

Qualità delle abitazioni: analisi esplorativa e classificazione

Anna Giabelli, Giacomo Saccaggi, Jennifer Santini, Joana Curri, Letizia Mandelli

ABSTRACT

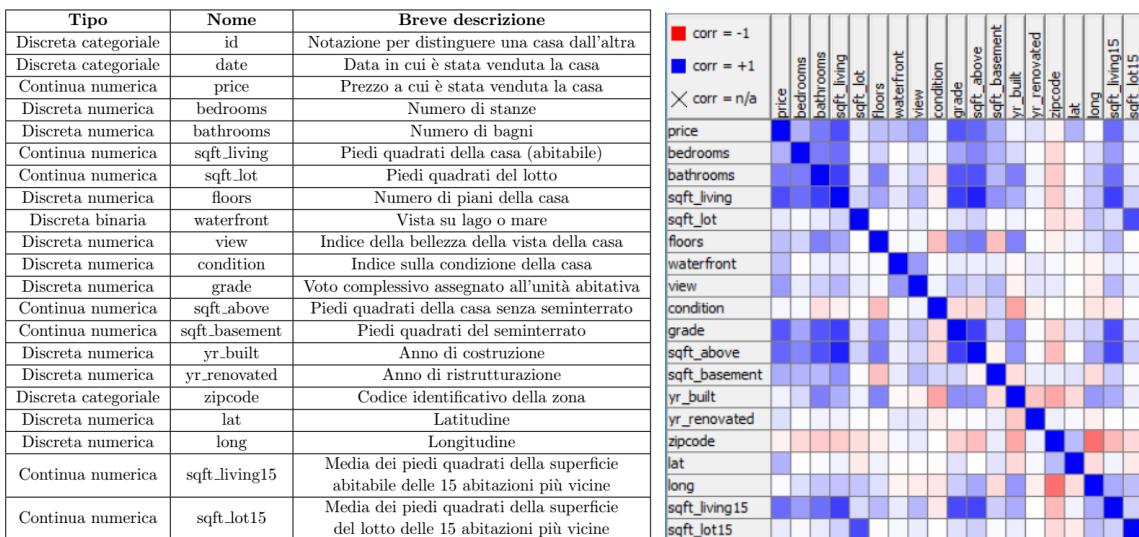
Il settore C2C^a della compravendita di beni immobili è un ambito molto complicato dove subentrano fattori soggettivi difficilmente prevedibili. Una delle variabili che influisce maggiormente su quanto un utente è disposto a pagare per un immobile, è la quotazione degli esperti. L'obiettivo di questo studio vuole essere quello di determinare in che modo sia possibile prevedere e da quali fattori dipende la valutazione dei periti circa la qualità di un edificio, così da fornire, agli utenti interessati a comprare o vendere una casa, delle linee guida di valutazione. A questo proposito, la domanda della ricerca è la seguente: in che modo è possibile valutare la qualità di un edificio e quali sono i parametri che più influenzano la votazione del perito? Per rispondere a tale domanda sono stati analizzati circa 20.000 dati relativi all'anno 2014/2015 sulla vendita di case nella contea di *King County*, valutate tramite una misura di giudizio detta *grade*. Durante questo progetto si andranno ad approfondire passo per passo tutti i passaggi delle analisi fatte, inizialmente cercando di prevedere la variabile *grade* per poi valutare gli attributi che più lo influenzano.

^aConsumer to consumer

1. INTRODUZIONE

Il dataset¹ oggetto del nostro studio riguarda le case vendute nell'arco di un anno tra Maggio 2014 e Maggio 2015. Le case si trovano tutte a King County, la contea più popolosa dello stato di Washington negli Stati Uniti. Il capoluogo di questa contea è Seattle che, insieme ad altre due contee, forma l'area metropolitana di Seattle.

Il dataset presenta 21.613 osservazioni e 21 attributi che spaziano dal prezzo dell'immobile a sue caratteristiche, come i piedi quadrati o la posizione geografica. Nello specifico gli attributi presenti nel dataset sono riportati nella tabella 1a.



(a) Tabella degli attributi

(b) Matrice di correlazione

Figura 1: Tabella degli attributi e correlazioni

Nel sito *kingcounty.gov*² la variabile *grade* è descritta come una classificazione per qualità costruttiva che si riferisce ai tipi di materiali utilizzati e alla qualità della lavorazione. Gli edifici di migliore qualità (*grade* più alto) hanno un costo di costruzione maggiore per unità di misura e hanno un valore più elevato.

Il grado di costruzione è determinato dall'assessore che valuta l'immobile per la determinazione delle tasse, quindi non cambia a meno che un'abitazione non venga ristrutturata o rinnovata ampiamente. È un compito molto soggettivo, tuttavia esso si basa su dettagliate linee guida fornite dalla contea. Chiameremo *grade* la qualità di un edificio.

Il nostro primo obiettivo consiste nel trovare un modo per prevedere tramite le caratteristiche delle case, quale sia il loro *grade*, così da vedere se sia possibile prevedere tale variabile senza il bisogno di inviare una persona a valutarne la qualità, o

¹Dove trovarlo? <https://www.kaggle.com/harlfoxem/housesalesprediction/data>

²<http://www.kingcounty.gov/depts/assessor/> /media/depts/Assessor/documents/AreaReports/2016/Residential/016.ashx

quanto meno così che si possano identificare in futuro dati anomali, magari errati. Vorremo, quindi, implementare un classificatore tramite il quale potremo identificare tali valutazioni, così che possano essere eventualmente ricontrolate. Sarebbe interessante fare tale previsione sulla base di una scala da 1 a 13, che coincide con quella utilizzata dalla contea, ma per i dati a nostra disposizione non abbiamo sufficienti osservazioni soprattutto per valori di *grade* alti e bassi. Per tale ragione, oltre che per semplificare gli algoritmi di classificazione utilizzati e per rendere più immediata l'interpretazione, abbiamo diviso questi 13 livelli della variabile *grade* in diverse categorie, basandoci sulla matrice di confusione ottenuta andando a classificare i dati su 13 stratificazioni e andando a unire le classi tra le quali il classificatore commetteva il maggior numero di errori. Tale procedimento sarà approfondito prossimamente.

