Carnati Emanuele, 781231
Buccoliero Antonello, 776830
De Cicco Alessandro, 834225

VENDITA DI CASE NELLA CONTEA DI KING, USA

PREDIZIONE DEL PREZZO DI VENDITA DELLE CASE TRAMITE LA REGRESSIONE

INDICE

Abstract

La contea di King è una contea dello stato di Washington, negli Stati Uniti.

È la contea più popolata dello stato di Washington, nonché la tredicesima negli Stati Uniti.
Il capoluogo di contea è Seattle, che è anche la città più grande dello stato.
In termini geografici, la contea si estende su una superficie totale di 5980 km², di cui 5480 km² è terra e 490 km² (circa l'8%) è acqua. La contea possiede un territorio molto ampio che si estende dalla costa fino all'entroterra, raggiungendo il punto più alto a 2426 metri, con il Mount Daniel.
Attraverso uno studio di regressione lineare, si vuole predire i prezzi delle case della contea basandosi su più caratteristiche fornite dal database, utilizzando modelli avanzati e metodi di Machine Learning.

Verrà mostrata la complessità del modello ed in quale modo possiamo selezionare il miglior modello predittivo usando le tecniche di validazione.

1. Introduzione

Questo elaborato presenta la predizione di una variabile continua come il prezzo attraverso metodi di machine learning.
Più nel dettaglio, viene studiata la possibilità di predire il prezzo delle case partendo dalle variabili presenti nel dataset. Questo obiettivo lo si è perseguito comparando le performance di diversi classificatori.

In primo luogo, passeremo attraverso un'approfondita esplorazione dei dati per identificare le caratteristiche più importanti ed esplorare la correlazione tra le variabili. Successivamente, i dati verranno normalizzati e sarà effettuata un'analisi iniziale di regressione, infine verranno applicati diversi algoritmi di apprendimento automatico e verrà calcolata la loro bontà.
Il modello sviluppato, aiuterà un probabile acquirente a selezionare la casa giusta, in termini di prezzo e caratteristiche.

Prima di iniziare l'analisi esploratoria dei dati, è utile mostrarne il contesto; un buon modo per farlo è attraverso le mappe, quindi sfruttando le coordinate delle singole case presenti sul database.
Tramite l'applicativo Tableau Desktop sono state mappate le posizioni delle case; raggruppando le case per zipcode, otteniamo una visione generale che ci aiuta a capire meglio il contesto dell'analisi.

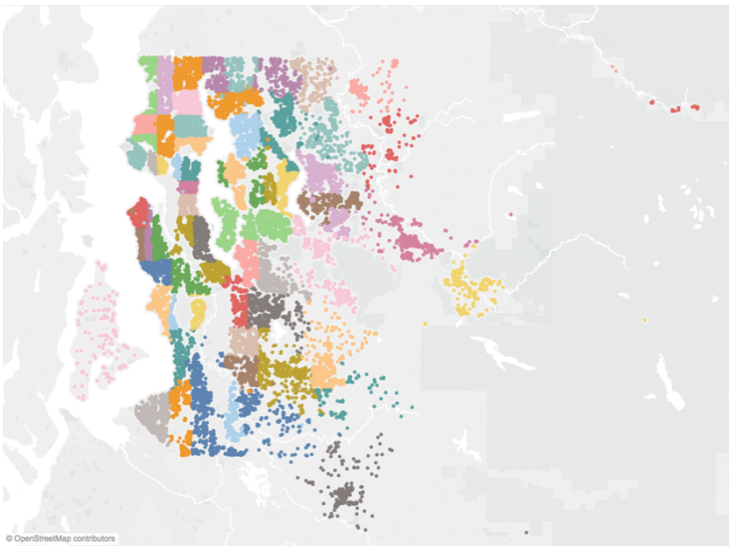

Fig. 1 - Geolocalizzazione delle case, suddivise per zipcode

Per la creazione del modello è stato utilizzato l'applicativo KNIME, piattaforma gratuita open source tramite la quale è stato creato un flusso di analisi dati.

