad esempio, considerando la fascia di prezzo più bassa (0), si avrà uno stesso numero di cellulari dotati o meno di bluetooth o di wi-fi. L'unica eccezione è costituita dalla variabile three_g, secondo la quale la maggioranza dei telefoni cellulare, anche condizionatamente alla fascia di prezzo presenta il 3G;
- è possibile ipotizzare che le caratteristiche in esame non influenzano la fascia di prezzo in quanto le distribuzioni condizionate di una stessa esplicativa sono pressoché uguali;
- le variabili esplicative di natura qualitativa non presentano alcun potere discriminante di rilievo. Sapendo ad esempio che un telefono cellulare possiede 4G, bluetooth e touchscreen non sarà possibile prevederne la fascia di prezzo di appartenenza.

L'ispezione del GRAFICO N. 1.3, riferito alle caratteristiche di tipo quantitativo, suggerisce invece che:

- L'unica variabile ad avere un potere discriminante particolarmente rilevante è RAM, che misura in Mb la capacità della RAM posseduta dal telefono cellulare: una volta conosciuto tale valore si suppone che sarà possibile prevedere con sufficiente accuratezza la fascia di prezzo di appartenenza del cellulare;
- le variabili px_width, px_height e battery_power presentano un discreto potere discriminante;
- la fascia di prezzo di appartenenza di un cellulare aumenta proporzionalmente con l'aumentare della capacità della RAM posseduta.

E' possibile assumere che, nell'adattare una procedura di classificazione, le variabili individuate in questa fase vengano utilizzate per discriminare le osservazioni: RAM, avendo un potere discriminante evidentemente maggiore rispetto alle altre, sarà la variabile più rilevante, seguita plausibilmente da px_width, px_height e battery_power.

GRAFICO N. 1.3: distribuzioni condizionate delle variabili esplicative quantitative rispetto alla fascia di prezzo



2. Analisi di classificazione: DECISION TREE SINGOLO

Per raggiungere l'obiettivo preposto e ottenere un classificatore delle osservazioni il più efficace possibile è necessario adottare le tecniche tipiche di un'analisi di classificazione statistica.
La procedura dell'**albero di classificazione** permette di ottenere un classificatore che, tramite una serie di domande, effettua progressivamente delle partizioni dell'insieme iniziale di osservazioni fino a giungere ad insiemi di osservazioni "pure", appartenenti ad un'unica fascia di prezzo.
Tale classificatore viene solitamente "addestrato" tramite delle osservazioni di training e successivamente valutato tramite osservazioni di test, ossia tramite dei telefoni cellulare di cui si conosce la fascia di prezzo di appartenenza, confrontando quella predetta con quella effettiva.

L'insieme iniziale delle osservazioni è stato suddiviso in maniera casuale in:

- TRAINING SET, contenente il 75% delle osservazioni;
- TEST (o VALIDATION) SET, contenente il rimanente 25% delle osservazioni.

TABELLA N. 2.1: distribuzione di price_range nel training, nel test set e nel dataset complessivo

price_range	TRAINING SET		TEST SET		DATASET COMPLESSIVO	
0	366	**24,4 %**	134	**26,8 %**	500	25,0 %
1	371	**24,7 %**	129	**25,8 %**	500	25,0 %
2	374	**24,9 %**	126	**25,2 %**	500	25,0 %
3	389	**26,0 %**	111	**22,2 %**	500	25,0 %
TOTALE	1.500	100 %	500	100 %	2.000	100 %

Dopo aver verificato che sia il training set sia il test set fossero bilanciati rispetto alla variabile di risposta (TABELLA N. 2.1), ossia che fosse presente in essi una stessa percentuale di osservazioni appartenenti ad ogni fascia di prezzo, è stato costruito dapprima un albero di classificazione di dimensioni elevate per poi individuare, tramite potatura, un albero ottimo più semplice, di dimensioni inferiori, ma comunque affidabile.

Dato un albero di dimensioni elevate, la **potatura** (o **pruning**) è una tecnica che permette di rimuovere da esso i soli split che generano il minore decremento nel tasso di errata classificazione, ossia i soli rami la cui presenza, a confronto con gli altri, comporta il miglioramento inferiore nella capacità dell'albero di classificare in maniera accurata le osservazioni.

UNA VOLTA LO STANDARD ERROR

Una delle principali tecniche per procedere alla potatura di un albero di dimensioni elevate consiste nel considerare come candidati alberi ottimi tutti quelli aventi una stima per cross validation del tasso di errata classificazione inferiore ad una soglia individuata come:

SOGLIA = MINORE STIMA CV DEL TASSO DI ERRATA CLASSIFICAZIONE + STANDARD ERROR CORRISPONDENTE

La scelta dell'albero ottimo assoluto tra gli ottimi "candidati" così individuati verrà effettuata considerando il più semplice, ossia quello di dimensioni inferiori.

