Analisi del mercato immobiliare di Seattle

Longo Gloria 864579, Alberto Porrini 826306, Luca Iarocci 894066, Roberto Ferrari 852220

Abstract
La domanda e l'offerta degli immobili americani è da anni ormai in continua crescita a causa dell'aumento della popolazione che si è sviluppato nell'ultimo decennio. Ciò porta alla naturale conseguenza di un aumento esponenziale dei prezzi delle abitazioni che diventano sempre più difficili da stimare e categorizzare. L'obiettivo di questo progetto è quello di capire, tramite l'addestramento di alcuni modelli statistici, se è possibile giungere a conclusioni di ramo economico e di natura predittiva riguardanti il prezzo delle case vendute nella King County area, partendo da dati storici delle vendite e mettendo in relazione le caratteristiche degli immobili e il prezzo di vendita . In secondo luogo invece si vuole condurre un operazione di clustering delle abitazione aventi caratteristiche comuni con lo scopo di scoprire se esse corrispondono a differenti aree della contea con determinate caratteristiche.

Keywords
Machine Learning — Real Estate — Clustering

Contents

Introduzione 1
1 Descrizione del dataset 1
1.1 Features selection . . . 2
2 Modelli di regressione 2
2.1 Holdout . . . 3
2.2 Cross Validation . . . 3
3 Valutazione delle performance 3
3.1 R^2 . . . 3
3.2 Root Mean Square Error . . . 3
4 Risultati e analisi 3
5 Clustering 4
5.1 Clusterizzazione Gerarchica . . . 4
5.2 K-means and K-medoids . . . 5
6 Validazione 5
6.1 Indici interni o non supervisionati . . . 5
6.2 Test Formale . . . 6
7 Analisi dei risultati 6
8 Analisi delle componenti principali 6
9 Analisi dei risultati dopo PCA 7
10 Conclusioni e sviluppi futuri 7
References 8

Introduzione

Il costo degli immobili nello stato del Washington è in continua crescita, basti pensare che nel 2021 il prezzo mediano delle case in vendita è incrementato del 23.9 % rispetto al 2020 [1].

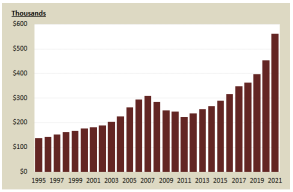

Figure 1. Andamento del prezzo mediano delle case vendute nello stato del Washington dal 1995 al 2021

I fattori che ne influenzano il valore sono molteplici, a partire dal numero di locali al suo interno fino ad arrivare all'anno di costruzione e al quartiere in cui sono situate. Grazie a questo fenomeno, aumenta sempre di più l'esigenza di sviluppare modelli statistici e algoritmi di apprendimento automatico in grado di facilitare e accelerare tutte le operazioni legate al mondo del real estate. Gli scopi di quest'analisi sono infatti l'implementazione di modelli statistici in grado di determinare il prezzo degli immobili partendo dalle loro caratteristiche fisiche e l'individualizzazione di alcuni cluster geografici che possano mettere in relazione le varie aree della contea accomunate da immobili dalle caratteristiche simili.

1. Descrizione del dataset

Per lo sviluppo di questa ricerca è stato utilizzato un dataset di 21613 osservazioni e 21 attributi estratto dalla piattaforma pubblica di Kaggle [2]. Esso contiene le informazioni delle case vendute nella King County area, Washington State da Maggio 2014 a Maggio 2015. Verrà in seguito illustrata un'analisi più approfondita degli attributi contenuti:

- Id: codice univoco per ogni casa venduta
- Data: data della vendita della casa
- Price: prezzo della casa venduta

- Bedrooms: numero di camere da letto
- Bathrooms: numero di bagni
- Sqft living: metratura dello spazio abitativo interno
- Sqft lot: metratura dello spazio catastale
- Floors: numero di piani
- Waterfront: variabile binaria per stabilire se l'appartamento affacciava o meno sul mare
- View: indice da 0 a 4 che indica quanto è bella la vista dall'abitazione
- Condition: indice da 1 a 5 sullo stato dell'immobile
- Grade: indice da 1 a 13, dove 1-3 rappresenta una qualità inferiore riguardante la costruzione e progettazione di edifici, 7 ha un livello medio di costruzione e progettazione e 11-13 ha un livello di alta qualità di costruzione e progettazione
- Sqft above: metratura dello spazio abitativo interno che si trova sopra il livello del suolo
- Sqft basement: metratura dello spazio abitativo interno che si trova sotto il livello del suolo
- Yr built: anno di costruzione della casa
- Yr renovate: anno dell'ultima ristrutturazione o 0 se non è mai stata ristrutturata
- Zipcode: In quale codice postale è situata
- Latitude: latitudine
- Longitude: longitudine
- Sqft living 15: metratura dello spazio abitativo interno per i 15 vicini più prossimi
- Sqft lot 15: metratura dei lotti di terreno dei 15 vicini più prossimi