Una volta divisi i dati nelle classi, per raggiungere il nostro obiettivo abbiamo utilizzato vari modelli di classificazione che fossero adatti a variabili ordinali e multi classe. Nelle analisi più comuni la variabile risposta è di tipo binario, ossia può assumere solo due valori; nel nostro caso trasformare la variabile di interesse in una variabile di questo tipo sarebbe stato estremamente riduttivo dal punto di vista interpretativo perché avrebbe significato suddividere la qualità delle abitazioni in un livello alto e in uno basso. Un'altra problematica della nostra variabile d'interesse, oltre a quella di essere multi classe, è quella di essere ordinale il che rende la classificazione ancora più delicata in quanto occorre tenere conto dell'ordine. Il fatto che l'algoritmo che utilizziamo tenga conto dell'ordine è un punto a favore del nostro studio, sebbene supponiamo che le classi nella quale la variabile *grade* è divisa siano equidistanti.

Il secondo obiettivo, strettamente collegato al primo, consiste nel capire quali siano le variabili esplicative più importanti per determinare la qualità della casa in base alla variabile *grade*. Per fare questo utilizzeremo due metodi allo scopo di selezionare le variabili in base alla loro capacità predittiva nei confronti della nostra variabile di interesse.

2. PREPROCESSING

2.1 Analisi descrittiva del dataset

Osserviamo innanzitutto che non vi sono missing values nel nostro dataset e che la somma di *sqft_above* e *sqft_basement* corrisponde a *sqft_living*. Andiamo quindi ad effettuare un'analisi più approfondita delle singole variabili e dei loro outlier, valori anomali. Le variabili *id* e *date* sono di scarso interesse per il nostro studio, quindi le trascuriamo. Riportiamo in figura 1b la matrice di correlazione delle altre variabili presenti nel dataset e in figura 2 alcuni grafici che riteniamo significativi ai fini del nostro studio: poiché esso è incentrato sulla variabile *grade* riportiamo i boxplot di *price*, *bedrooms*, *bathrooms* e *sqft_living* suddivisi per *grade* e gli scatterplot di tali variabili colorati a seconda della variabile *grade*, come da legenda.

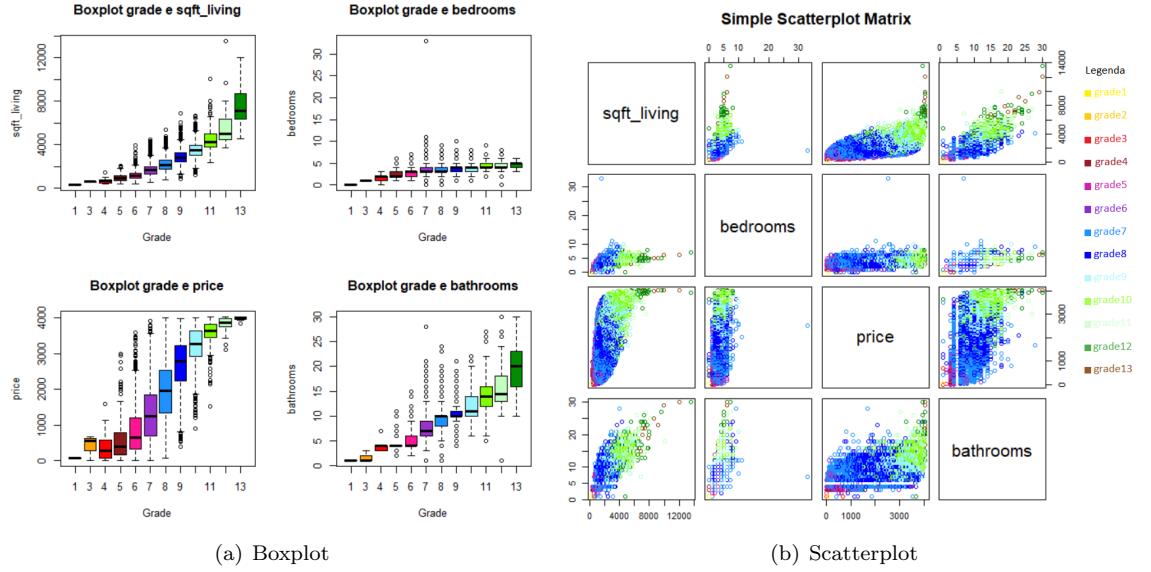


Figura 2: Boxplot di *price*, *bedrooms*, *bathrooms* e *sqft_living* suddivisi per *grade* e scatterplot colorati a seconda della variabile *grade*

Prima di procedere con la pulizia, ci soffermiamo su una breve spiegazione di tali variabili, esclusa la variabile *grade* sulla quale ci siamo già soffermati nell'introduzione essendo la variabile presa in esame per la nostra analisi. Cominciamo dalla variabile *bathrooms*, che è ottenuta come la somma delle seguenti quantità:

- 1 = locale con bagno vasca, doccia, wc, lavello
- 0.75 = locale con doccia, wc, lavello
- 0.5 = locale con wc e lavello
- 0.25 = locale con wc

Vediamo inoltre che la variabile *floor* in 2079 case non assume valore intero, in quanto l'edificio dispone di una mansarda o di un piano ammezzato: si tratta, infatti, di un “mezzo piano” tale che la metratura di tale livello è inferiore a una certa

percentuale, variabile a seconda del luogo, di quella del livello inferiore. Nel caso in cui la variabile *yr_renovated* assuma valore 0, significa che la casa non ha subito ristrutturazioni, in caso contrario viene indicato l'anno di ristrutturazione. Per quanto riguarda *sqft_lot*, esso indica i piediquadri³ del lotto cioè dalla somma del terreno sul quale è edificata la casa e dei piediquadri del giardino; in particolare nel caso di 789 abitazioni i piediquadri della casa sono maggiori di quelli dell'intero lotto (case su più piani con un giardino non molto ampio). *Waterfront* può assumere valore 0 se non c'è vista lago o mare e 1 viceversa, *view*, attributo che indica la qualità della vista, assume valori da 0 a 4, *condition* da 0 a 5. Tale ultima variabile esplicativa descrive la condizione della casa, dal pessimo all'eccellente, in relazione all'anno in cui è stata costruita. Ad esempio, se consideriamo una casa costruita nel 1930, ben costruita e ben tenuta, potrebbe essere assegnata una classe *condition*= 5, che indica che essa è in condizioni ottime.