Il GRAFICO N. 2.1 permette di visualizzare l'andamento della stima per cross-validation del tasso di errata classificazione relativo (X-val relative error), ossia da moltiplicare per quello alla radice, al variare della grandezza dell'albero in termini di numero di foglie (size of tree) e del peso dato alla complessità dell'albero stesso (cp).

L'**albero ottimo** viene individuato considerando il più semplice tra quelli aventi valore del tasso di errata classificazione al di sotto della soglia determinata tramite il metodo di una volta lo standard error
(zoom del GRAFICO N. 2.1).
La potatura si effettua considerando il valore di cp in corrispondenza di questo albero.

GRAFICO N. 2.1: tasso di errata classificazione in funzione della complessità dell'albero

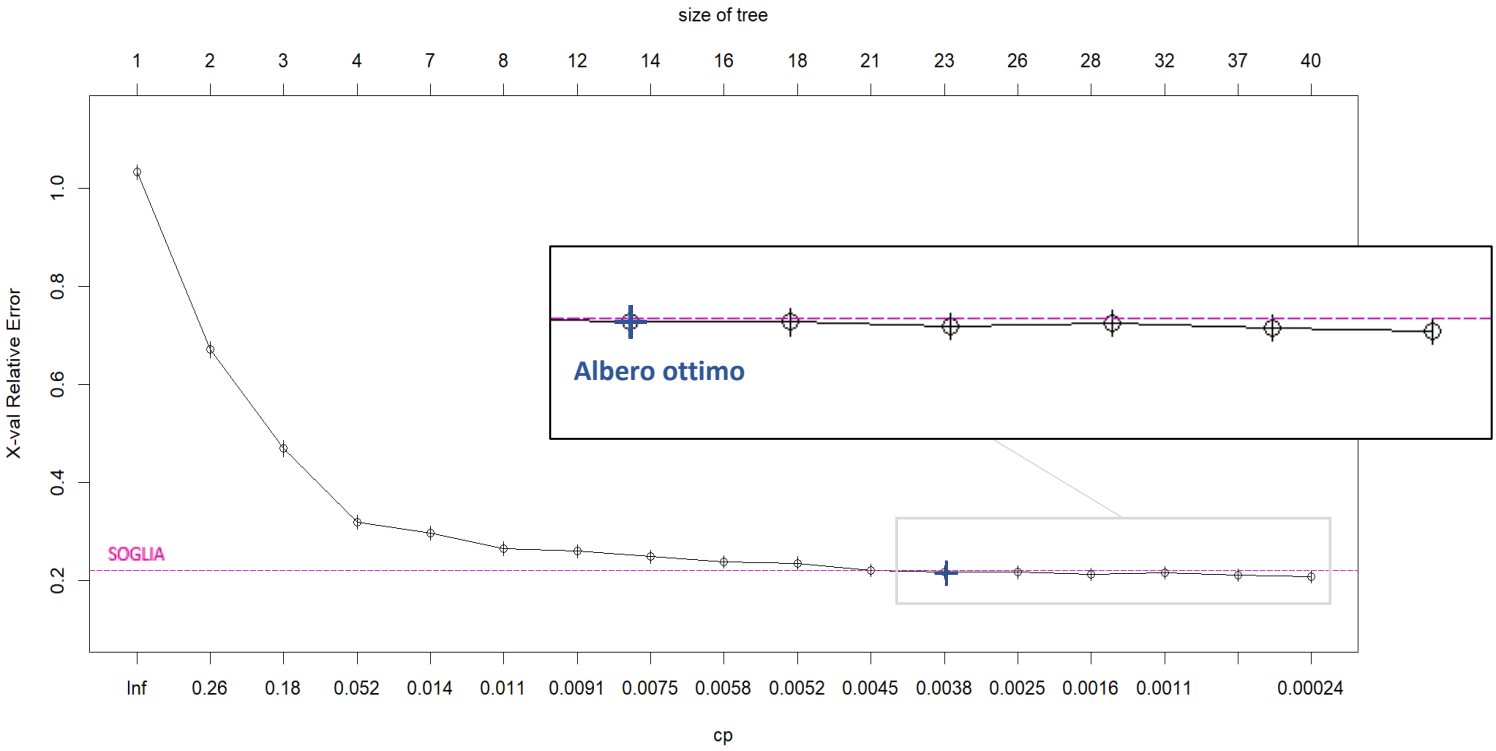


TABELLA N. 2.2: caratteristiche dell'albero ottimo

CARATTERISTICHE DELL'ALBERO OTTIMO POTATO	
Numero di split	22
Numero di foglie	23
Peso dato alla complessità dell'albero	0,0036
Stima per risostituzione del tasso di errata classificazione	0,10
Stima per cross validation del tasso di errata classificazione	0,16