2. Dataset e preprocessing

2.1. Il dataset selezionato

I dati utilizzati per creare il modello di predizione sono stati scaricati dal sito di dataset kaggle.com, all'indirizzo https://www.kaggle.com/harlfoxem/housesalesprediction.
Tali dati identificano le case presenti nella contea di King attraverso 19 attributi:

id (identificativo di una casa), date (data in cui la casa è stata venduta), price (prezzo di vendita della casa), bedrooms (numero di camere da letto nella casa), bathrooms (numero di bagni nella casa), sqft_living (metratura della casa), sqft_lot (metratura del lotto), floors (numero di piani nella casa), waterfront (casa che ha una vista sull'acqua), view (numero di visite alla casa), condition (condizione della casa), grade (grado generale che è stato assegnato alla casa, basato sul sistema di rating di King County), sqft_above (metratura della casa a parte il seminterrato), yr_built (anno di costruzione), yr_renovated (anno in cui la casa è stata rinnovata), zipcode (codice postale), lat (latitudine), long (longitudine), sqft_living15 (metratura della casa media delle 15 case più vicine), sqft_lot15 (metratura del lotto media delle 15 case più vicine).

Il database in questione contiene 21.613 record, una buona quantità di case su cui basare la costruzione del modello di predizione del prezzo.
Contiene anche numerosi attributi, alcuni dei quali inutili ai fini del nostro progetto; per evitare di incorrere in risultati distorti, sono state applicate metodologie di preprocessing per ridurre la dimensionalità e considerare solo gli attributi significativi.

2.2 Preprocessing ed analisi esploratoria dei dati

Il workflow (flusso) di analisi legge inizialmente il file CSV originale (File Reader), sul quale vengono eseguite le prime operazioni di preprocessing: eliminazione delle colonne "id" (attributo di tipo stringa dal quale non possono essere estratte informazioni utili al fine del modello) e "date" (date comprese fra gli anni 2014 e 2015, aventi valore insignificante).
Considerando il numero non eccessivo di valori presenti nel nostro dataset, non è stata applicata nessuna tecnica di campionamento per ridurre il numero di record in analisi, per non incorrere in un nuovo dataset poco rappresentativo.
Inizialmente si vuole mostrare la distribuzione della variabile target "price", attraverso un boxplot:

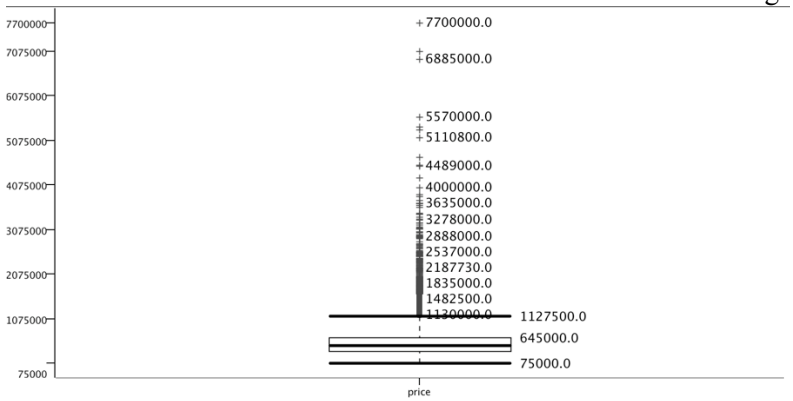

Fig. 2 – Boxplot dell'attributo "price"

Analizzando il boxplot presente in *figura 2* possiamo notare un gran numero di outliers al di sopra del valore estremo del boxplot ($1.127.500,0), con poche case aventi il prezzo sopra i $ 5.000.000,00.
Se ignoriamo gli outliers, la maggior parte delle case sono situate nella fascia di prezzo che va da $75.000 a $1.127.500. Inoltre, possiamo notare come la distribuzione fra il primo ed il terzo quartile varia tra $321.950,0 e $645.000,0, con un range interquartile del valore di $323.050,0 ed una mediana di $450.000,0.

Dopo aver introdotto la variabile "price", andiamo a considerare la correlazione esistente fra la nostra variabile target e le altre variabili indipendenti presenti all'interno del dataset.