1.1 Features selection

Il dataset scelto per lo sviluppo di questo progetto non contiene nessun valore nullo, questo ha consentito dunque di conservare tutte le osservazioni senza doverne escludere alcuna a priori. In seguito è stata attuata un'analisi riguardate la correlazione delle variabili esplicative con la variabile dipendente rappresentata in seguito attraverso una matrice di correlazione. Dall'analisi sono state esclude le variabili *Id* dal momento che contiene un codice identificativo differente per tutte le osservazioni e *Date* che contiene la data in cui la casa è stata venduta, ritenuta non rilevante dal momento che il dataset contiene le vendite di un singolo anno.

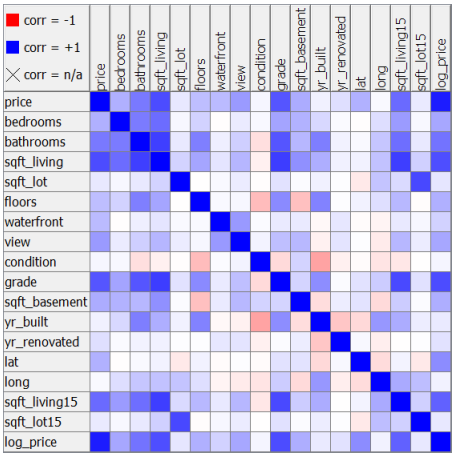

Figure 2. Correlazione

La variabile *Sqft above* risulta avere una correlazione molto elevata dello 0.6 con la variabile *Price*, questo risultato però potrebbe essere dovuto ad una dipendenza con la variabile *Sqft living*. Infatti *Sqft above* (metratura dello spazio abitativo interno che si trova sopra il livello del suolo) sembra essere la differenza tra *Sqft living* (metratura dello spazio abitativo interno) e *Sqft basement* (metratura dello spazio abitativo interno che si trova sotto il livello del suolo), a causa di questa dipendenza lineare, è stata esclusa la variabile *Sqft above* dalle analisi di regressione.

Per ottenere dei risultati più affidabili e più precisi, è stata classificata la variabile contente l'anno dell'ultima ristrutturazione della casa *Yr renovate* in una variabile binaria avete come valore 0 se la casa non ha subito ristrutturazioni o 1 altrimenti. Questa scelta è stata fatta dal momento che solamente 914 case su 21613 hanno subito una ristrutturazione, dunque il 4,23% .

Le variabili *Grade*, *Yr built* e *Condition* sono state decifrate come variabili numeriche e non come variabili qualitative per evitare di riscontrare problemi con le future operazione di cluster. Questa scelta non ha portato variazioni negli indici di valutazione delle performance.

2. Modelli di regressione

Sono stati implementati diversi modelli di regressione per il valore monetario degli immobili. Essi verranno di seguito elencati con i rispettivi nomi dei nodi di Knime:

- *Simple Regression Tree Learner*: è un algoritmo euristico, dunque nonostante non garantisca di giungere a risultati ottimali, permette di ottenere soluzioni approssimate e ragionevoli.
- *Random Forest Learner (Regression)* : anch'esso un algoritmo euristico
- *Linear Regression based on Weka 3.7*: algoritmo di regressione che permette di stimare le relazioni tra variabili. Determina il legame funzionale tra le variabili

indipendenti e la variabile dipendente. Questo nodo a differenza di quello di default di Knime accetta anche variabili categoriche con molte classi come ad esempio lo Zipcode

Per tutti i seguenti modelli sono state utilizzate le tecniche di *Holdout* e di *Cross validation*

2.1 Holdout

Nello sviluppo di algoritmi di Machine Learning, è di fondamentale importanza valutare il modello per stimare le capacità di generalizzazione dei pattern appresi durante l'addestramento. Tuttavia, poiché le istanze future hanno valori di destinazione sconosciuti e non è possibile verificare ora l'accuratezza delle previsioni per esse, è necessario utilizzare alcuni dei dati di cui si conosce già la risposta come proxy per i dati futuri. Una strategia comune consiste nell'utilizzare tutti i dati etichettati disponibili e frazionarli in sottoinsiemi per l'addestramento e la valutazione, di solito con un rapporto del 70-80 % per l'addestramento e del 20-30 % per la valutazione, questa tecnica è chiamata *Holdout*. Data la grande quantità di dati a nostra disposizione, per questo progetto si è optato per una divisione dell'80% per il training set e 20% per il test set. In questo modo il dataset di training sarà utilizzato per addestrare i modelli mentre il dataset di test sarà utilizzato per valutare la qualità predittiva del modello ottenuto.