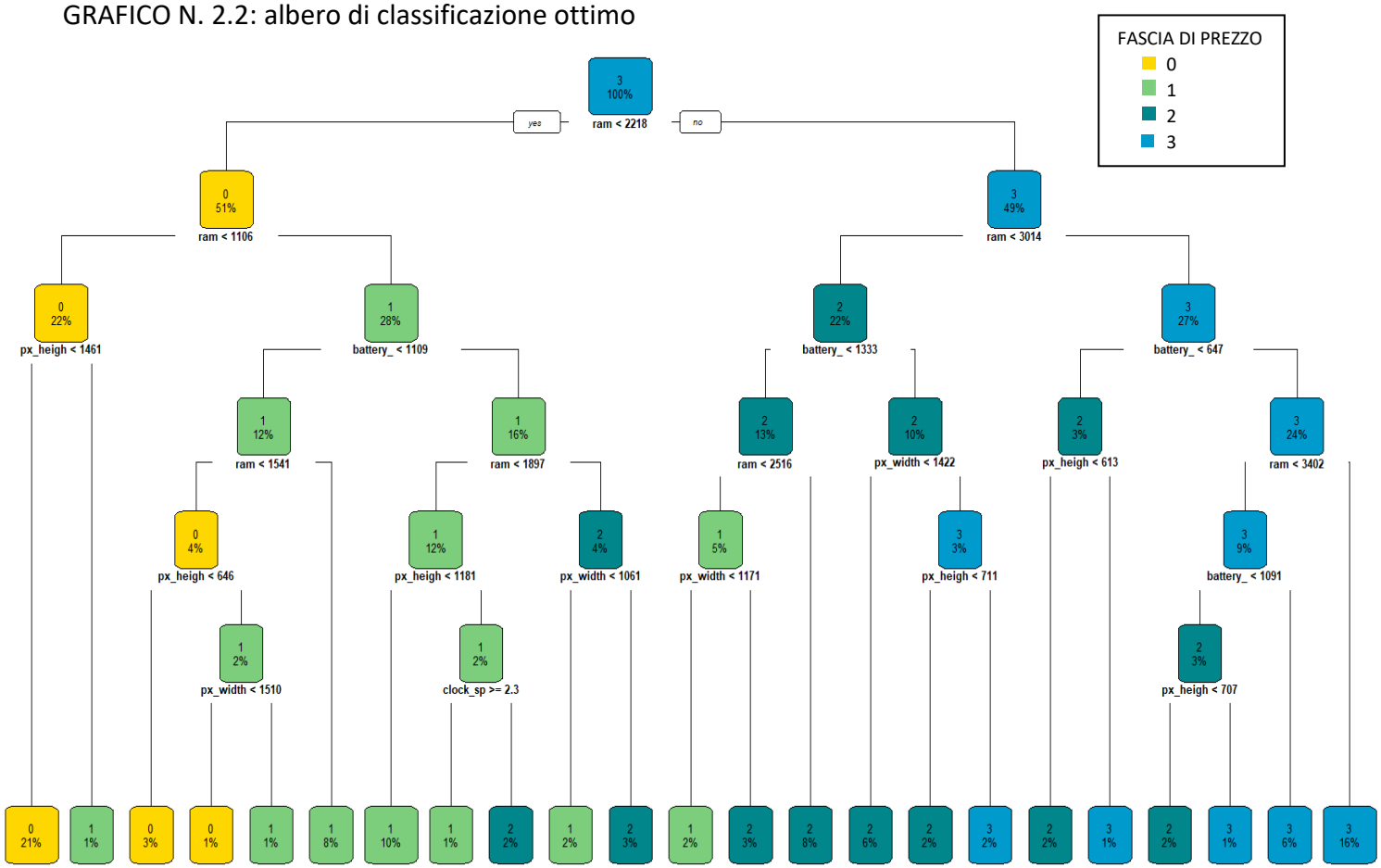
Un albero decisionale può essere rappresentato graficamente tramite un **grafo**, ovvero una struttura ad albero composta da:

- nodi, ossia i gruppi di osservazioni da suddividere;
- rami, ossia le condizioni che permettono di ripartire il nodo d'origine;
- foglie, ossia i nodi finali costituiti da unità il più possibile omogenee rispetto alla variabile d'interesse.

Il seguente grafico (GRAFICO N. 2.2) costituisce una rappresentazione dell'albero ottimo individuato tramite potatura. Ciascun nodo d'origine viene ripartito in maniera binaria secondo una determinata regola riferita ai valori di una variabile esplicativa.
All'interno di ciascun nodo sono riportate informazioni riferite a:

- fascia di prezzo, ossia la scelta di classificazione effettuata dall'albero ed evidenziata sia dal colore del nodo sia dalla modalità della variabile stessa;
- percentuale di osservazioni dell'insieme iniziale contenute nel nodo.