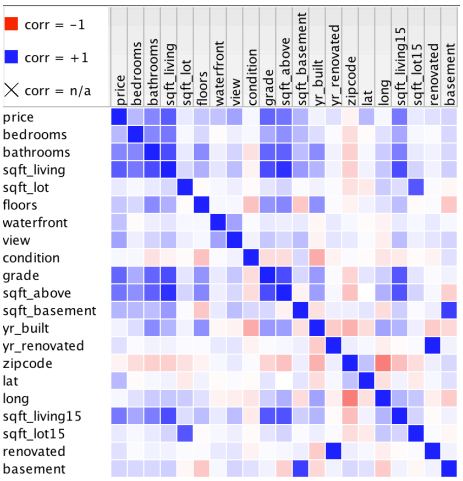

Fig. 3 – Matrice di correlazione

Row ID	price
price	1
sqft_living	0.702
grade	0.667
sqft_above	0.606
sqft_living15	0.585
bathrooms	0.525
view	0.397
sqft_basement	0.324
bedrooms	0.308
lat	0.307
waterfront	0.266
floors	0.257
basement	0.18
yr_renovated	0.126
renovated	0.126
sqft_lot	0.09
sqft_lot15	0.082
yr_built	0.054
condition	0.036
long	0.022
zipcode	-0.053

Fig. 4 – Coefficiente lineare di Pearson fra variabile target e variabili indipendenti

Il dataset viene prima normalizzato e poi applicata la funzione "Linear Correlation".

Analizzando la matrice di correlazione ottenuta (*figura 3*), le variabili "bathrooms", "sqft_living", "grade", "sqft_above" ed "sqft_living15" risultano avere il valore del coefficiente lineare di Pearson "r" più elevato, e quindi ci potrebbe essere un'eventuale relazione lineare rispetto la variabile target. I valori sono rispettivamente 0.525, 0.702, 0.667, 0.606 e 0.585 (*figura 4*).

Bisogna far distinzione fra variabili continue, categoriche e binarie.

In riferimento alle variabili continue, per misurare la forza e la direzione della relazione, possiamo usare il coefficiente lineare di Pearson ed analizzare la relazione fra la variabile "price" e le variabili "sqft_living", "sqft_living_15" e "sqft_above".

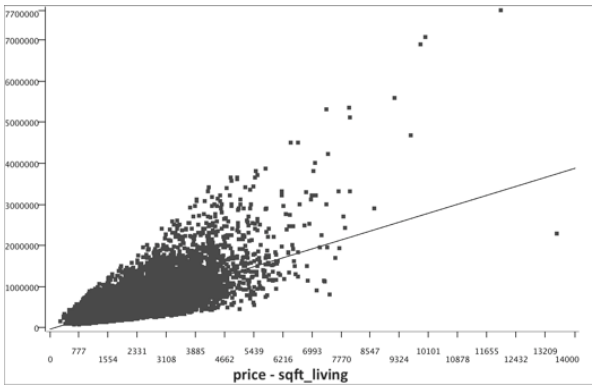

Fig. 5 – Scatterplot della regressione lineare fra "price" ed "sqft_living"

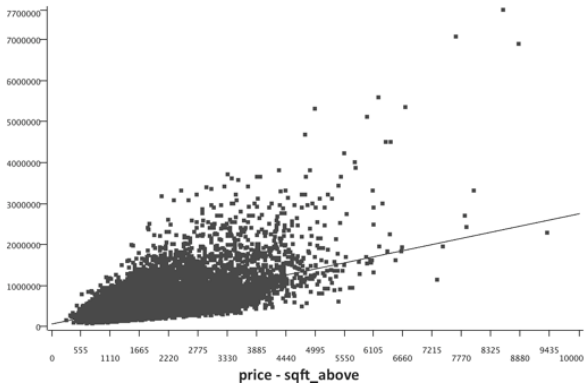

Fig. 6 – Scatterplot della regressione lineare fra "price" ed "sqft_above"

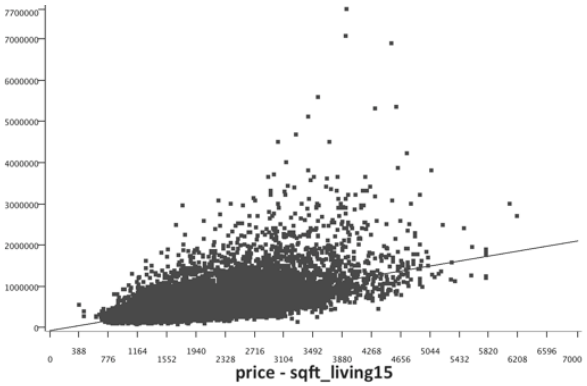

Fig. 6 – Scatterplot della regressione lineare fra "price" ed "sqft_living15"

Row ID	sqft_living	sqft_above	sqft_living15
sqft_living	1	0.877	0.756
sqft_above	0.877	1	0.732
sqft_living15	0.756	0.732	1

Fig. 7 – Matrice di correlazione fra le variabili "sqft_living", "sqft_above" ed "sqft_living15"

Le tre variabili, oltre ad essere fortemente correlate positivamente col prezzo, sono anche fortemente correlate positivamente tra di loro, come si può notare dalla matrice di correlazione associata (*Figura 7*).

