Data-Mining 3cfu

House Price Prediction

Baldanza Matteo[1,2]

[1]Laurea Triennale in Scienze Statistiche Ed Economiche
[2]Studente Magistrale in Scienze Statistiche e Data Science

January 7, 2022

Abstract

L'obiettivo dello studio è stata la minimizzazione del MAE (mean absolute error) attraverso l'utilizzo di modelli statistici e di machine learning. Il lavoro si è concentrato prevalentemente nell'analisi esplorativa e nel feature engineering mentre il tuning dei parametri non è stato molto approfondito sia per questioni computazionali sia per miglioramenti non sostanziali. Si riportano quindi in questo articolo le osservazioni, le intuizioni, le idee e i risultati ottenuti durante lo studio del dataset in questione.
Si ottiene con la peggior previsione possibile (modello nullo) un MAE pari a 0.179 mentre un modello completo senza alcuna analisi porta ad un MAE di 0.091.
Il modello su cui si è riposto il maggior interesse è il modello lineare che con iterazioni e splines porta delle performance molto buone (MAE = 0.053) e all'altezza di modelli molto più complessi (Xgboost, MAE = 0.050). Si è riusciti infine ad ottenere un MAE pari a 0.0494 tramite ensemble tra xgboost e modello lineare.

1 Introduzione

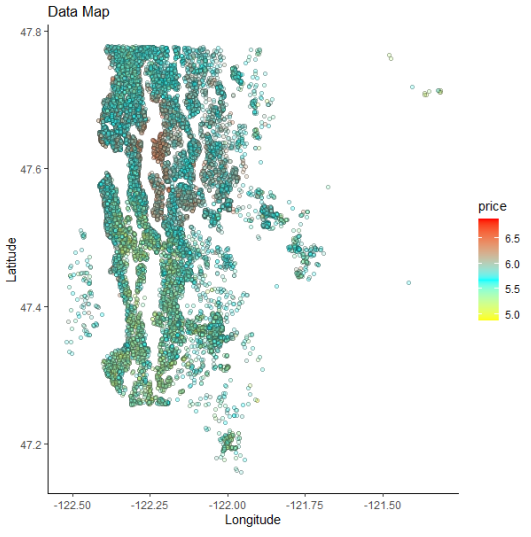
Il dataset in questione è costituito da 21613 osservazioni e 19 features. La variabile dipendente è "price" che per un evidente assimetria è stata trasformata mediante trasformazione logaritmica con base pari a 10. Si dispone di un training set composto da 17293 unità statistiche e di un test set composto dalle restanti 4320. Nel test a disposizione non è presente la variabile risposta e lo scopo del lavoro è quello di trovare la miglior previsione possibile seguendo come metrica il MAE. A tal proposito l'idea iniziale è stata quella di suddividere il training in una parte di train e in una di test che onde evitare confusioni chiameremo, senza mancanza di formalità, validation. E' stato utilizzato come metodo un campionamento stratificato sulla variabile risposta price. Il validation *non sarà* utilizzato per alcuna considerazione, è come se non fosse mai stato osservato e verrà usato solo per le valutazioni finali delle performance dei modelli.
Si riassumono per maggiore chiarezza i valori di mae di partenza.

	LM NULLO	**LM PIENO**
MAE	0.0179	0.091

2 Analisi Esplorativa

Le variabili a disposizione sono le segunti: *"date sold, bedrooms, bathrooms, sqft living, sqft lot, floors, waterfront, view, condition, sqft above, sqft basement, yr built, year renovated, zip code, latitude, latitude, longitude, nn sqft living, nn sqft lot"*.

Per la valutazione della componente spaziale si propone il seguente grafico:



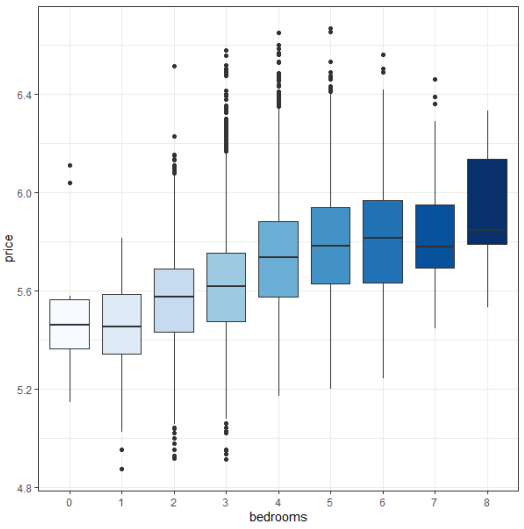
Si osserva in maniera evidente che le case con prezzi alti sono collocate in specifiche regioni spaziali (colore rosso nella mappa). E' interessante notare come ci siano degli spazi vuoti nella mappa, molto probabilmente attribuili alla presenza del mare (è presente infatti la variabile waterfront). Questa piccola scoperta ha posto subito l'attenzione sulla variabile zip code. E' infatti normale aspettarsi che quest'ultima e la combinazione latitudine/longitudine diano lo stesso tipo di informazione a livello spaziale. Sebben il pensiero iniziale fosse questo i dati hanno smentito ciò, suggerendo un utilizzo di entrambe le features. Si è infatti notato che all'interno di ogni paese il prezzo cambia (ovvio riscontro reale, ogni paese/città ha i suoi prezzi al metro quadro). Il problema è che lo zip code da solo non basta a descrivere i vari prezzi. Ci possono essere zone dello stesso paese con prezzi diversi, basti pensare al quartiere di Milano Bicocca, seppur il cap sia lo stesso le residenze vicine all'università hanno prezzi maggiori. Si è cosi optato per un mantenimento di entrambe.

Per quanto riguarda la variabile bedrooms si propone una tabella riassuntiva di quanto trovato nel training.

0	1	2	3	4	5	6
11	131	1779	6285	4390	1028	166

7	8	9	10	11	33
28	8	4	1	1	1

È presente una casa con 33 bedrooms. Studiando bene questo valore si è scoperto che il prezzo attributo non giustifica il valore delle variabili indipendenti. Valutando il prezzo delle case vicine si è notato come il 33 sia molto presumibilmente un errore di imputazione. Con infatti un valore pari a 3 il prezzo si allinea alla media e si evita cosi di perdere un'informazione (si sarebbe potuta eliminare). Per quanto riguarda invece il numero di unità basse per le case con stanze maggiori di 8 si è deciso di sostituire tutti i valori maggiori di 9 con 8. Dai boxplot condizionati emerge che i prezzi hanno più o meno le stesse mediane (preferibile come metrica rispetto alla media per i numerosi outliers). Un'ultima osservazione viene posta sulle case con zero camere. Il tempo speso per indagare la questione ha portato alla scelta di mantenere tali osservazione che possono essere spiegate da negozi,capannoni o altre strutture che sono state e/o che erano in vendita. Ci sono poi anche un numero di "case" con zero bagni e zero stanze contemporaneamente, ma non è stata scelta l'esclusione di queste rilevazioni anche perchè si è notato nello stesso test la presenza di situazioni analoghe. Queste modifiche/decisioni hanno portato ad un, seppur piccolo, miglioramento nel fit dei modelli (si mostra nell'immagine qui riportata come il prezzo cambi rispetto al numero delle stanze)



Per quanto riguarda la variabile bahtrooms non si propone una tabella riassuntiva per via dei troppi valori assunti. In questo caso come per la situazione precedente è stato studiato il boxplot condizionato al prezzo. Cosi facendo si è notato che i valori con decimali <0.5 hanno un prezzo mediano simile agli interi più piccoli. Si è deciso cosi di applicare a loro un'approssimazione per difetto. Vale esattemente l'opposto per i valori con decimali maggiori/uguali a 0.5. Si sono infatti approssimati questo valori per eccesso.

Per quanto riguarda i floors si propone una tabella contenente anche i valori del prezzo (mediana).

1	1.5	2	2.5	3	3.5
6831	1236	5276	98	386	6
5.59	5.72	5.74	5.89	5.69	5.72

La decisione su cosa fare è stata più volte cambiata nel corso del lavoro ma di questo se ne parlerà nella prossima sezione.

La variabile sqft living che esprime lo spazio abitabile di casa in metri quadri ha un'importante effetto nelle performance dei modelli. E' evidente infatti pensare che al crescere dei metri quadri il prezzo di una casa salga rispettivamente. I dati in effetti suggeriscono proprio una forte correlazione positiva. Ci sono due osservazioni che hanno valori di living elevatissimi, superiori ai $10000m^2$. Si è scelto così di togliere dal dataset tali unità statistiche.