GRAFICO N. 2.2: albero di classificazione ottimo



3. Analisi di classificazione: RANDOM FOREST

La random forest appartiene alla famiglia di tecniche di analisi di classificazione degli **ensemble methods** e, in genere, permette di superare alcuni limiti dell'albero di classificazione singolo.
È in grado di ridurre la variabilità della classificazione fornita da un unico albero in quanto, in questo caso, l'esito della classificazione viene ottenuto combinando i risultati provenienti da più alberi.
Tali alberi vengono costruiti su gruppi **bootstrap** del dataset di partenza, ottenuti campionando con ripetizione le osservazioni iniziali.
Ad ogni split, la scelta della variabile da utilizzare non viene effettuata sull'insieme totale delle variabili esplicative presenti, ma su un sottoinsieme casuale di esse, in modo tale da evidenziare il potere discriminante di variabili che, altrimenti, potrebbero venire mascherate dall'effetto altrui.

OTTIMIZZARE UNA PROCEDURA DI RANDOM FOREST

L'ottimizzazione di una random forest prevede la scelta del valore ottimo dei seguenti parametri:

- numero di variabili esplicative da campionare ad ogni split, per ogni albero;
- numero di alberi da creare con campioni bootstrap.

La scelta della random forest ottima avviene confrontando la bontà delle prestazioni delle foreste al variare di tali parametri, quantificata considerando l'errore di classificazione calcolato sulle osservazioni appartenenti al test set, e tenendo inoltre conto del fatto che, all'aumentare del valore di tali parametri, aumenta lo sforzo computazionale da sostenere nell'implementazione della procedura.

Sono state messe a confronto le prestazioni di random forest aventi:

- numero di variabili esplicative da campionare ad ogni split variabile da 1 a 20;
- numero di alberi variabile da 1 a 800.

NUMERO DI VARIABILI ESPLICATIVE

Dopo aver implementato le 16.000 foreste aventi combinazioni diverse dei parametri citati, la scelta finale riguardo il **valore ottimo del numero di variabili esplicative** ha coinvolto le foreste rappresentate nel GRAFICO N. 3.1 e aventi un valore del parametro m (numero di variabili) pari a:

- **m = 5**, valore del parametro solitamente adottato e corrispondente all'incirca alla radice quadrata del numero di variabili esplicative di cui si dispone;
- **m = 13, 14, 15, 16**, valore del parametro nelle random forest aventi un tasso di errore inferiore in corrispondenza della dimensione massima, pari a 800 alberi;
- **m = 20**, valore del parametro corrispondente alla procedura di bagging, in cui non si effettua un campionamento delle variabili esplicative poiché, ad ogni split, esse vengono considerate per intero.