Prendendo in considerazione le variabili categoriche, ad eccezione dell'unica variabile "condition", che risulta poco correlata col prezzo, le variabili "bathrooms", "bedrooms", "floors" e "views" sono moderatamente correlate col prezzo, con la variabile "grade" più correlata di tutte (*figura 4*).

In ultima analisi, l'unica variabile binaria presente nel dataset di origine è "waterfront", che risulta leggermente correlata col prezzo (0,266).
Tuttavia, questa bassa correlazione non viene riscontrata in un'analisi più approfondita di tale variabile in relazione al prezzo.
Infatti, il prezzo varia di molto in base alla presenza di un corso d'acqua che affianca la proprietà.
Una possibile causa di tale differenza è il fatto che la variabile "price" non risulta omogeneamente distribuita fra i due gruppi, infatti soltanto 163 case su 21.613 presentano la variabile uguale a 1.

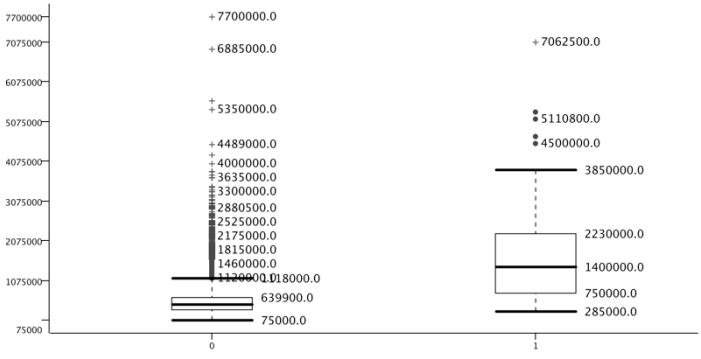

Fig. 8 – Boxplot di confronto fra la variabile "waterfont" e "price"

Dal momento che le variabili “sqft_basement” e “yr_renovated” presentano un gran numero di zeri, sono state create due variabili binarie “basement” e “renovated”, aventi valore 1 o 0. Il comportamento di tali variabili risulta diverso da quello di “waterfront”, siccome hanno una correlazione relativamente bassa con il prezzo senza condizionare eccessivamente la variazione del prezzo della casa.

In conclusione di questa prima analisi, possiamo affermare di aver analizzato dettagliatamente i tre gruppi di variabili presenti nel nostro dataset, evidenziando le relazioni più forti con la variabile del prezzo, che ci serviranno per la prossima analisi di correlazione lineare, per una prima costruzione del modello predittivo.

3. Analisi tramite vari classificatori

3.1 Analisi di regressione lineare

Il fine di questa analisi è quello di predire il prezzo delle case utilizzando i tool di regressione lineare di KNIME; basando il modello sulle considerazioni fatte in precedenza, si è ritenuto opportuno sfruttare le variabili “sqft_living” e “grade” in quanto sono le più correlate e la variabile “waterfront”, che risulta avere un impatto decisivo sul prezzo della casa.

KNIME mette a disposizione numerosi nodi di regressione lineare; il nostro workflow ha la finalità di testare tutti i nodi di regressione e valutare il modello migliore.

Dopo aver svolto tutte le operazioni di pre-processing necessarie, il dataset viene partizionato in training set e test set, in modo casuale con con una percentuale rispettivamente del 67% per il training set e del 33% per il test set.

Sono stati utilizzati 4 metodi di correlazione lineare, tutti appartenenti alla sezione “Weka” di KNIME : SMOreg (nodo che implementa la support vector machine per la regressione), MultiLayerPerceptron (un classificatore che usa il metodo di backpropagation per sottoporre le reti neurali all’interno di Weka ad apprendimento), SimpleLinearRegression (modello di regressione lineare univariata, dove viene selezionata quella tra le variabili indipendenti che ha minor errore quadratico) e LinearRegression (nodo di regressione lineare semplice).