2.2 Cross Validation

La tecnica del *k-fold cross validation* ha come input un unico parametro k che si riferisce al numero di gruppi in cui deve essere suddiviso un dato campione di dati. La convalida incrociata, detta *k-fold*, consiste nella suddivisione dell'insieme di dati totale in k parti di uguale numerosità e, a ogni passo, la k^a parte dell'insieme di dati viene a essere quella di convalida, mentre la restante parte costituisce l'insieme di addestramento. Così si allena il modello per ognuna delle k parti, riducendo il problema di overfitting, ma anche di campionamento asimmetrico (e quindi affetto da distorsione) del campione osservato, tipico della suddivisione dei dati in due sole parti (ossia addestramento/convalida). I risultati di un test di cross-validazione k-fold sono spesso riassunti con la media dell'indice di performance utilizzato per valutare il modello. Il valore attribuito alla variabile k per quest'analisi è pari a 5.

3. Valutazione delle performance

Dopo aver addestrato un algoritmo predittivo è importante valutarne la sua performance attraverso degli indicatori che tengano conto del tipo di approccio utilizzato e degli obiettivi da raggiungere. Per questa analisi sono state utilizzate le metriche R^2 e *Root Mean square error*

3.1 R^2

Il coefficiente di determinazione, più comunemente R^2, è un indice che misura il legame tra la variabilità dei dati e la correttezza del modello statistico utilizzato. Intuitivamente, esso è legato alla frazione della varianza non spiegata dal modello, infatti si calcola con la seguente espressione:

$$R^2 = 1 - \frac{RSS}{TSS},$$

dove:

- $RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ è la devianza residua (Residual Sum of Squares);
- $\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ è la devianza totale (Total Sum of Squares);

- y_i sono i dati osservati;
- $\bar{y}$ è la loro media;
- $\hat{y}_i$ sono i dati stimati dal modello.

3.2 Root Mean Square Error

La radice dell'errore quadratico medio (Root Mean Square Error, RMSE) indica la radice della discrepanza quadratica media fra i valori dei dati osservati ed i valori dei dati stimati. Questo indice fornisce dunque una misura per giudicare la qualità di uno stimatore in termini della sua variazione e della sua distorsione per questo deve essere il più basso possibile.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n}}.$$

4. Risultati e analisi

Verranno illustrati in questo paragrafo i risultati ottenuti

dell'analisi della prima domanda di ricerca posta inizialmente, verranno paragonati quindi i vari valori degli indici di performance ottenuti dai tre differenti modelli predittivi per la variabile output *Price*. Durante lo sviluppo del progetto si è scelto di utilizzare come variabile dipendente il logaritmo naturale del prezzo degli immobili piuttosto che il prezzo nel suo valore assoluto per consentire di cogliere meglio la differenza tra valore osservato e valore atteso in rapporto al valore osservato.

Nella fig. 3 sono rappresentati i valori dell'RMSE calcolato sulla variabile output *Price* dei tre modelli tramite boxplot. Essendo l'RMSE la radice della media della somma delle distanze al quadrato tra valore atteso e valore osservato, possiamo usare come metro di paragone la media del prezzo delle abitazioni che è di circa 540088,14$. Osservando il grafico si può subito notare che il Random Forest è il modello che da i risultati migliori in termini di radice dell'errore quadratico medio dal momento che possiede il valore di circa 138489,19$ che paragonato alla media del prezzo degli immobili è un valore molto promettente. Anche il Linear regression ha registrato un valore piuttosto rilevante con un RMSE di circa 162102.22 mentre il Decision Tree se comparato agli altri due modelli non ha ottenuto delle buone performance con un RMSE molto alto di circa 182497,66.

Nella fig. 4 invece, vengono riportati i valori dell'RMSE calcolato sul logaritmo naturale della variabile *Price* dei tre

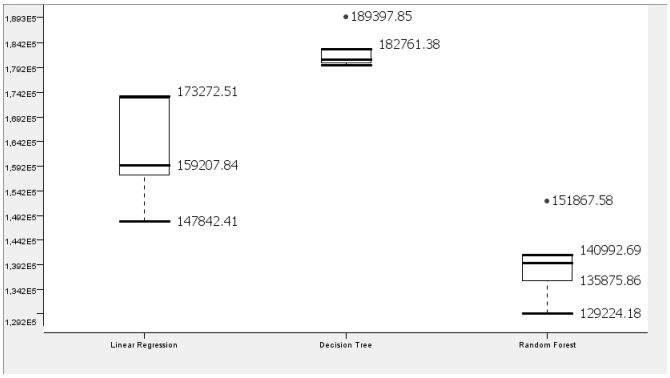

Figure 3. Confronto RMSE price

modelli predittivi. Si nota subito un miglioramento delle performance del modello lineare che raggiunge valori molto simili a quelli del Random Forest. Ad ogni modo il Decision Tree rimane il modello con le performance più basse dal momento che registra sempre valori molto alti rispetto agli altri modelli.