Sqft lot che esprime lo spazio totale della casa in vendita è stato mantenuto con alcune modi-

fiche che verranno trattate fra poco. Anche in questo caso come nel precedente sono state eliminate le due osservazioni più estreme in quanto peggiorative nelle performance dei modelli.

Sono state esplorate anche tutte le altre variabili. Per quanto riguarda la variabile waterfront, il cambio del prezzo è significato al variare del valore da lei assunto (assume valore 1 se la casa è posizionare davanti al mare). Lo stesso vale per view (qualita della vista) e condition (condizione casa).

Si chiude questa prima parte esplorativa riportando un grosso problema presente nel dataset. La variabile sqft living altro non che è la somma di sqft above e di sqft basemente, rispettivamente spazio in metri quadri di casa abitabili sopra e sotto il piano terra. Il che rappresenta ovviamente un grosso problema in fase di utilizzo dei modelli.

3 Feature engineering

Il processo di feature engineering è stato eseguito con tanti cicli. Tante volte una partenza da un'idea potenzialmente buona è stata scartata dai modelli e si è dovuti tornare al punto di partenza per delle rivalutazioni. Si cercherà quindi di spiegare al meglio i vari approcci utilizzati.

Prima di partire dal problema appena presentato è bene spiegare come è stata valutata questa fase. Alcune trasformazioni di variabili non sono state applicate a tutti i modelli che verranno presentati. Altre invece sono state fate a tutti. In particolare il modello di regressione lineare e i metodi basati sugli alberi non condividono, in tale lavoro, tutto lo stesso tipo di feature engineering. E' bene quindi suddividere in più paragrafi cosa è stato fatto.

3.1 Scelte condivise

Partiamo dal presentare le scelte che inequivocablmente hanno portato beneficio ad ogni modello utilizzato.

- Riprendiamo la variabile **floors**. Il primo approccio, subito scartato, è stato di non prendere alcuna decisione e mantenere la variabile in questo modo. Un'altra possibilià è quella di dividere i piani in una variabile categoriale composta da case ad un piano, case a 2.5 piani e poi tutte le restenti. Il motivo è che i prezzi cambiano in maniera significativa in queste tre macro categorie. Sebben l'idea fosse ottima il riscontro dei modelli e delle tecniche di valutazione delle performance hanno portato ad una scelta finale di una variabile a due categorie, case con un piano e poi tutte le restanti

- Per quanto riguarda la variabile **year renovated** si è pensato di trasformarla in questo modo: "year renovated-yr built". In questo modo le case che non sono state rinnovate avranno un valore pari a 0 mentre le altre avranno valori tanto più grandi quanto il tempo trascorso dalla loro ristrutturazione. Sebben all'inizio si fosse optato per una binaria con valore "yes" in caso di ristrotturazione e "no" in caso contrario si è notato che la performance peggiora

- E' stata creata una variabile di nome **garden** data dalla differenza tra sqft lot e sqft living. Se infatti la dimensione del lot è la dimensione totale dell'abitazione e la dimensione del living è la dimensione abitabile della casa, si è pensato che la differenza sia attribuile alla parte esterna, ovvero il giardino

- Sulla base delle informazioni a disposizione è stata creata una nuova variabile **"pr m qu"** indicante il prezzo medio al metro quadro per zip code (paese). Si può infatti presumere che una buona approssimazione di questo valore possa essere trovato facendo la media di tutte le case della stessa città rapportate ai metri quadri di casa vivibile.

3.2 Scelte Modelli Lineari

- Possiamo finalmente parlare del problema lasciato in disparte in precedenza ovvero della forte correlazione tra above, basement e sqft living. Sebben metodi come Random Forest e Xgboost non soffrano molto questa problematica la regressione linerare si. L'approccio banale sarebbe quello di eliminare le due variabili che formano il sqft living ma in questo modo si perderebbe tanta informazione. Si è scelto quindi di trasformare la variabile sqft basement in una variabile dicotomica con categorie, "yes" e "no", nel caso in cui rispettivamente il valore assunto sia maggiore e minore di zero. Il motivo è dettato dal fatto che essendo il basement la parte "sotteranea" di casa si pensa che un'abitazione con tale parte abbia più valore di una casa senza. Tale variabile ha ottenuto infatti un buon riscontro nei modelli

- E' stata creata una variabile **date sold diff** formata dalla differenza tra l'anno

di vendita della casa e l'anno della sua costruzione. Questa scelta motivata dal pensiero che più una casa è vecchia minore o maggiore (se magari storica) il suo valore potrebbe essere rispetto ad una nuova. Anche in questo caso i modelli hanno ottenuto un incremento notevole di performance

- Sono state trasformate come factor (categoriali) le seguenti variabili: "bedrooms, zip code,condition, yr built ,bathrooms,view,waterfront", alle quali si aggiungono ovviamente tutte quelle presentate in precedenza (floors, basement....)