Analizzando i seguenti modelli, riteniamo che i due modelli più significativi siano il MultiLayerPerceptron ed il LinearRegression.
Possiamo affermare che il modello MultiLayerPerceptron risulta avere il coefficiente di correlazione multipla (R^2) maggiore (0,612).
Tale coefficiente è uguale all’indice della bontà di adattamento, che spiega la variabilità totale della funzione target.
Per questo motivo può esser considerato come il modello che dovrebbe fornire una previsione più attendibile.
Possiamo notare come il grafico dei valori reali e quelli predetti segua un andamento lineare, ad eccezione di qualche predizione sconsiderata, che può esser tralasciata contando il gran numero di case che fanno parte dell’analisi (*figura 9*).
Il grafico del confronto fra i valori del prezzo ed i residui standardizzati (normalizzazione Z-Score), invece, mostra come la distribuzione dei residui, quindi i valori della differenza fra il prezzo reale e quello predetto, crescano all’aumentare del prezzo (*figura 10*).

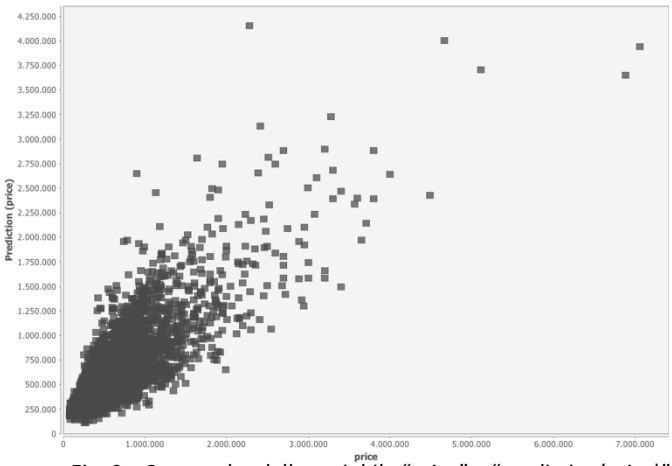

Fig. 9 – Scatterplot della variabile "price" e "prediction(price)" generato dal modello MultiLayerPerceptron

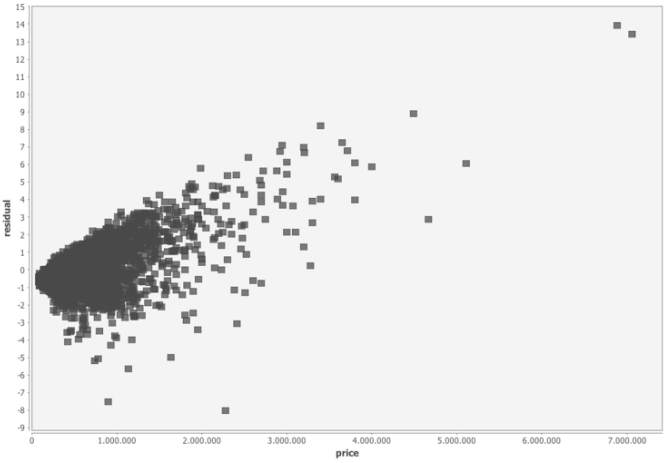

Fig. 10 – Scatterplot dei residui rispetto alla variabile "price"

Prendendo in considerazione l'analisi sul nodo LinearRegression, andiamo a mostrare l'output del nodo "Weka Predictor", che studia la variazione del prezzo predetto in base alle tre variabili prese in considerazione.

Price = -584954.9212 + 168.1871*sqft_living + 777701.5335*waterfront + 100336.879*grade

I quattro coefficienti di regressione parziale (computati dal modello) sono rispettivamente:

- 168.187, che rappresenta la variazione che subisce l'attributo target tenendo fisso le variabili indipendenti "waterfront" e "grade", ed aumentando di un'unità la variabile "sqft_living".
- 777701.5335, che rappresenta la variazione che subisce l'attributo target tenendo fisso le variabili indipendenti "sqft_living" e "grade", ed aumentando di un'unità la variabile "waterfront".
- 100336.879, che rappresenta la variazione che subisce l'attributo target tenendo fisso le variabili indipendenti "sqft_living" e "waterfront", ed aumentando di un'unità la variabile "grade".
- -584954.9212, rappresenta l'intercetta, cioè il valore che assume la variabile target quando le variabili indipendenti sono nulle.