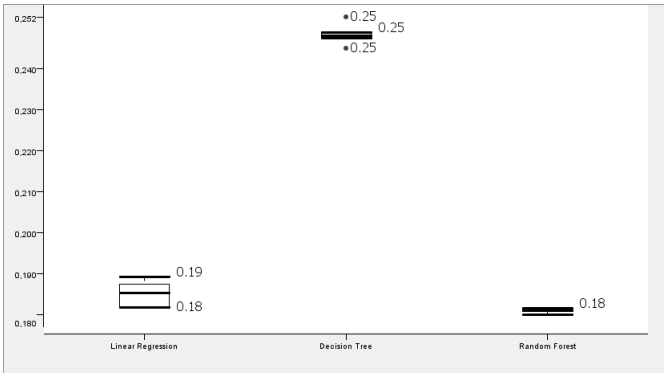

Figure 4. Confronto RMSE log price

In fig. 5 vengono confrontati i valori dell'indice di performance R^2 dei tre diversi modelli e delle due diverse modalità della variabile predittiva. Si nota subito come il modello Decision Tree è quello che ha le peggiori performance in termini di R^2 comparato agli altri due con un indice dello 0.75 per la variabile *Price* e dello 0.77 per il logaritmo della variabile *Price*. Il Linear regression invece ha ottenuto dei valori molto promettenti con un R^2 dello 0.805 per la variabile *Price* e dello 0.876 per il logaritmo della variabile *Price*. Infine il Random Forest è il modello che ha ottenuto i valori più alti, con 0.855 per la variabile *Price* e 0.882 per il logaritmo naturale.

In conclusione dunque il modello predittivo Random Forest risulta il migliore in termini di performance sia parlando di R^2 che in termini di RMSE. I risultati inoltre, confermano che fare previsione sul logaritmo naturale della variabile dipendente da dei risultati più affidabili paragonati alla variabile dipendente nel suo valore assoluto.

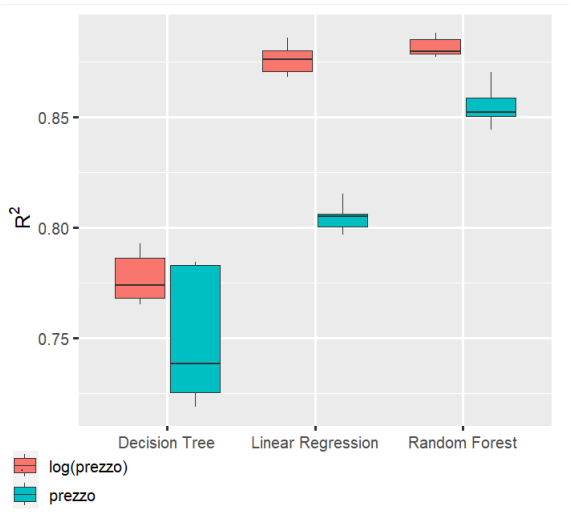

Figure 5. Confronto R^2

5. Clustering

Questa seconda sezione dell'elaborato ha come obiettivo quello di classificare gli zipcode contenenti abitazioni con caratteristiche simili tra loro attraverso la procedura di clustering. Si vuole cercare quindi di creare dei gruppi omogenei all'interno ed eterogenei tra di loro con lo scopo di indagare se i differenti cluster corrispondono a zone differenti della King County Area [3]. Verranno dunque illustrate le procedure che hanno portato alla scelta e alla costruzione di cluster per poi concludere con una visualizzazione grafica di quest'ultimi per rappresentare meglio i risultati dell'analisi.

Per lo sviluppo dell'analisi, sono state escluse tutte le variabili che potessero dare un'informazione relativa alla posizione geografica della casa, esse sono: *Lat*, *Long*, *Zipcode*, *Sqft living 15*, *Sqft lot 15*

Le variabili selezionate sono state dapprima normalizzate in un intervallo che va da 0 a 1 e secondariamente sono state raggruppate secondo la mediana.