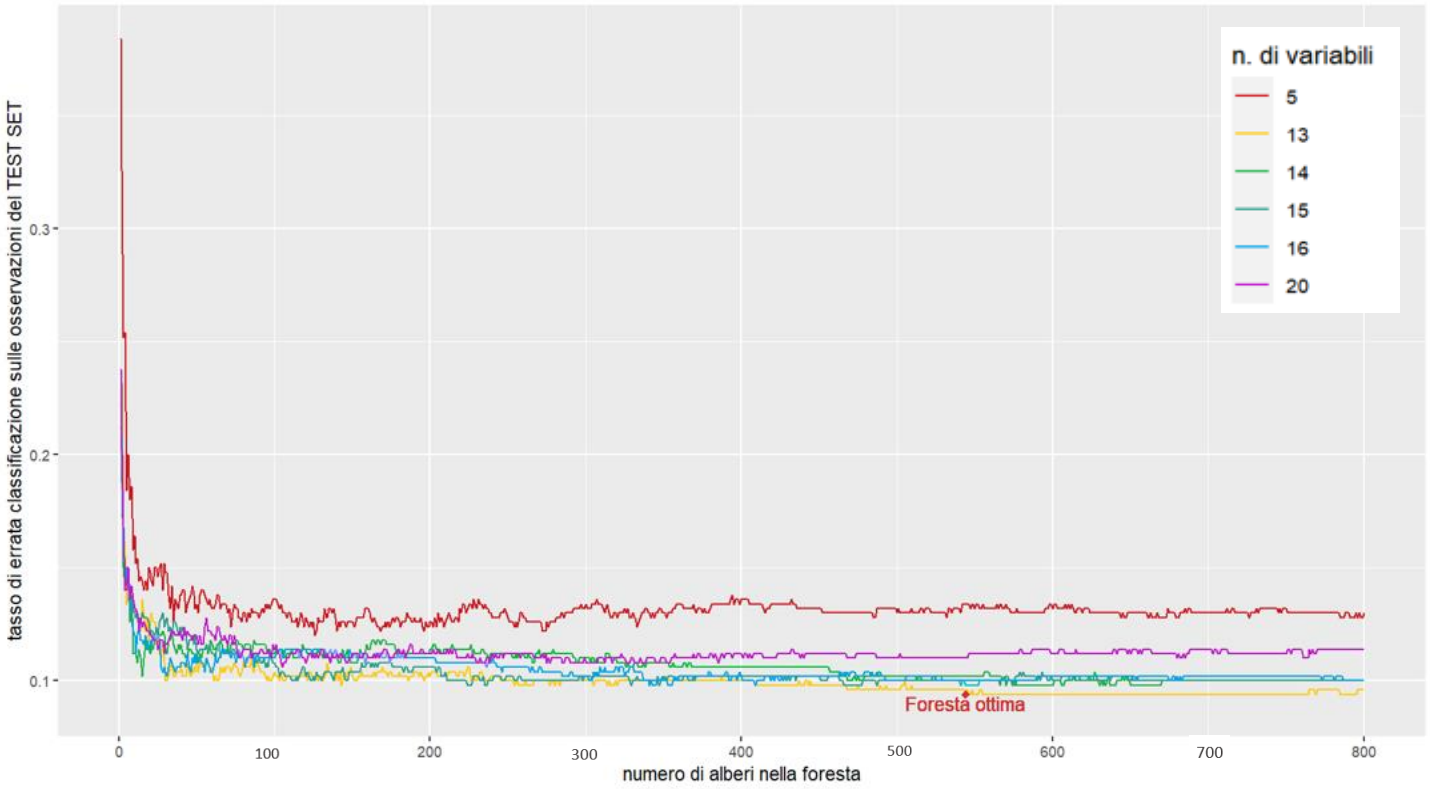
NUMERO DI ALBERI

Il **valore ottimo del numero di alberi** contenuti nella random forest, invece, è stato determinato considerando la dimensione minima a partire dalla quale le prestazioni delle random forest aventi numero di variabili ottimo presentano una performance stabile:
il livello di precisione raggiunto da una foresta avente 800 alberi potrà dunque essere all'incirca raggiunto anche da una foresta di dimensioni inferiori, con un evidente risparmio in termini di sforzo computazionale.

Il GRAFICO N. 3.1 riportato di seguito mostra l'andamento del tasso di errata classificazione calcolato sulle osservazioni contenute nel test set, al variare del numero di alberi presenti nella random forest e del numero di variabili esplicative utilizzate.

Permette di individuare il valore dei parametri che determina risultati migliori.

GRAFICO N. 3.1: confronto delle prestazioni delle random forest



La visione del GRAFICO N. 3.1 suggerisce che:

- sia il bagging sia il valore del parametro m solitamente utilizzato portano a dei risultati meno attendibili;
- a partire da una dimensione pari a 544, le prestazioni delle random forest i cui alberi sono stati realizzati campionando ad ogni split un insieme di 13 variabili esplicative, si stabilizzano.
  Avendo raggiunto una condizione di stabilità, l'inclusione nella foresta di un numero superiore di alberi non comporterà alcun miglioramento rilevante nella bontà della classificazione effettuata.

TABELLA N. 3.1: caratteristiche della random forest ottimizzata

CARATTERISTICHE DELLA RANDOM FOREST OTTIMA	
Numero di variabili esplicative	13
Numero di alberi contenuti nella forest	544

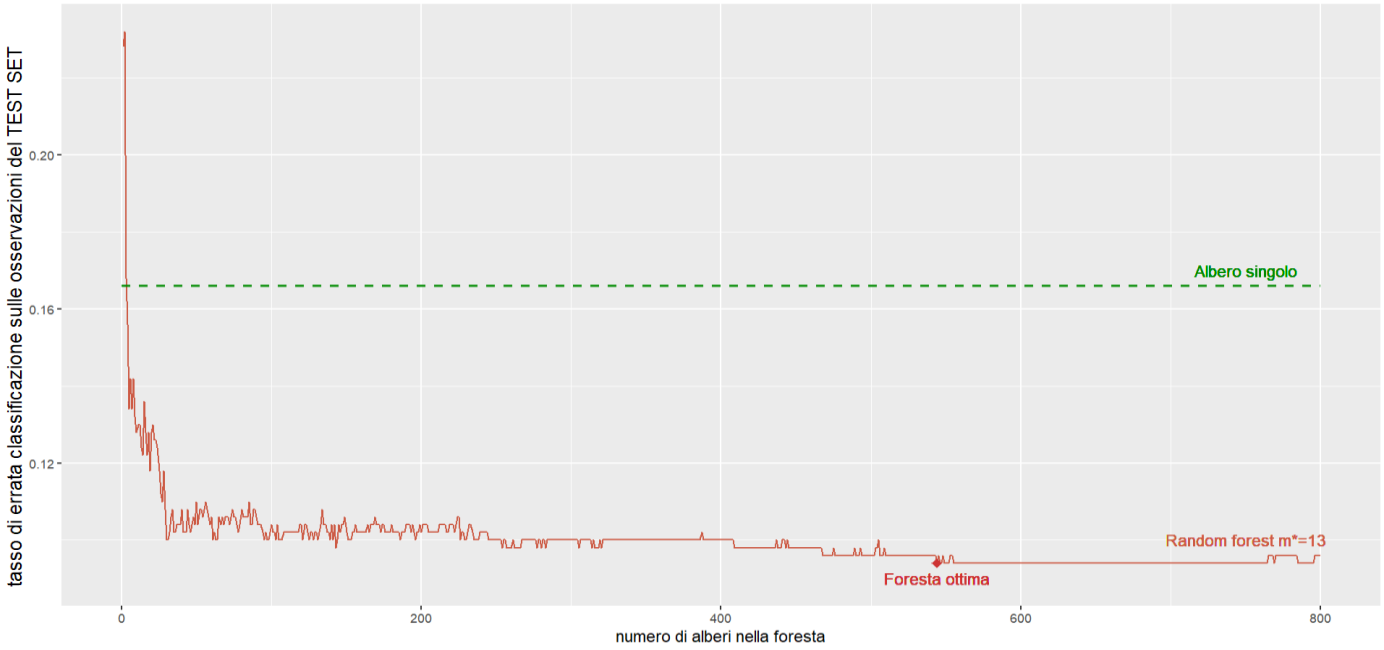
4. Confronto tra prestazioni: DECISION TREE SINGOLO e RANDOM FOREST