2.2 Pulizia dei dati

Gli strumenti che abbiamo utilizzato per identificare dati anomali sono stati inizialmente i boxplot, in quanto per alcune variabili già da essi vedevamo l'assenza di outlier sospetti e in tal caso abbiamo deciso di non apportare alcuna operazione di pulizia. In un secondo momento abbiamo guardato alcuni scatterplot, tra i quali quello in figura 2b che ci ha permesso di togliere una abitazione con 33 stanze e *sqft_living* simile a quella delle abitazioni con 2/3/4 stanze, come si vede dalla figura 3a. Un altro dato anomalo corrispondente a *sqft_living*= 13540 cioè molto grande, che ha *price*= 2280000 cioè abbastanza basso. Controllando latitudine e longitudine lo abbiamo identificato e abbiamo scoperto essere una clinica veterinaria per cavalli, l'abbiamo quindi considerato come un dato anomalo che avrebbe potuto influenzare negativamente la nostra analisi e abbiamo deciso di rimuoverlo.

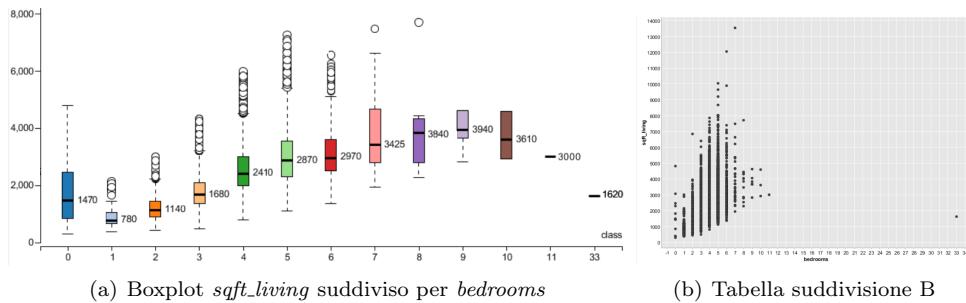


Figura 3: Boxplot e scatterplot per identificare 33 stanze come dato errato

Inoltre abbiamo deciso di rimuovere le abitazioni con 0 bagni e 0 stanze perché hanno una qualità alta, come vediamo nella figura 4, e a nostro parere sono anch'essi dati anomali. Abbiamo controllato anche i dati con *bathrooms*= 0 e *bedrooms*= 1, ma essi sono compatibili con la scala di qualità data dalla contea. Abbiamo più approfonditamente analizzato il dato *Row10481* ma esso si trova vicino a un lago e ha un grande parco, decidiamo di tenerlo in considerazione nella nostra analisi.

Row ID	D id	S date	D price	I bedrooms	D ▲ bat...	I grade
Row6994	2,954,400,190	20140624T000000	1,295,650	0	0	12
Row875	6,306,400,140	20140612T000000	1,095,000	0	0	7
Row9773	3,374,500,520	20150429T000000	355,000	0	0	8
Row3119	3,918,400,017	20150205T000000	380,000	0	0	8
Row9854	7,849,202,190	20141223T000000	235,000	0	0	7
Row1423	9,543,000,205	20150413T000000	139,950	0	0	7
Row10481	203,100,435	20140918T000000	484,000	1	0	7
Row1149	3,421,079,032	20150217T000000	75,000	1	0	3
Row5832	5,702,500,050	20141104T000000	280,000	1	0	3
Row19452	3,980,300,371	20140926T000000	142,000	0	0	1

Figura 4: Tabella dei dati con 0 bagni

Abbiamo così concluso la pulizia del nostro dataset e il numero di record è passato da 21.613 a 21.605. Abbiamo potuto procedere in questo modo minuzioso in quanto non erano presenti missing data e, pur riconoscendo che su 21.613 lavorare sul singolo dato potrebbe sembrare eccessivo, abbiamo pensato di procedere in questo modo così da sperare di ottenere un modello di classificazione più preciso.

3. MODELLO DI PREVISIONE DELLA VARIABILE GRADE

3.1 Osservazione preliminare

Nel decidere di compiere questa analisi, come spiegato ampiamente nell'introduzione, desideriamo predire la variabile *grade*, stabilita sulla base delle caratteristiche della casa. Il prezzo viene dato a posteriori e con una semplice analisi⁴ otteniamo

³1 foot = 1 piede = 0.3048 metri. Decidiamo di lasciare tale unità di misura in quanto il nostro modello di previsione si basa su un dataset americano e potrà essere usato con efficienza in quel contesto.

⁴La figura 5 è stata ottenuta utilizzando il *Tree Ensemble Learner (Regression)*, un nodo KNIME che sfrutta l'algoritmo di *RandomForest*; lo abbiamo preferito rispetto al *Tree Ensemble Learner*, un altro nodo di KNIME, in quanto quest'ultimo non tiene conto del fatto che *price* sia una variabile continua e quindi la analizzerebbe come una multi classe. Successivamente affronteremo in modo più approfondito il funzionamento di questo algoritmo ai fini di identificare gli attributi più influenti nella previsione della variabile *grade*.

la figura 5 che mostra come esso dipenda in gran parte dalla variabile *grade*. Abbiamo, quindi, deciso di rimuovere tale variabile esplicativa dalla nostra analisi, in quanto è nostro interesse che tale modello di previsione si basi strettamente sulle caratteristiche oggettive della casa, disponibili prima della vendita.