3.3 Scelte modelli con Alberi

Si riportano ora le differenze con il modello lineare. Sono state provate le stesse sue modifiche nei seguenti algortimi ma nessun miglioramento è stato apportato.

- Per gestire le variabili sqft basement e sqft above sono state create rispettivamente due nuove variabili **perc cas** e **perc case 2** formate dalla percentuale del sqft living attribuile ad ognuna delle due feature. In questo modo si legano le informazioni dandone una nuova interpretazione e vengono rimosse completamente dal dataset le variabili base.

- Per quanto riguarda la variabile **zip code** sebben nel modello lineare sia considerata tutta come factor qui sono state messi come nulli gli zip code che si presentano con una frequenza minore dello 0.05

- La variabile **date sold** è stata convertita in numerico. Ad esempio la data "25/04/2015" diviene 20150425. Attribuire un peso anche alla date di vendita attribuisce un miglioramento alle previsioni. La serie storica delle date ha quindi valore statistico per questo tipo di modelli.

Al termine di entrambi i processi le variabili sqft living, sqft lot, nn sqft lot e nn sqft living sono state trasformate mediante scala logaritmica a base 10. Questo è stato fatto sia per rispettare la scala di trasformazione della risposta sia per un evidente miglioramento della distribuzione (si nota infatti una forte vicinanza alla normalità).

Un recipiente a parte è stato creato per i modelli Knn e Svm. Si sono infatti normalizzati i dati in quanto metodi basati sulle distanze. I metodi in questione non sono stati indagati maggiormente per via delle loro scarse performance e per questo non saranno nominati successivamente.

4 Scelta dei modelli/tuning

L'approccio è stato il medesimo per tutti i modelli tranne che per il modello lineare (non contiene iperparametri). Si è utilizzata una cross validation con 5 folds. Il numero è stato scelto seguendo ciò che la letteratura suggerisce [1]. Di solito è buona cosa utilizzarne 10, ma in questo caso per motivi computazionali si è scelto di optare per un valore pari a 5. I risultati non dovrebbero risentirne particolarmente. Il metodo in questione è stato utilizzato sia per valutare la scelta degli iperparametri sia per calcolare l'errore di generalizzazione (generalized error). Si ricorda che l'errore di generalizzazione viene calcolato con metriche come holdout, cross validation o bootstrap ed è una buona approssimazione dell'errore del modello sui dati futuri. Confrontando poi il valore ottenuto con l'errore empirico (previsione sui dati di train) si possono ottenere diverse informazioni. In caso di vicinanze dei due valori le stime risultano robuste e ci si attenderà con i nuovi dati un valore più o meno simile mentre potrebbe essere un campanello di allarme overfitting se sono distanti.
Si separano anche in questo caso i vari modelli.

4.1 Modello Lineare

Il featuring eninereening mostrato nello scorso paragrafo è stato utilizzato per la costruzione del nuovo modello di regressione utilizzato per costruire le previsioni. Si osserva che essendo la risposta normalmente distribuita e accorgendosi di una andamento lineare tra price e molte variabili, l'utilizzo di tale modello è divenuto quasi obbligatorio. Come già spiegato in precedenza più volte le trasformazioni delle variabili sono state cambiate ma quello che ne risulta alla fine, con le modifiche presentate in precedenza, è un miglioramento sostanziale.

Generalized Error	Empirical Errror
0.0588	0.0575

Si nota subito un miglioramento rispetto allo 0.09 del modello completo di partenza. I due valori sono molto vicini il che ci assicura buona stabilità nelle previsioni (verificheremo questo sul validation più avanti).
Ottenuto questo buon valore si è andati a costruire un modello lineare con la presenza di iterazione. Il motivo è spiegato tramite il semplice grafico proposto qui sotto.

La pendenza della retta è completamente differente il che suggerisce la presenza di iterazione significativa tra la variabile sqft living e waterfront (0=No vista mare, 1=Si vista mare).