5.1 Clusterizzazione Gerarchica

In questo progetto viene fatto uso di algoritmi di clustering gerarchici agglomerativi o bottom up; questa classe di algoritmi assegna inizialmente un cluster differente ad ogni osservazione e, in un processo iterativo, combina a due a due i cluster più simili tra loro fino a ricondurre tutte le osservazioni ad un unico cluster. La similarità tra due cluster dipende da due fattori: dalla metrica utilizzata per determinare la distanza tra due osservazioni e dal criterio di collegamento (linkage) scelto. Al fine di determinare il criterio ottimale per il caso in analisi sono stati confrontati tra loro i risultati ottenuti da quattro diversi criteri di linkage: single, complete, average e Ward. La valutazione dei clustering così ottenuti è avvenuta tramite il calcolo del Cophenetic Correlation Coefficient (il cui valore va massimizzato), portando dunque alla scelta del criterio di collegamento medio (average linkage) secondo il quale la distanza tra due cluster è calcolata come la media

della distanza di tutte le possibili coppie di osservazioni appartenenti ai cluster. I risultati ottenuti fanno utilizzo della distanza euclidea in qualità di metrica.

5.2 K-means and K-medoids

Altre due tecniche di clustering utilizzate in questo progetto sono k-medoids, realizzata tramite l'algoritmo PAM (Partition Around Medoids), e k-means. Si tratta di tecniche in linea di principio molto simili e appartenenti alla classe degli algoritmi partizionali, portando dunque ad un'assegnazione univoca di ogni punto dello spazio delle variabili ad un unico cluster. Entrambi operano secondo uno schema iterativo partendo da un set di k punti scelti casualmente nello spazio delle variabili (o nell'insieme delle osservazioni nel caso del PAM), ognuno di questi funge da centro del k-esimo cluster a cui vengono associate tutte le osservazioni più vicine secondo la metrica utilizzata (nel contesto, coerentemente a quanto scelto per il caso gerarchico, si è optato per la distanza euclidea). La posizione dei centri viene progressivamente modificata con l'obiettivo di minimizzare l'errore quadratico medio, ovvero la distanza tra punti assegnati ad un determinato cluster e il suo centro. Gli algoritmi terminano nei seguneti casi: o quando l'assegnazione del cluster non varia più, o raggiunta una soglia prefissata di iterazioni (nelle applicazioni presentate questa è stata fissata a 100). La differenza principale tra k-means e k-medoids riguarda la natura dei centri che nel caso k-means risultano essere punti fittizi, la cui posizione nello spazio delle variabili è la media delle componenti dei punti appartenenti al cluster, mentre nel caso k-medoids risultano sempre essere punti appartenenti al dataset di partenza. Questo fa sì che da un punto di vista teorico k-means risulti più robusto rispetto a rumore e possibili outlier. I due algoritmi condividono anche due difetti:

- il numero dei cluster da realizzare deve essere specificato a priori
- l'inizializzazione della posizione dei centri può influenzare il risultato del clustering

Una soluzione al problema del numero ottimale di clustering è data dell'utilizzo di indici interni, mentre per quanto riguarda l'inizializzazione si è optato per ripetere la clusterizzazione n volte con inizializzazioni diverse e selezionando il risultato che minimizza l'errore quadratico medio (nello specifico n è stato fissato a 25 producendo risultati consistenti nel corso dei test).

6. Validazione

Gli algoritmi di clusterizzazione utilizzati in questo progetto hanno lo scopo di classificare le aree geografiche della contea in diversi gruppi con caratteristiche simili tra loro. Per determinare quale tra questi sia il migliore e quanti cluster realizzare, è necessario utilizzare degli indici di valutazione in grado di calcolare la bontà e la precisione dell'algoritmo implementato. Tenendo in considerazione la natura delle tecniche utilizzate, si è ritenuto opportuno utilizzare indici interni, o non supervisionati in quanto non sussistono ipotesi a priori sulla struttura dei dati. A seguire i risultati ottenuti sono stati sottoposti a un test formale per determinarne la significatività statistica.

6.1 Indici interni o non supervisionati

Gli indici interni di valutazione misurano la bontà dei risultati ottenuti da un clustering senza l'ausilio di informazioni esterne. In questo progetto sono stati presi in considerazione i seguenti:

- Il coefficiente di *Silhouette*, definito come:

$$Si = \frac{b_i - a_i}{\max(a_i, b_i)} \in [-1, +1]$$

  dove

  - a_i corrisponde alla distanza media dell'i-esima osservazione rispetto alle altre osservazioni del cluster;
  - b_i corrisponde alla distanza media dell'i-esima osservazione rispetto alle osservazioni di ogni cluster che sia diverso dal cluster di partenza.