Terminate le considerazioni fatte tramite l'ausilio di modelli basati sulla regressione lineare, procediamo con un'analisi alternativa, tramite la creazione di un modello di classificazione non binaria.

3.2 Analisi per classi

Essendo la variabile "price" di tipo continuo, le case possono essere raggruppate in classi dipendenti dal prezzo della casa; in questo modo è possibile effettuare una classificazione non binaria, costruendo un modello di predizione della classe (range di prezzo della casa presa in considerazione).

Nel nostro modello sono state create 5 classi:

- Classe 1, prezzo che varia da € 0,00 a € 300.000,00
- Classe 2, prezzo che varia da € 300.000,00 a € 400.000,00
- Classe 3, prezzo che varia da € 400.000,00 a € 500.000,00
- Classe 4, prezzo che varia da € 500.000,00 a € 700.000,00
- Classe 5, prezzo superiore a € 700.000,00

Le classi in questo modo assumono una distribuzione omogenea.

Il workflow realizzato comprende due nodi per la predizione della classe: Logistic e Naive Bayes.

Entrambi i nodi implementano la "k-folds cross validation"; consiste nella suddivisione del dataset totale in *k* parti di uguale numerosità.
Ad ogni passo, la k-esima parte del dataset viene ad essere il validation dataset, mentre la restante parte costituisce il training dataset; così, per ognuna delle *k* parti si allena il modello, evitando quindi problemi di overfitting, ma anche di campionamento asimmetrico del training dataset. In altre parole, si suddivide il campione osservato in gruppi di egual numerosità, si esclude iterativamente un gruppo alla volta e lo si cerca di predire; ciò al fine di verificare la bontà del modello di predizione utilizzato.

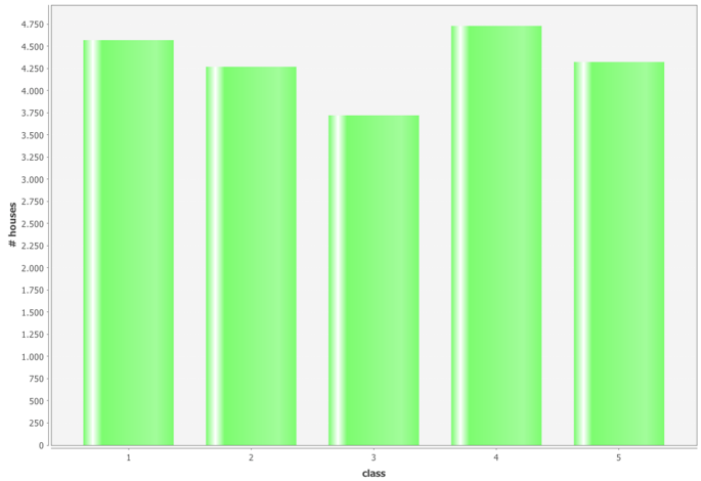

Fig. 11 – Grafico a barre che mostra la distribuzione della variabile classe nel dataset

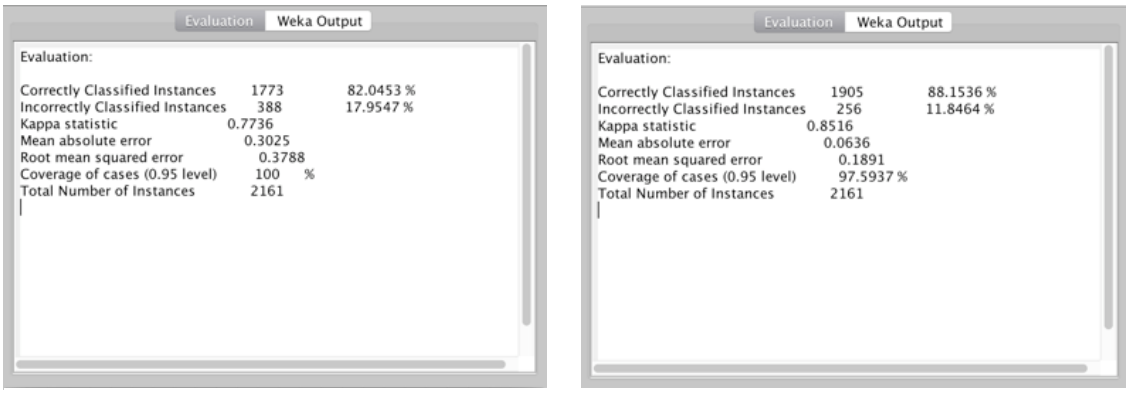

Fig. 12 – Confronto fra gli output di weka predictor dei modelli Logistic (sinistra) e Naïve Bayes (destra)