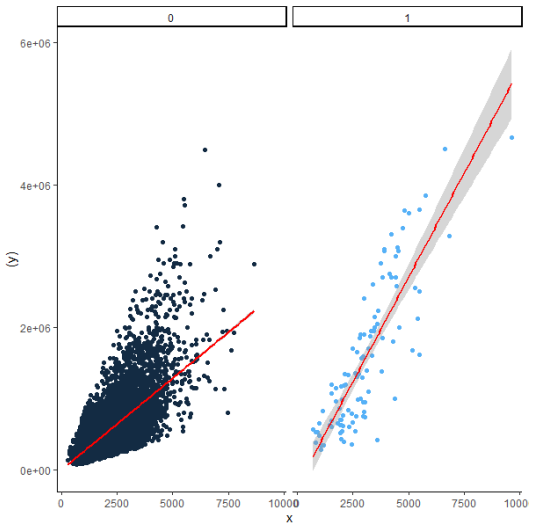


Figure 1: Iterazione sqft living con waterfront

Non si propongo tutti i grafici effettuati per ovvie questioni ma si è notata la presenza di moltissime variabili con questa particolarità. Si procede quindi mostrando i risultati di questo "nuovo" modello.

Generalized Error	Empirical Errror
0.0578	0.0544

In questo caso si nota un miglioramento piccolissimo dell'errore di generalizzazione rispetto al modello senza iterazioni mentre un miglioramento più significativo è quello dell'errore empirico. I due valori in se in questo caso appaiono piu lontani il che porta più imprecisione sul cosa aspettarsi nel validation (si attende comunque un valore che stia all'interno di questi due, più vicino al generalized error). Sebben il miglioramento sia poco significativo si è deciso di mantenere questo modello in quanto un miglioramento di mae dello 0.003 ai fini dello scopo del lavoro è ben accetto.

Si è valutato infine un modello basato sulle iterazioni e sullo splice delle variabili latitudine e longitudine. Gli spline sostituiscono il predittore numerico esistente con un set di colonne che consentono a un modello di emulare una relazione flessibile e non lineare. Man mano che vengono aggiunti più termini splines ai dati, la capacità di rappresentare in modo non lineare la relazione aumenta. Sfortunatamente, può anche aumentare la probabilità di rilevare tendenze di dati che si verificano per caso (ovvero, adattamento eccessivo) [2]. I gradi di liberà dello spline sono stati considerati come hyperparametri e sono stati trovati in base alla minimizzazione del mae (risultato pari a 50 per latitudine e 25 per longitudine). Si presentano cosi i risultati con il modello lineare al massimo della sua libertà.

Generalized Error	Empirical Errror
0.055	0.0517

I risultati migliorano ancora una volta il che porta supporto all'idea di inserire lo spline. I due valori appaiono distanti e può far pensare alla presenza di troppo adattamento ai dati (overfitting). Sebben questo possa esser possibile valuteremo l'opzione una volta effettuate le previsione sul validation.

4.2 Random Forest

Il primo approccio non lineare è stato tramite Random Forest. Con lo stesso principio precedente si sono trovati gli hyperparametri, risultati pari a 9 per il valore di mtry, ovvero numero di variabili scelte casualmente ad ogni split (di solito la letteratura suggerisce un valore pari alla radice quadrata del numero delle variaibli) e 4 per il numero minimo di punti che devono trovarsi in un nodo prima di un ulteriore suddivisione. Si sono cosi ottenuti in questo caso i seguenti risultati

Generalized Error	Empirical Errror
0.0574	0.0228

Si nota subito la distanza tra i due valori. Gli alberi come di consueto si sono adattati troppo al set di dati. È utile notare che senza il generalized error, attendendosi un valore sul validation di previsione simile a quello dell'empirical error, si commetterebbe un errore abissale. Si è invece cosi consapevoli che il valore di previsione nel validation rifletterà il valore di 0.0574 e non 0.0228 che rimane comunque un buon risultato (modello lineare sembrerebbe migliore)

4.3 XGBOOST

Essendo uno dei migliori modelli si è ricorsi anche al suo utilizzo aspettandosi un miglioramento rispetto al metodo appena sopra descritto. Tramite tecnica di repeated cross validation sono stati trovati i seguenti hyperparametri che hanno minimizzato il mae sul training set (è stata impostanza una griglia di ricerca manuela molto piccola per questioni computazionali):

- Trees: 500
- Min n = 37
- Tree Depth: 15
- Learning Rate: 0.0400719

- Loss reduction: 210^{-6}

Si sono cosi ottenuti i seguenti valori:

Generalized Error	Empirical Errror
0.0519	0.028

I risultati tramite xgboost sono migliori di tutti i modelli precedenti. Il generalized error è infatti basso assicurando una buona previsione nel validation. Anche in questo caso c'è un ottimo adattamento ai dati di train il che provoca quella distanza con l'altro valore.