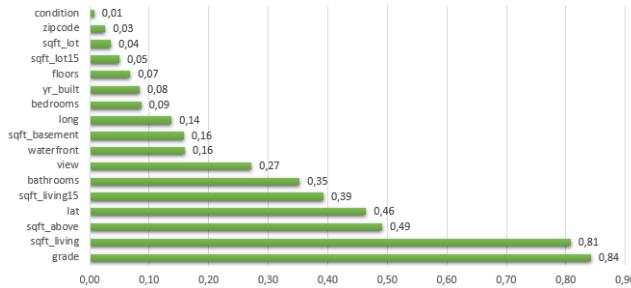


Figura 5: Grafico delle *importance* degli attributi nello spiegare *price*

3.2 Criteri per la suddivisione della variabile *grade* in classi e algoritmo utilizzato

Abbiamo deciso di suddividere la variabile *grade* in classi e per fare questo ci basiamo sulle matrici di confusione ottenute con diversi metodi: utilizziamo come classificatore l'*OrdinalClassClassifier*, un algoritmo costruito appositamente da Frank e Hall per variabili ordinali multiclass come quella presa in esame nel nostro lavoro. Abbiamo escluso la regressione lineare in quanto la variabile di previsione e molte delle variabili esplicative sono discrete. Questo algoritmo è presente in un nodo Weka e lo presentiamo in breve nelle note strutturandolo in 4 punti, rimandando all'articolo di Frank e Hall⁵ per ulteriori approfondimenti.

Decidiamo di confrontare 6 metodi creati utilizzando diversi classificatori all'interno dell'algoritmo precedentemente mostrato: *NBTree*, *MultiClass*, *OrdinalClass*, *J48*, *NaiveBayes* e *RandomForest*. Osserviamo le tabelle contenenti l'accuratezza e la Kappa di Cohen, coefficiente statistico che rappresenta il grado di accuratezza e affidabilità della classificazione. Entrambi hanno un range tra 0 e 1 e il classificatore è tanto migliore quanto più tali indici sono vicini a 1. Utilizziamo il campionamento stratificato con training set pari al 67% dei dati e test set pari al 33%.

Row ID	Acc_NBTree	Acc_MultiClass	Acc_Ordinal...	Acc_J48	Acc_NaiveBa...	Acc_Random...
Overall	0.567	0.516	0.617	0.617	0.534	0.698

(a) Tabella delle accuracy ottenute coi diversi algoritmi

Row ID	CK_NBTree	CK_MultiClass	CK_OrdinalCl...	CK_J48	CK_NAiveBa...	CK_Random...
Overall	0.405	0.269	0.466	0.466	0.33	0.575

(b) Tabella degli indici Kappa di Cohen ottenute coi diversi algoritmi

Figura 6: Tabelle degli indici di bontà del classificatore con diversi algoritmi

Consideriamo la matrice di confusione con accuratezza maggiore ossia quella ottenuta col metodo *RandomForest* con 60 alberi:

Row ID	4	5	6	7	8	9	10	11	12	13
4	4	2	4	0	0	0	0	0	0	0
5	2	29	37	12	0	0	0	0	0	0
6	0	22	375	271	5	0	0	0	0	0
7	0	0	140	2417	381	23	2	0	0	0
8	0	0	0	430	1399	165	7	0	0	1
9	0	0	0	19	258	506	76	4	0	0
10	0	0	0	3	24	131	192	24	0	0
11	0	0	0	0	1	12	60	52	7	0
12	0	0	0	0	0	1	7	17	4	0
13	0	0	0	0	0	0	2	1	1	0

Figura 7: Matrice di confusione prima della suddivisione di *grade* in classi

⁵L'articolo si trova per esempio al link: <https://researchcommons.waikato.ac.nz/bitstream/handle/10289/64/content.pdf>

1. *Trasformation of attributes*: il primo step di questo algoritmo consiste nel convertire le k classi ordinali, in k-1 classi binarie. Tale trasformazione avviene binarizzando in modo sequenziale le classi ordinali inserite nel dataset.

2. *Derived Datasets*: dal primo passaggio deriviamo nuovi Dataset; uno per ognuno dei k-1 attributi.

3. *Classifiers*: successivamente, l'algoritmo *OrdinalClassClassifier* è applicato per generare un classificatore (il classificatore scelto viene indicato nella configurazione del nodo) per ogni dataset creato allo step precedente.

4. *Predicted Classes*: per predire il valore delle k classi ordinali dobbiamo stimare la probabilità di queste, usando i k-1 modelli creati. Per la riuscita di questa operazione, si va a confrontare il classificatore di ciascun modello per poi scegliere quello con la probabilità maggiore. Il confronto prevede la sottrazione della probabilità condizionata della i-esima classe per i suoi attributi, alla probabilità condizionata della (*i*-1)-esima classe per gli attributi della *i*-esima classe (con $1 < i < k$); ovvero la sottrazione del classificatore riferito al Dataset (*i*-1)-esimo (creato allo step 2-3), meno lo stesso riferito al dataset *i*-esimo.

Possiamo osservare sia da questa *confusion matrix* sia da quelle ottenute con gli altri algoritmi di classificazione che sono possibili diverse suddivisioni in classi, riportiamo quelle da noi analizzate:

Classi in cui suddivido	Diverse suddivisioni della variabile <i>grade</i>					
	3 classi	4 classi	5 classi	6 classi A	6 classi B	6 classi C
grade= 1	1-6	1-5	1-4	1-3	1-4	1-3
grade= 2	7-9	6-7	5-6	4-5	5-6	4-5
grade= 3	10-13	8-9	7-8	6	7	6-7
grade= 4		10-13	9-10	7-8	8	8-9
grade= 5			11-13	9-10	9-10	10-11
grade= 6				11-13	11-13	12-13

Nel fare tale suddivisione in livelli abbiamo tenuto conto sia della matrice di confusione sia della scelta che ci sembrava più ragionevole guardando i nostri dati: da una parte abbiamo considerato il fatto che le frequenze assolute con le quali compaiono nelle fasce più alte (*grade*= 11, 12, 13) e basse (*grade*= 1, 2, 3) sono molto minori, rispetto alla parte centrale, dall'altra il significato della variabile, *grade* spiegato nell'introduzione, per il quale è sensato che, per esempio, le classi 1-3 non vengano mai suddivise. Decidiamo di analizzare diverse casistiche in quanto noi preferiremmo classificare in più livelli, poiché ci permette una classificazione in base alla qualità più simile a quella fatta dalla contea, ma dobbiamo anche valutare la bontà di tale classificazione perché, come abbiamo visto nella tabella 7, più classi abbiamo più probabile sarà sbagliare a classificare. Presenteremo i risultati ottenuti nel paragrafo successivo.