  Valori positivi e quanto più prossimi a 1 sono indicativi di una buona riuscita del clustering

- Il coefficiente di *Dunn*, definito in [0, + inf] come il rapporto tra la minima distanza di osservazioni non appartenenti allo stesso cluster e la massima distanza tra osservazioni dello stesso cluster

$$\text{Dato un clustering } \mathscr{C} = \{C1, \ldots, Cn\}$$

$$D(\mathscr{C}) = \frac{\min_k, C_l \in \mathscr{C}, C_k \neq C_l \left(\min_{i \in C_k, j \in C_l} \text{dist}(i, j)\right)}{\max_{C_m \in \mathscr{C}} \text{diam}\,(C_m)}$$

  Valori più alti del coefficiente di *Dunn* sono indicativi di una buona riuscita del clustering

- Il coefficiente di *Connectivity*, definito in [0, +infinito] come:

$$\text{Conn}(\mathscr{C}) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i, nn_{i(j)}},$$

  dove

  - N corrisponde al numero totale di osservazioni;
  - L corrisponde al parametro che indica il numero dei nearest neighbour da utilizzare;
  - $nn_{i\,(j)}$ corrisponde al j-esimo nearest neighbour della i-esima osservazione.
  - $x_i, nn_{(j)} = 0$ se i e $nn_{(j)}$ appartengono allo stesso cluster e 1/j altrimenti

  Valori più bassi del coefficiente di connectivity sono indicativi di una buona riuscita del clustering.

6.2 Test Formale

Al fine di vagliare la validità statistica dei risultati ottenuti è stato eseguito un test di ipotesi avente come ipotesi nulla l'assenza di una struttura nei dati analizzati. Questo passaggio ha importanza da un punto di vista formale in quanto gli algoritmi di clustering generano una suddivisione delle osservazioni a prescindere dalla loro distribuzione, che può potenzialmente essere del tutto casuale e non presentare alcuno schema. Nello specifico i risultati ottenuti in termini di indice di silhouette dagli algoritmi di clustering sul dataset in analisi sono stati confrontati con quelli ottenuti da un dataset fittizio generato tramite una simulazione di Monte Carlo.

Per questo test si è scelto il quantile di livello 0,99 (significatività alfa= 0,01).

7. Analisi dei risultati

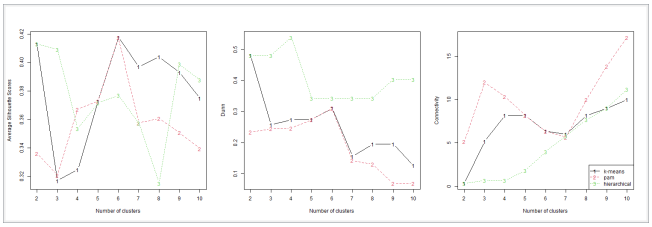

Figure 6. Silhuoette Dunn Connectivity

La valutazione degli indici interni sull'intervallo da 2 a 10 cluster evidenzia una tendenza simile rispetto ai tre indici da parte degli algoritmi k-means e PAM, con il primo che tende a performare meglio del secondo nella quasi totalità dei casi. Partendo dall'indice di Silhouette si osserva come K-means e PAM assumano il valore massimo in 6, l'algoritmo gerarchico fa registrare il miglior risultato nel caso di solo 2 cluster mostrando comunque un massimo relativo in corrispondenza di 6. In termini di indice di Dunn la tecnica di clustering gerarchica risulta favorita; l'indice registra i valori massimi in corrispondenza di un clustering a 2 e 4 per k-means e gerarchico rispettivamente, mentre per PAM l'indice assume il suo apice in 6 dove si osserva un massimo relativo per k-means. Anche in termini di connectivity la tecnica di clustering gerarchica ottiene risultati migliori ad eccezione del caso a 10 cluster. L'indice mostra un andamento non decrescente per il clustering gerarchico, risultano invece più interessanti le curve tracciate da k-means e PAM che mostrano un avvallamento in corrispondenza di 6 e 7. In conclusione si è optato per realizzare 6 cluster tramite la tecnica k-means. La scelta è motivata dal massimo assoluto registrato in corrispondenza di 6 cluster dell'indice di Silhouette per k-means e PAM e dalla generale tendenza, così come dai risultati paragonabili seppur marginalmente peggiori osservati per Dunn e connectivity rispetto al clustering gerarchico.

k=6	Silhouette	Dunn	Connectivity
K-means	0.537	0.196	3.667
PAM	0.532	0.089	5.333
Hierarchical	0.521	0.262	2.667