4.4 Altri modelli

Si riportano per completezza i risultati dei modelli Svm e Knn al solo scopo informativo. Si è raggiunto un generalized error per entrambi pari a 0.062 il che ha portato alla scelta di escluderli da qualsiasi analisi successiva.

5 Modello finale

I modelli portati quindi nella fase finale del lavoro sono stati Random Forest, Xgboost e Modello Lineare con splines. Prima di procedere all'idea finale è bene confrontare le previsioni dei modelli sui "finti" nuovi dati, validation. Si potrà cosi osservare se i risultati trovati in precedenza sono robusti rispetto a nuove osservazioni.

Modello	Gen. Error	Validation Error
Lm	0.055	0.0537
RF	0.057	0.0537
Xgb	0.052	0.0500

Tutti i modelli vanno incontro ad over performance, ovvero un miglioramento del mae rispetto a quanto mostrato nei dati di training. Questo potrebbe essere spiegato dal buon adattamento rispetto ai nuovi dati. Si nota come Xgboost confermi le sue ottime perfomance cosi come il modello lineare che pareggia il valore rispetto al Random Forest.

La scelta finale è stata quella di formare un ensemble di modelli. Riassumendo velocemente gli enseble utilizzano più algoritmi per ottenere prestazioni predittive migliori di quelle che potrebbero essere ottenute da qualsiasi modello da solo. Questo metodo tende a produrre risultati buoni solo quando c'è una significativa diversità tra i vari modelli usati. [3]

A tal proposito si è scelto di combinare per ovvie ragioni il modello lineare con iterazioni e splines con xgboost (modelli completamente differenti) con l'esclusione di random forest. Il primo approccio facendo una semplice medie tra le due previsioni ha portato ad un MAE di 0.0499.
Ovviamente questo non è il miglior risultato ottenibile. Modificando infatti la semplice media con una media ponderata tra le previsione dei due modelli si può abbassare ancora di più il valore della metrica. E' stata cosi ricercata la miglior combinazione lineare che minimizzasse il MAE ottenendo un valore moltiplicativo pari a 0.26 per il modello lineare e 0.74 per xgboost. Il risultato è stato il seguente:

$$MAE = 0.0494$$

.

L'obbiettivo iniziale era ovviamente quello di prevedere al meglio i dati di test. Dato che le metriche sono state calcolate in precedenza, possiamo considerare i dati del validation come qualcosa di noto e utile ai fini della previsione finale. Si è deciso infatti di utilizzare tutto il training e il validation per la previsione del prezzo delle case sul test set, utilizzando ovviamente una previsione combinata tra modello lineare, con iterazioni e splines pesato al 28% e un modello xgboost pesato al 72%. I risultati dovrebbero essere molto simili a quelli ottenuti sul validation e comunque non distanzi ai valori dei generalized error.

Conclusioni

La parte più importante dell'intero lavoro è stato il feature engineering che ha permesso di abbassare notevolmente l'errore di previsione. Il tuning da solo infatti non permette un miglioramento sostanziale dei risultati. Si è scoperto che il miglior approccio è basato su una media ponderata tra modello lineare con iterazioni e splines pesato al 28% e un modello xgboost pesato al 72%. Quello che si trova è un MAE sul validation pari a 0.0497. La previsione finale è stata effettuata unendo il training e il validation a disposizione aspettandosi cosi risultati molto simili a quelli ottenuti o comunque non lontani dai valori dei generalized error.

References

[1] Ludvig Renbo Olsen. Multiple-k: Picking the number of folds for cross-validation, cran-r.

[2] MAX KUHN and JULIA SILGE. Tidy modeling with r, version 0.0.1.9010 (2022-01-04).

[3] L. Kuncheva and C Whitaker. *Measures of diversity in classifier ensembles, Machine Learning, 51, pp. 181-207, 2003.*