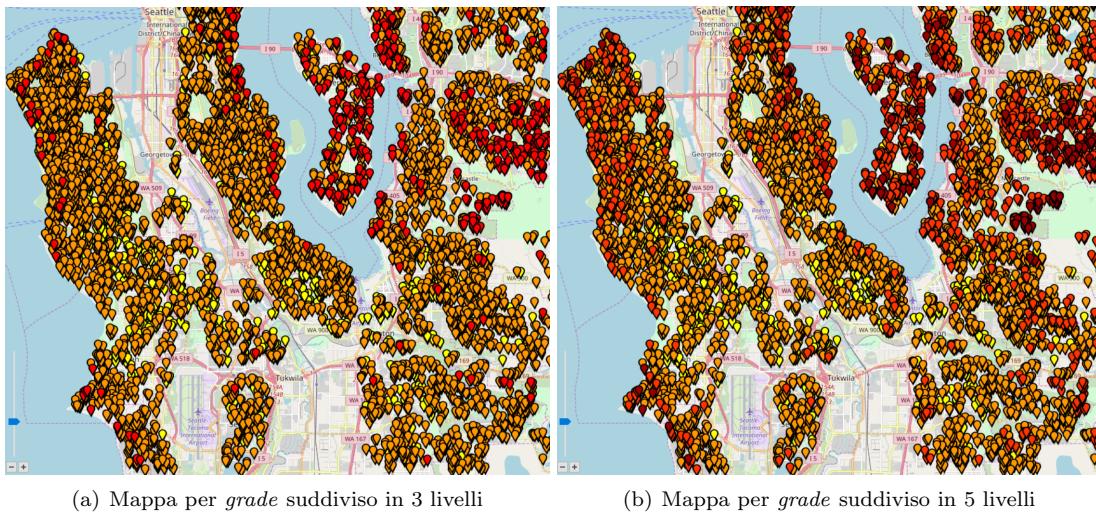


Figura 8: Mappe a confronto per diverse suddivisioni di *grade*

In queste mappe possiamo notare come vengano suddivise le case a seconda della diversa scelta dei livelli, dal giallo al bordeaux, che corrispondono rispettivamente a qualità 1 e qualità 5. Possiamo notare come dal grafico con 3 livelli sia facile identificare a colpo d'occhio le zone più prestigiose e quelle con case di qualità inferiore, tuttavia la suddivisione in 5 è in grado allo stesso tempo di fare tale suddivisione in modo più preciso, basti confrontare, ad esempio la zona ovest dell'immagine: qui possiamo osservare come a nord sulla costa vi siano case di qualità maggiore rispetto a quelle presenti sulla costa ovest, che dalla suddivisione in 3 classi apparivano della stessa qualità.

3.3 Suddivisione in classi effettuata e classificazione

Per ognuna delle possibili suddivisioni in livelli abbiamo effettuato 6 classificazioni utilizzando l'algoritmo presentato in precedenza, *OrdinalClassClassifier*, con i 6 algoritmi di classificazione *NBTree* (*NBT*), *MultiClass* (*Mc*), *OrdinalClass* (*Oc*), *J48*, *NaiveBayes* (*NB*) e *RandomForest* (*RF*), adatti al nostro problema. Per partizionare il dataset abbiamo utilizzato il metodo *holdout* con training set 67%, sul quale andremo a costruire la regola, e test set 33%, sul quale tale regola verrà testata. Confrontiamo innanzitutto le diverse tabelle ottenute dalla suddivisione in 6 classi in diversi modi, tutti plausibili guardando la matrice di confusione in figura 7:

Row ID	D NBT	D Mc	D Oc	D J48	D NB	D RF
P (Grade=1.0)	0.813	1	0.561	0.561	0.999	1
P (Grade=2.0)	0.95	0.932	0.772	0.772	0.885	0.958
P (Grade=3.0)	0.904	0.923	0.821	0.821	0.756	0.951
P (Grade=4.0)	0.871	0.88	0.822	0.822	0.811	0.927
P (Grade=5.0)	0.891	0.932	0.838	0.838	0.829	0.954
P (Grade=6.0)	0.965	0.98	0.906	0.906	0.972	0.972

(a) Tabella suddivisione A

Row ID	D NBT	D Mc	D Oc	D J48	D NB	D RF
P (Grade=1.0)	0.997	0.993	0.595	0.595	0.944	0.995
P (Grade=2.0)	0.936	0.934	0.843	0.843	0.775	0.96
P (Grade=3.0)	0.85	0.84	0.786	0.786	0.727	0.909
P (Grade=4.0)	0.81	0.811	0.734	0.734	0.767	0.881
P (Grade=5.0)	0.893	0.932	0.839	0.839	0.829	0.954
P (Grade=6.0)	0.966	0.98	0.903	0.903	0.972	0.972

(b) Tabella suddivisione B

Row ID	D NBT	D Mc	D Oc	D J48	D NB	D RF
P (Grade=1.0)	0.788	1	0.828	0.828	0.999	1
P (Grade=2.0)	0.951	0.932	0.797	0.797	0.885	0.958
P (Grade=3.0)	0.902	0.902	0.842	0.842	0.787	0.945
P (Grade=4.0)	0.869	0.868	0.804	0.804	0.806	0.922
P (Grade=5.0)	0.936	0.958	0.822	0.822	0.814	0.964
P (Grade=6.0)	0.977	0.977	0.784	0.784	0.99	0.968

(c) Tabella suddivisione C

Figura 9: Confronto diverse suddivisioni delle variabili in 6 livelli

Vediamo che il classificatore migliore è il *RandomForest* con 60 alberi, escludiamo la suddivisione B in quanto è presente un valore al di sotto di 0.9 a differenza delle altre due, inoltre confrontiamo le medie di tali aree sottese dalle curve ROC e vediamo che la classificazione migliore è quella ottenuta con la suddivisione A. La scelta dei 60 alberi è stata fatta procedendo per tentativi, poiché ci sembrava il miglior compromesso tra accuratezza e tempi computazionali.