Figure 7. Valori degli indici

La valutazione tramite test di ipotesi del clustering k-means a 6, facendo registrare un valore medio di silhouette (0,42) maggiore rispetto al quantile di ordine 0,99 ottenuto tramite simulazione di Monte Carlo (0,14), porta al rifiuto dell'ipotesi nulla di assenza di struttura dei dati e avvalora i risultati ottenuti fig. 8

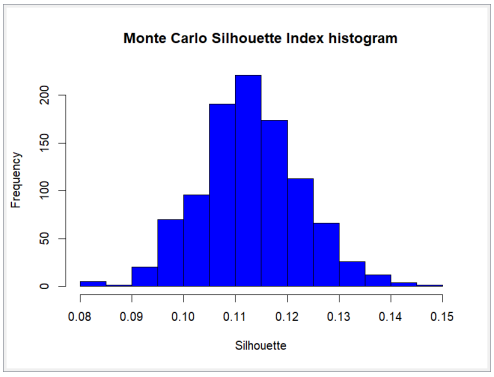

Figure 8. Test ipotesi vs Monte Carlo

8. Analisi delle componenti principali

L'analisi delle componenti principali, o PCA, è un metodo di riduzione della dimensionalità che viene spesso utilizzato per semplificare i dati, trasformando le variabili di partenza in un insieme piccolo a piacere di variabili artificiali non correlate contenenti la maggior frazione possibile delle informazioni dell'insieme di partenza.[4] Tale risultato è ottenuto a partire dalla scomposizione spettrale della matrice delle covarianze delle variabili di partenza che restituisce un set di altrettante direzioni ortogonali (autovettori) i cui autovalori associati rappresentano l'ammontare della variabilità totale rappresentata dalle rispettive direzioni. Nel caso in analisi l'applicazione di questo metodo ha un duplice effetto in quanto: da una parte facilita l'addestramento degli algoritmi di apprendimento automatici e dall'altra bilancia, nel caso specifico del clustering, il rapporto tra variabili (originariamente 10) e osservazioni in analisi (70).

Per l'implementazione dell'analisi delle componenti principali sono stati utilizzati, a seguito di una normal-

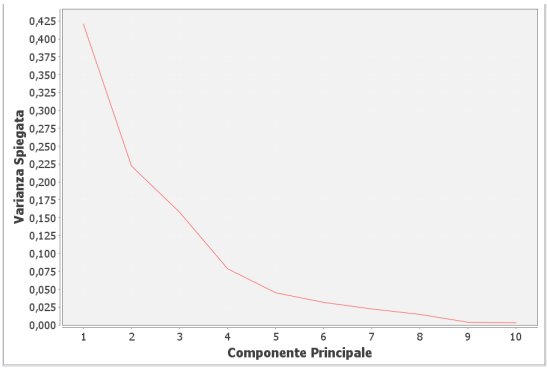

Figure 9. Analisi delle Componenti Principali

izzazione tra 0 e 1 tramite nodo *normalizer*, il nodo di knime *PCA Compute* portando al seguente risultato in fig. 9. Non c'è un metodo univoco per stabilire il numero preciso di componenti principali da conservare; di seguito se ne elencano due:

- Scree-plot: grafico degli autovalori (o varianza spiegata) in funzione del numero di PC; poiché gli autovalori sono decrescenti, il grafico assume l'aspetto di una spezzata con pendenza negativa. Una variazione di pendenza significativa indica
- il numero di componenti da tenere in considerazione. Quota di varianza totale spiegata: si scelgono le componenti principali che tengono conto di una quota prefissata sufficientemente elevata di varianza totale spiegata

Per lo sviluppo di quest'analisi si è scelto di estrarre le prime quattro componenti principali, che spiegano all'incirca l' 88% della variabilità dei dati di partenza.