I due modelli vengono testati su un dataset filtrato, del quale sono state considerate solamente le variabili più correlate.
Logistic implementa il classificatore "MultiClassClassifier", mentre Naive Bayes usa il classificatore "NaiveBayes", tramite il quale risolve il problema della classificazione a più classi senza usare la trasformazione "1 against all" (come per Logistic).
L'analisi di accuratezza dei due nodi mostra come Naive Bayes sia più performante, ottenendo un punteggio di 88,153% contro 82,045% del nodo Logistic, come mostrano gli output dei nodi "weka predictor" dei due modelli (*figura 12*).

Di seguito, una dimostrazione di come la distribuzione della variabile Prediction (class) sia in linea con i valori reali.

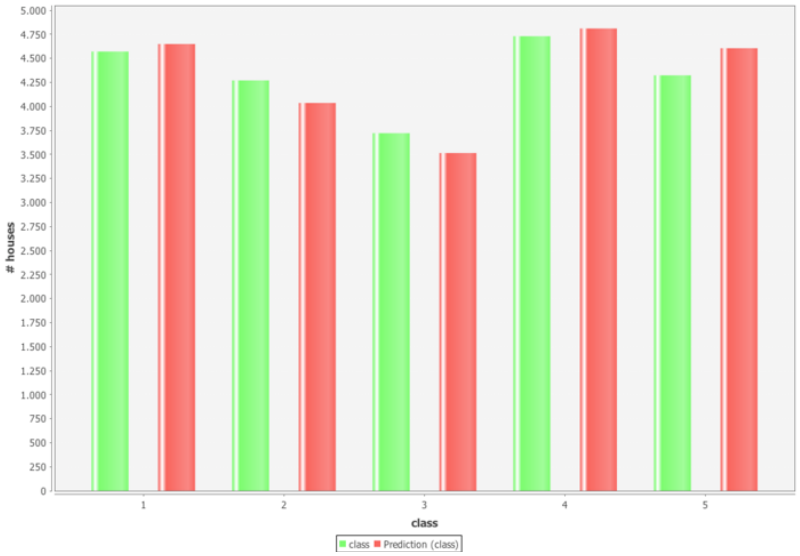

Fig. 13 – Grafico a barre che mostra di quanto varia la distribuzione nel dataset della variabile class e la sua predizione tramite il modello Naïve Bayes.

Il nodo Weka Predictor permette di aggiungere una colonna per ogni classe alla tabella di output, contenente la normalizzazione della classe per ciascuna distribuzione.
Tale funzionalità ci consente di sfruttare i grafici delle curve ROC, tramite il nodo di KNIME, i quali mostrano importanti informazioni riguardo l'accuratezza del modello di predizione della classe; una curva ROC disegna il numero dei record positivi inclusi in un subset selezionato sull'asse verticale, espressi come percentuale del numero totale di record positivi (%TP), contro il numero di record negativi inclusi nel subset, espressi come percentuale del numero totale di record negativi (%FP), sull'asse orizzontale.
Le curve ROC passano per i punti (0,0) e (1,1), avendo due condizioni che rappresentano le curve limite:
- la retta che taglia il grafico a 45° rappresenta il caso del classificatore casuale, la cui area sottostante ha il valore di 0,5.
- la linea che congiunge i punti (0,0), (0,1) ed (1,1) rappresenta il classificatore perfetto, la curva la cui area sottostante è pari a 1.

L'accuratezza viene misurata dall'area sottostante la curva, denominata AUC (Area under curve); di seguito verranno mostrati due grafici rappresentanti la classe meglio e meno predetta dal classificatore Logistic (quindi i valori maggiori e minori delle aree sottostanti la curva ROC), in quanto i valori AUC del modello Naive Bayes sono tutti superiori a 0,98.