La scelta della procedura da utilizzare, tra il singolo albero di classificazione e la random forest precedentemente implementati, viene effettuata valutandone le prestazioni, ossia la loro capacità di fornire una classificazione il più possibile accurata e priva di errore.

Tale confronto è stato effettuato tramite:

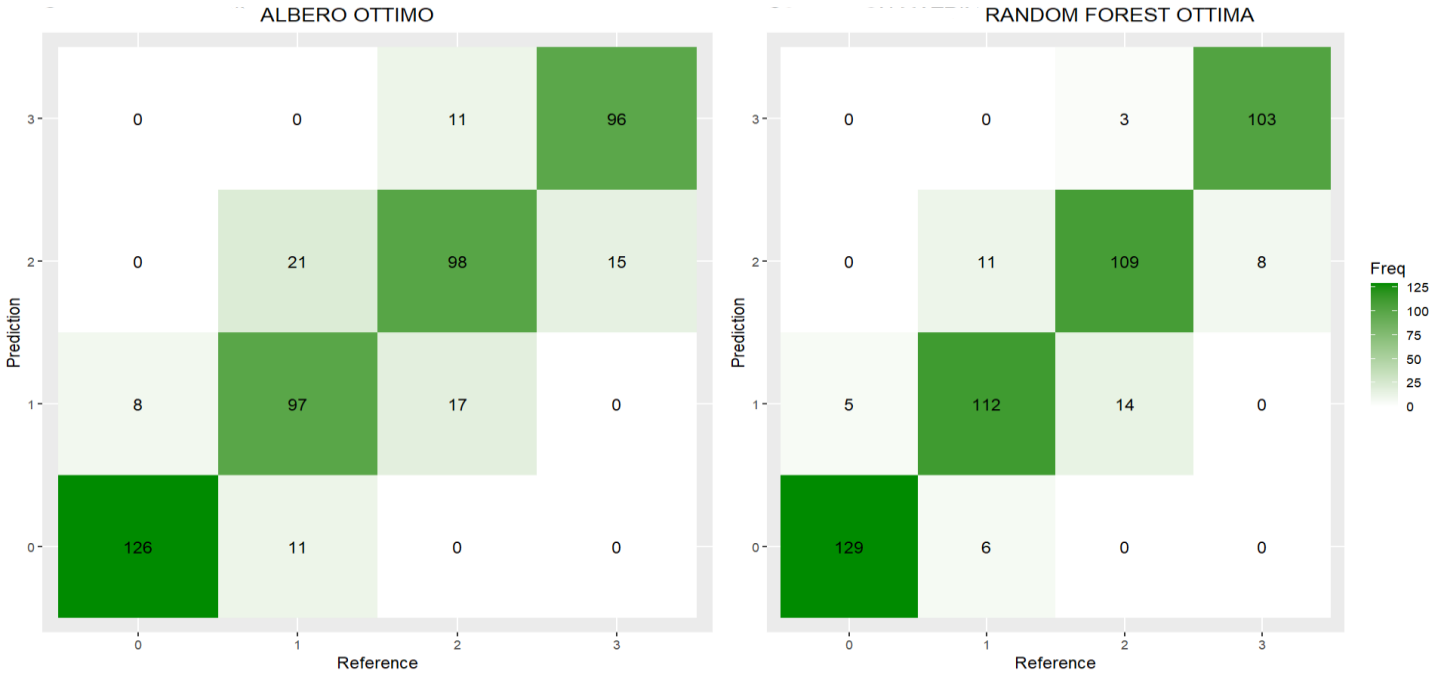
- rappresentazione grafica (GRAFICO N. 4.1);
- confusion matrix (GRAFICO N. 4.2);
- valore delle statistiche che permettono di valutare la bontà di una procedura (GRAFICI N. 4.3 e 4.4).