Row ID	D NBT	D Mc	D Oc	D J48	D NB	D RF
P (grade=1.0)	0.932	0.933	0.848	0.848	0.905	0.958
P (grade=2.0)	0.883	0.895	0.821	0.821	0.846	0.934
P (grade=3.0)	0.941	0.968	0.848	0.848	0.951	0.976
P (grade=4.0)						
P (Grade=1.0)	0.955	0.933	0.799	0.799	0.905	0.96
P (Grade=2.0)	0.902	0.902	0.842	0.842	0.787	0.944
P (Grade=3.0)	0.868	0.868	0.804	0.804	0.806	0.922
P (Grade=4.0)	0.95	0.965	0.849	0.849	0.948	0.969
P (Grade=5.0)						

(a) Tabella classificazione in 3 livelli

(b) Tabella classificazione in 4 livelli

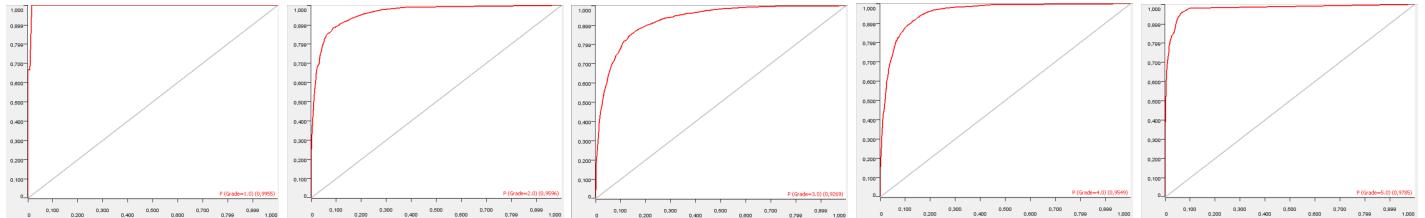
(c) Tabella classificazione in 5 livelli

Figura 10: Confronto diverse classificazione ottenute con suddivisione in 3, 4 e 5 livelli.

Row ID	D Mean 6 livelli	D Mean 5 livelli	D Mean 4 livelli	D Mean 3 livelli
Area Under C...	0.96	0.963	0.949	0.956

Figura 11: Tabella medie delle curve ROC della variabile *grade* con diverse suddivisioni

Nella tabella 11 riportiamo il calcolo della media delle aree sottese dalle curve ROC ottenute col classificatore migliore, che è il *RandomForest* nel caso delle diverse suddivisioni, come si vede dalla tabella 10 e dalla tabella 9, indice della bontà complessiva del metodo. Oltre alla media delle aree sottese di tali curve, vogliamo anche un classificatore che funzioni bene su tutti i valori assunti dalla variabile *grade*, quindi faremo anche questo ulteriore controllo che, guardando i valori ottenuti, corrisponderà a chiedere che tali aree in ognuno dei valori della variabile *grade* sia maggiore di 0.9. Sceglieremo di tenere la classificazione più precisa che è ottenuta con 5 livelli, come vediamo dalla tabella 11. Nel considerare tale suddivisione in 5 livelli, notiamo che ad essa possiamo facilmente associare, a livello interpretativo, le quantità *basso*, *medio-basso*, *medio*, *medio-alto*, *alto*. In conclusione, sceglieremo il classificatore migliore ossia *Ordinal Class Classifier* con algoritmo *RandomForest* con 60 alberi e variabile *grade* suddivisa in 5 livelli; riportiamo le curve ROC con esso ottenute, che confrontano ciascun livello di *grade* con tutti gli altri:



(a) Curva ROC *grade*= 1, (b) Curva ROC *grade*= 2, (c) Curva ROC *grade*= 3, (d) Curva ROC *grade*= 4, (e) Curva ROC *grade*= 5, AUC=0.9955 AUC=0.99% AUC=0.9269 AUC=0.9549 AUC=0.9785

Figura 12: Curve ROC per la classificazione di *grade* divisa in 5 livelli, *Ordinal Class Classifier* con algoritmo *RandomForest*

4. VARIABILI INFLUENTI NELLA DETERMINAZIONE DI *GRADE*

4.1 *RandomForest* variable importance

L'algoritmo *RandomForest* può essere usato, grazie al nodo *Tree Ensemble Learner*, come visto in precedenza, per determinare l'*importance* delle variabili nei problemi dove si attua una regressione o una classificazione.

Il valore dell'importanza degli i-esimi attributi viene valutato sui primi tre livelli ossia nelle prime due suddivisioni, in quanto esse risultano essere le più decisive per la creazione del modello. Per riuscire a comprendere quanto gli attributi spieghino la variabile *grade* abbiamo adottato tale nodo in quanto questo analizza *grade* come una variabile multi classe, tuttavia tale strumento non considera il fatto che l'attributo sia ordinale. Per raggiungere l'obiettivo di scoprire quali variabili siano più importanti per la previsione di *grade*, abbiamo utilizzato la seconda tabella di output, la quale ci fornisce informazioni su quante volte ciascun attributo venga utilizzato nei primi 3 livelli, *number of splits (level x)*, e il numero di volte in cui un attributo era presente nel modello, *number of candidates (level x)*. Per calcolare l'influenza che le variabili hanno sulla variabile *grade* sceglieremo di calcolare una media dell'importanza dei primi 3 livelli per ogni attributo:

$$importance = \frac{1}{3} \sum_{x=0}^2 \frac{\#split(level(x))}{\#candidates(level(x))}$$

I risultati di questa analisi sono:

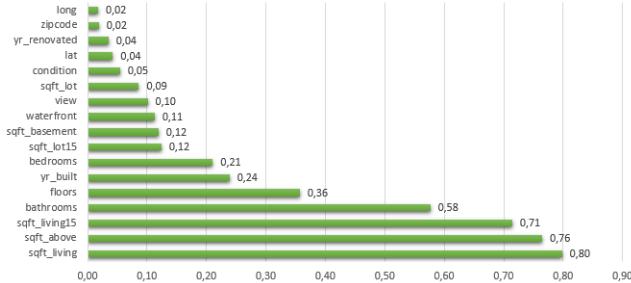


Figura 13: Tabella e grafico che a seconda dell'*importance* evidenziano le variabili che spiegano maggiormente *grade*

4.2 Features Selection - Wrapper

Vorremmo ora capire quali variabili siano più importanti nel modello di previsione da noi identificato come il migliore. Per rispondere a questa domanda abbiamo deciso di usare la tecnica di *Features Selection* che serve per identificare gli attributi ridondanti, cioè quelli che contengono informazioni già disponibili grazie ad altri attributi, e quelli irrilevanti, cioè quelli che contengono informazioni non utili per risolvere il problema considerato. Prima di passare al vero e proprio svolgimento di questa operazione abbiamo deciso di fare una partizione con un campionamento stratificato il quale ci permette di avere una maggiore rappresentabilità dei dati, e dividiamo training set pari al 67% e test set pari al 33%. Sono disponibili diversi approcci per applicare la *Features Selection* e per rispondere in modo corretto ed efficiente al nostro problema, abbiamo deciso di applicare il metodo *Wrapper* poiché esso misura l'utilità delle caratteristiche sulla base della performance del classificatore e quindi ci suggerisce di tenere solo le variabili che permettono di predire meglio *grade* sulla base del classificatore.

Il criterio con cui abbiamo scelto *OrdinalClassClassifier* con impostato il classificatore *RandomForest* per implementare tale metodo è quello della valutazione della sua performance predittiva prima della *Features Selection* in confronto alla performance degli altri classificatori. Riportiamo il modello che esce dal *Backward Feature Elimination Filter* dopo il confronto con il test set nel caso di 8 variabili in figura 14. Scegliamo tale modello perché ci sembra un giusto compromesso tra numerosità delle variabili ed errore di previsione.

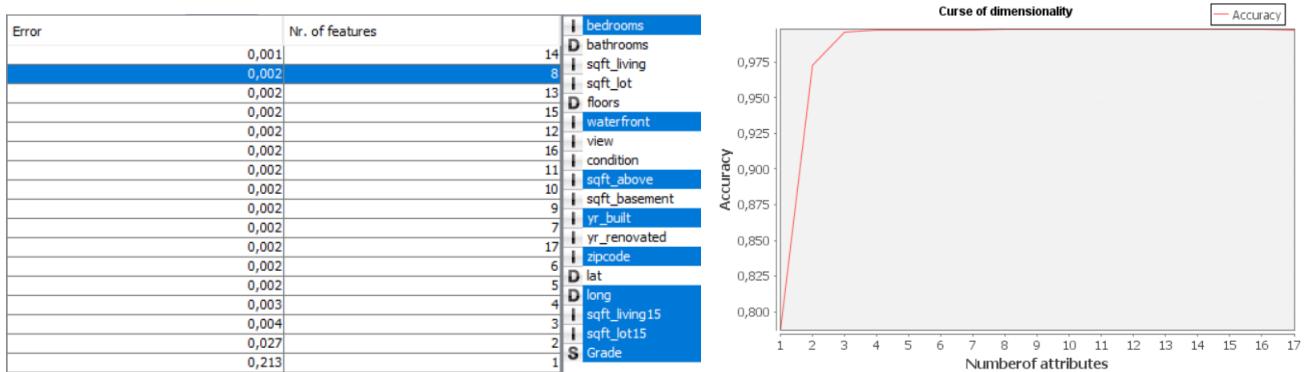


Figura 14: Selezione del numero di variabili che teniamo dalla *Features Selection*

Si può notare che le variabili che compaiono nel modello, in ordine di comparsa, sono:

1. Sqft_above: rappresenta i footquadri della parte della casa abitabile senza considerare il seminterrato. Tale variabile è anche la seconda in ordine di *importance*, il che conferma la sua influenza nel predire *grade*
2. Long: insieme a zipcode ci indica un'importante rilevanza della zona
3. Yr_built
4. Zipcode
5. Bedrooms
6. Waterfront
7. Sqft_living15: insieme a Sqft_lot15 ci fa intuire che il valore della casa dipende molto dalla media della grandezza delle case dei 15 vicini più prossimi, che in un certo senso è un indice della zona nella quale una certa abitazione si trova
8. Sqft_lot15

Per valutare la bontà della *Features Selection* implementata abbiamo eseguito una procedura di *Cross Validation* sia sul dataset contenente solo le 8 variabili selezionate, sia sul dataset contenente tutte le variabili, usando la partizione che non era stata usata per implementare la *Features Selection*, altrimenti la validazione non avrebbe avuto significato. In figura 15 si vede come in seguito alla *Features Selection* gli errori di classificazione in percentuale commessi nella procedura di cross validation siano aumentati di poco: ad esempio, la mediana degli errori passa da 15,57 a 16,69.