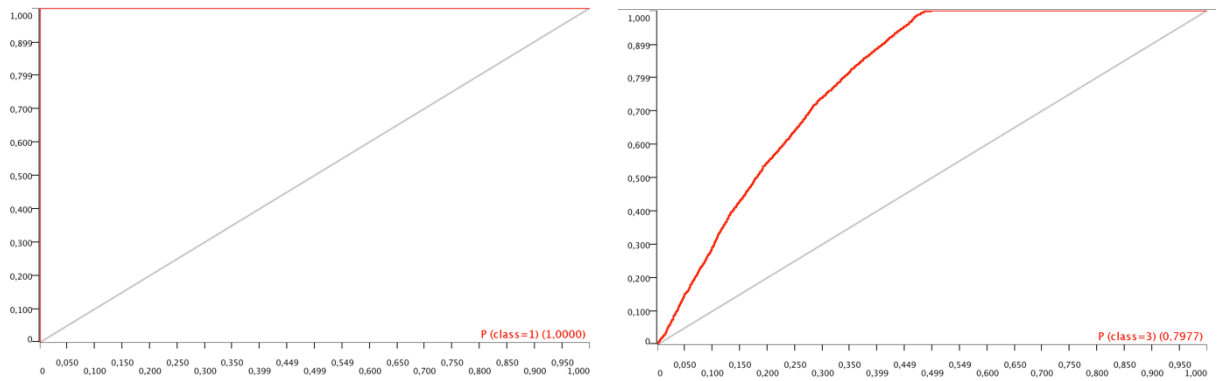

Fig. 14 – ROC curve rappresentante la predizione della classe 1 e 3 del modello Logistic.

Per il modello Logistic, le classi 1 e 5 hanno il valore AUC perfetto 1.0. Per quanto riguarda la classe 3, il valore AUC è di 0,79, un risultato accettabile e non totalmente negativo, che però condiziona inevitabilmente l'accuratezza generale del modello.

4. Conclusioni

Tramite il seguente elaborato è stato introdotto ed analizzato il workflow costruito su KNIME; inizialmente sono state applicate varie operazioni di preprocessing sul dataset di origine, che abbiamo ritenuto affidabile e completo, in quanto privo di dati mancanti.
Da una prima analisi di correlazione dei dati, abbiamo riscontrato un basso numero di variabili fortemente correlate con la variabile obiettivo target, le quali sono state utilizzate nell'analisi di regressione lineare e nell'analisi per classi.

I risultati dei modelli di predizione del prezzo costruiti hanno dimostrato complessivamente una buona accuratezza e quindi un basso margine di errore.

Possiamo, dopo questa applicazione di tecniche di machine learning, concludere che:

- La variabile "price" delle case può essere predetta tramite un'analisi di regressione lineare, ottenendo risultati soddisfacenti. Le predizioni più complete ed affidabili vengono effettuate dai nodi MultiLayerPerceptron (un classificatore che utilizza il metodo di backpropagation per classificare le istanze) e LinearRegression (nodo che implementa una regressione lineare semplice).
- La variabile "price" può essere raggruppata in classi (determinate dalla fascia di prezzo delle case) sulle quali possono essere applicati dei modelli di predizione come il Logistic o Naive Bayes. Abbiamo constatato che il modello Naive Bayes è più completo ed affidabile, un comportamento che viene sottolineato dalla distribuzione della variabile predetta e dalle curve ROC delle singole classi predette.

Si potrebbero apportare molti miglioramenti a questo studio, estendendo l'analisi sui modelli di regressione, quindi implementando nuove tecniche di regressione più complesse.

Inoltre, abbiamo notato come la variabile latitudine presenti una leggera correlazione col prezzo; infatti, più ci si sposta verso nord, più aumenta il prezzo.
Tale informazione può esser sfruttata per effettuare un'analisi geografica più approfondita; a parer nostro, la componente geografica influisce sul prezzo di una casa e può essere utile analizzarla attraverso strumenti che non abbiamo sfruttato in questo elaborato.

5. Riferimenti

Kaggle dataset: https: //www.kaggle.com/harlfoxem/housesalesprediction.

Lezioni di KNIME su e-learning: http://elearning.unimib.it/course/view.php?id=17327.

Informazioni geografiche sulla contea di King: https://en.wikipedia.org/wiki/King_County,_Washington.