GRAFICO N. 4.1: prestazioni random forest e albero singolo ottimizzati



Il GRAFICO N. 4.1 mette a confronto il tasso di errata classificazione calcolato sul test set con riferimento sia al singolo albero ottimo di classificazione sia alla random forest. Utilizzare l'ensemble method della random forest, in questo caso, porta ad una riduzione del tasso di errata classificazione (performance migliore).

GRAFICO N. 4.2: confusion matrix albero singolo e random forest ottimizzati

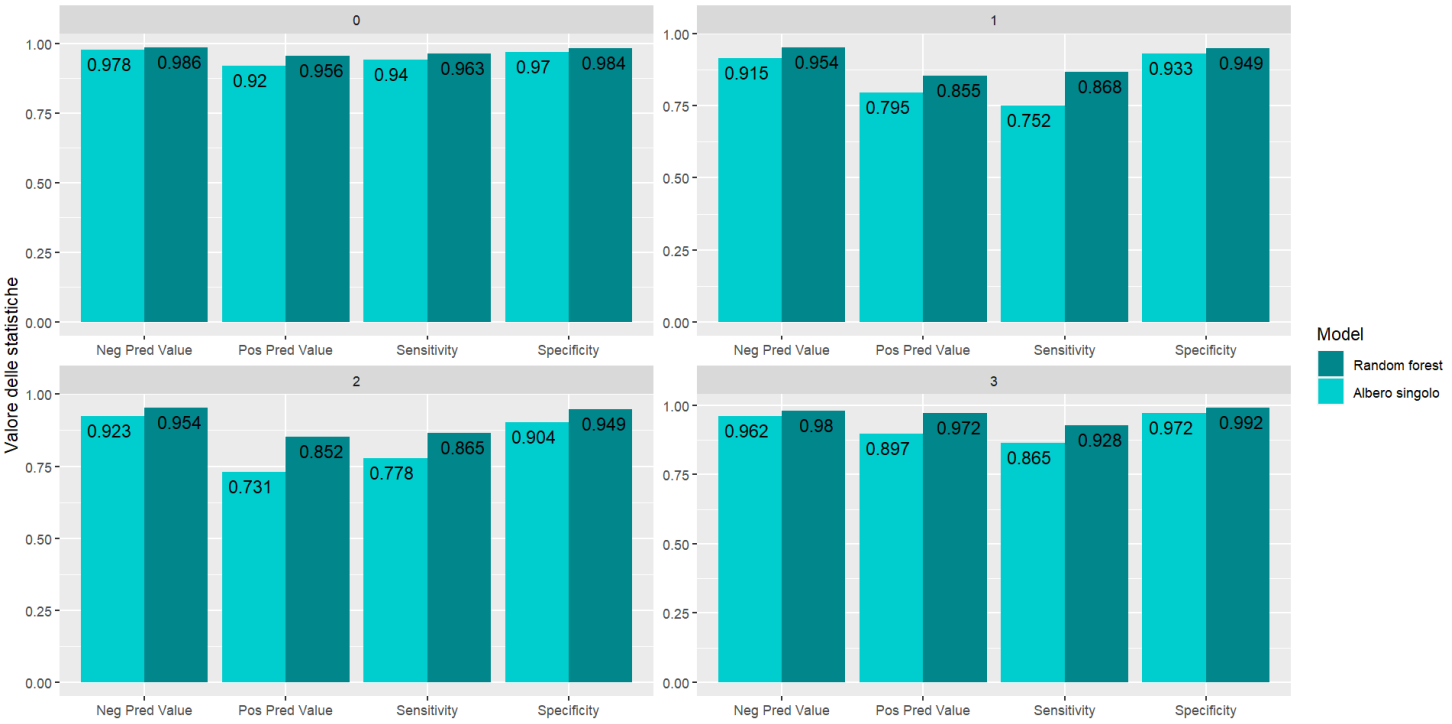


Le confusion matrix riportate nel GRAFICO N. 4.2 presentano:

- orizzontalmente la vera fascia di prezzo dei telefoni appartenenti al test set;
- verticalmente la fascia di prezzo prevista dalla procedura di classificazione adottata.

Le frequenze all'interno di ciascuna cella sono state rappresentate sia tramite numeri sia tramite diverse sfumature di verde, d'intensità direttamente proporzionale alla frequenza stessa di riferimento.
La **diagonale secondaria** riporta il numero di telefoni cellulare appartenenti al test set e correttamente classificati dalla procedura.
È possibile notare che la diagonale secondaria della confusion matrix della random forest ottima presenta tonalità più intense di verde rispetto al singolo albero di classificazione, in quant5o, tramite essa, un maggior numero di telefoni cellulare sono stati correttamente classificati.