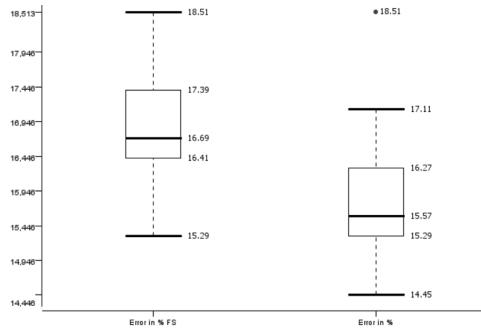
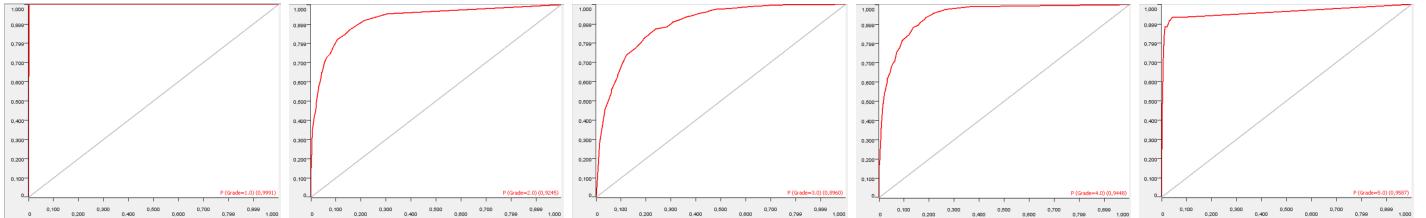


Figura 15: Confronto dei boxplot delle distribuzioni degli errori prima e dopo la *Features Selection*



(a) Curva ROC $grade=1$, AUC=0.9991 (b) Curva ROC $grade=2$, AUC=0.9245 (c) Curva ROC $grade=3$, AUC=0.8960 (d) Curva ROC $grade=4$, AUC=0.9448 (e) Curva ROC $grade=5$, AUC=0.9587

Figura 16: Curve ROC per la classificazione di *grade* effettuata solo con gli attributi selezionati dalla *Features Selection*

La classificazione implementata con le variabili selezionate dalla *Features Selection* fornisce risultati soddisfacenti, dato che la media delle aree sottostanti le curve ROC è 0,945. La *Features Selection*, quindi, ha fornito dei buoni risultati e ci ha portato a dire che le variabili più significative in termini di capacità predittiva del valore della variabile *grade* sono quelle riportate nel precedente elenco.

5. CONCLUSIONI

L'intento del nostro lavoro è stato fornire informazioni utili e innovative per qualsiasi stakeholder sulle vendite delle case tra il 2014 e il 2015 nella contea di Seattle.

Il primo obiettivo del nostro lavoro è stato trovare un modo per prevedere tramite le caratteristiche delle case, quale fosse il loro *grade*. Per procedere siamo stati costretti, per mancanza di dati nei valori più estremi di *grade* a raggruppare alcuni livelli di tale variabile e per compiere questo processo nel modo migliore possibile abbiamo osservato la matrice di dispersione: abbiamo, così, delineato 4 possibili suddivisioni in 3, 4, 5 e 6 livelli. Successivamente abbiamo creato e testato le regole di classificazione per tutte le suddivisioni possibili, analizzando la variabile oggetto di studio come un attributo multi classe ordinale. Per far questo abbiamo studiato e utilizzato *OrdinalClassClassifier*, un nodo di KNIME, usando al suo interno 6 algoritmi di classificazione *NBTree*, *Multi Class*, *Ordinal Class*, *J48*, *Naive Bayes* e *RandomForest*, adatti al nostro problema. La migliore classificazione risulta essere quella ottenuta con l'algoritmo di classificazione *RandomForest* suddividendo *grade* in 5 livelli.

Il secondo obiettivo è stato determinare quali fossero le variabili che influenzano di più la qualità di una casa. Abbiamo affrontato la risoluzione di questo problema tramite due modalità differenti: la prima è la *RandomForest variable importance* e la seconda è la *Features Selection*. Entrambi questi due approcci rispondono alla domanda posta inizialmente, il primo analizzando gli alberi dei primi tre livelli della foresta e osservando le variabili più influenti mentre il secondo creando dei modelli con differente numerosità di variabili e valutando a quanto ammonta la perdita di informazione dopo la loro rimozione nei diversi casi. La *RandomForest variable importance* ha tempi computazionali relativamente brevi e fornisce un output semplice e chiaro a differenza della *Features Selection* che ha tempi computazionalmente più lunghi e l'output, in risposta alla nostra domanda, risulta difficilmente traducibile, ma l'applicabilità della prima modalità risulta dubbia in quanto possiede solo una funzione informativa e non permette di creare un vero e proprio modello che sia facilmente trasportabile e generalizzabile mentre con il secondo approccio siamo in grado di generalizzare un modello che sia in grado di determinare la variabile *grade* attraverso un numero minore di variabili.

Il nostro lavoro costituisce un punto di inizio di un'analisi che può da un lato indirizzare l'utente a valutare una casa e dall'altro può aiutare il *Glossary of Terms* a scoprire eventuali errori o evitare di far uscire un perito per valutare ogni casa. Con una maggior quantità di dati raccolti negli anni si potrebbe forse pensare di rifare tale modello di classificazione su un range di *grade* da 1 a 13, tenendo conto dell'anno nel quale il dato è stato raccolto. Si potrebbe, inoltre, utilizzare al posto del metodo *holdout* la *Cross Validation*, poiché con essa pensiamo si possa ottenere un incremento della bontà del classificatore.

Riferimenti bibliografici

- [1] <https://www.kaggle.com/harlfoxem/housesalesprediction/data>
- [2] <https://www.knime.com>
- [3] <https://researchcommons.waikato.ac.nz/bitstream/handle/10289/64/content.pdf>
- [4] <http://grepcode.com/file/repo1.maven.org/maven2/nz.ac.waikato.cms.weka/ordinalClassClassifier/1.0.1/weka/classifiers/meta/OrdinalClassClassifier.java>
- [5] <http://www.uokufa.edu.iq/staff/ehsanali/Tan.pdf>
- [6] https://en.m.wikipedia.org/wiki/King_County,_Washington
- [7] <http://www.kingcounty.gov/depts/assessor/media/depts/Assessor/documents/AreaReports/2016/Residential/016.ashx>
- [8] <http://raincityguide.com/category/seattle-real-estate-guide/housing-market/>
- [9] <http://raincityguide.com/category/seattle-employers/>
- [10] <http://raincityguide.com/category/seattle-real-estate-guide/local-information/>
- [11] <http://raincityguide.com/2006/07/07/what-is-a-25-bathroom/>