GRAFICO N. 4.3: statistiche per valutare la bontà di albero singolo e random forest ottimizzati

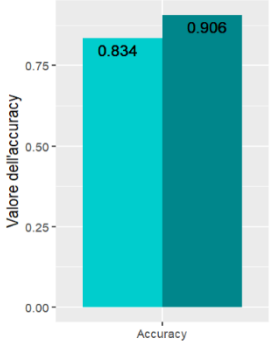


Per ciascuna fascia di prezzo i-esima sono state calcolate e riportate nel GRAFICO N. 4.3, le seguenti statistiche:

- **Specificity**, riferita alla capacità della procedura di classificare un cellulare non di fascia i-esima, in una fascia di prezzo diversa dalla i-esima;
- **Sensibility**, riferita alla capacità della procedura di classificare un cellulare di fascia i-esima nella stessa fascia di prezzo i-esima;
- **Positive predictive value**, riferito alla probabilità di avere cellulari veramente appartenenti alla fascia i-esima tra tutti quelli assegnati dalla procedura alla fascia i-esima;
- **Negative predictive value**, riferito alla probabilità di avere cellulari veramente non appartenenti alla fascia i-esima tra tutti quelli non assegnati dalla procedura alla fascia i-esima.

GRAFICO N. 4.4: accuracy

Con riferimento modello nel suo complesso è stata invece calcolata l'**accuracy** (GRAFICO N. 4.4), una misura della capacità della procedura di individuare la fascia di prezzo corretta dei telefoni cellulare. E' possibile notare che il valore di tali statistiche è sempre maggiore nel caso di Random forest e che, pertanto, questa fornisce risultati migliori rispetto al singolo albero di classificazione.

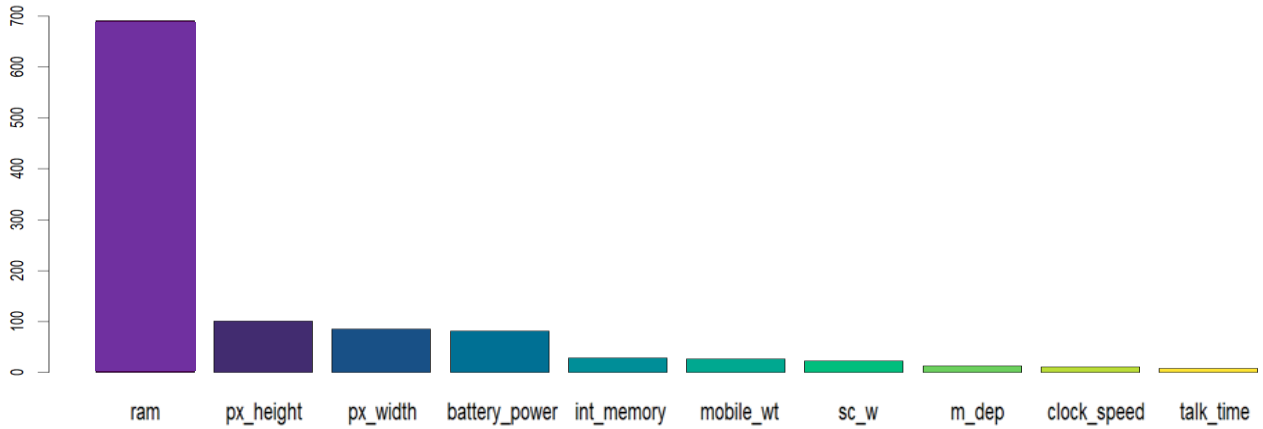


5. Confronto tra le variabili utilizzate:

DECISION TREE SINGOLO, RANDOM FOREST E ANALISI ESPLORATIVA

L'importanza di ciascuna delle variabili esplicative utilizzate nell'albero di classificazione singolo ottimo è stata determinata tenendo conto sia del numero di volte in cui una variabile appare come primary split sia del numero di volte in cui appare come surrogate split (split alternativi da utilizzare nel caso in cui l'unità presenti un NA). Come previsto in fase esplorativa, tra le variabili più importanti ritroviamo, tra le altre, RAM, px_height, px_width e battery_power (non tutte sono presenti nell'albero perché non tutte sono primary split).

GRAFICO N. 5.1: dieci variabili più importanti nell'albero di classificazione ottimo



Il GRAFICO N. 5.2 mostra l'importanza delle variabili esplicative determinata secondo due diverse misure:

- la 1° tiene conto della variazione nell'indice di Gini dovuto all'impiego o meno di una certa variabile;
- la 2° tiene conto della variazione nell'accuracy dovuta ad una permutazione delle sue modalità.

L'ispezione di tale grafico suggerisce che, anche in questo caso, i risultati dell'analisi esplorativa sono in linea con quelli dell'analisi di classificazione effettuata tramite random forest. Tra le variabili esplicative più importanti, più discriminanti sono risultate essere le stesse ipotizzate inizialmente tramite rappresentazione grafica esplorativa: RAM, battery_power, px_width, px_height.

GRAFICO N. 5.2: dieci variabili più importanti nella random forest ottima