9. Analisi dei risultati dopo PCA

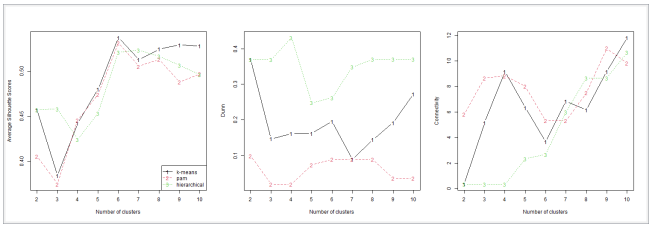

Figure 10. Silhuoette Dunn Connectivity dopo PCA

Analizzando l'indice di silhouette si osserva come i tre algoritmi mostrano un andamento simile al variare del numero di cluster con K-means e PAM che assumano in 6 valore massimo, e l'algoritmo gerarchico che fa registrare il suo apice in corrispondenza di 7 cluster, ottenendo comunque in 6 il suo secondo miglior risultato. In termini di indice di Dunn la tecnica di clustering gerarchica è risultata nuovamente favorita; l'indice registra, al pari del caso precedentemente studiato, valori massimi in corrispondenza di un clustering a 2 e 4 per k-means e gerarchico rispettivamente; l'algoritmo PAM assume il suo massimo nell'intervallo 6-8. In 6 si osserva anche in questo caso un massimo relativo per k-means. In termini di connectivity la tecnica di clustering gerarchica mostra un andamento non decrescente, e le curve tracciate da k-means e PAM mostrano un avvallamento in corrispondenza di 6. Emerge una tendenza da parte degli algoritmi in analisi a registrare valori di Dunn e silhouette paragonabili al massimo del rispettivo indice in corrispondenza dell'intervallo 7-10 cluster. Questo tipo di soluzioni, nonché sfavorite dall'indice di connectivity, conducono a cluster di numerosità ridotta (inferiore a 3 elementi già nel caso di 7 cluster) e sono state scartate in quanto ritenute meno significative al fine di costruire un clustering del mercato immobiliare capace di generalizzare le caratteristiche degli elementi che compongono le proprie classi. In conclusione si è optato per realizzare 6 cluster tramite la tecnica gerarchica, che registra in corrispondenza di tale numero di cluster valori paragonabili agli altri algoritmi in termini di silhouette e valori migliori in termini di Dunn e connectivity.

k=6	Silhouette	Dunn	Connectivity
K-means	0.537	0.196	3.667
PAM	0.532	0.089	5.333
Hierarchical	0.521	0.262	2.667

Figure 11. Valori degli indici dopo PCA

Anche in questo caso la valutazione tramite test di ipotesi del clustering gerarchico a 6 sviluppata sulle prime 4 componenti principali del dataset, fa registrare un valore medio di silhouette (0,52) maggiore rispetto al quantile di ordine 0,99 ottenuto tramite simulazione di Monte Carlo (0,28), portando così al rifiuto dell'ipotesi nulla di assenza di struttura dei dati e rafforza i risultati ottenuti. È possibile infine visualizzare la clusterizzazione gerarchica ottenuta tramite un dendrogramma dandoci la possibilità non solo di valutarne visivamente la qualità, ma anche di comprendere le relazioni tra i vari cluster nonché la loro struttura interna.

10. Conclusioni e sviluppi futuri

Grazie ai metodi di cluster utilizzati abbiamo ottenuto sei gruppi eterogenei tra loro ed omogenei al loro interno che ci hanno permesso di giungere a conclusioni concrete e di ramo economico. In fig. 13 sono rappresentati i diversi radar chart dei cluster ottenuti con il metodo kmeans.

Un'ulteriore visualizzazione e riconferma della validità

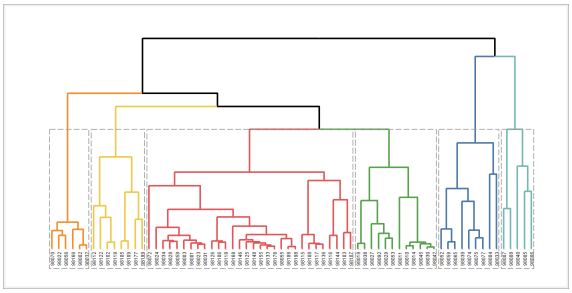

Figure 12. Dendrogramma

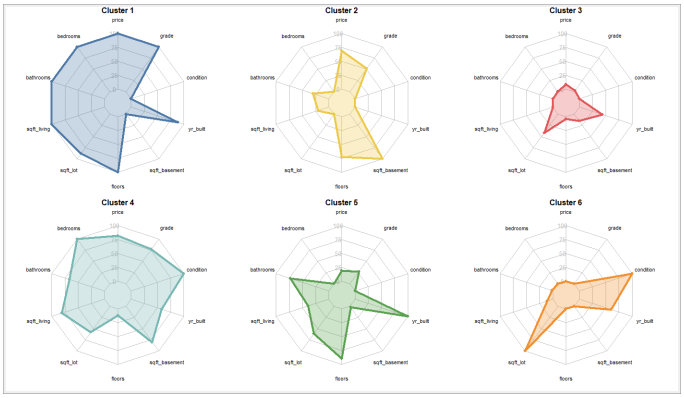

Figure 13. Radar chart cluster

del clustering ottenuto è data dalla rappresentazione delle osservazioni sul piano definito dalle prime due componenti principali mentre la terza componente principale è visualizzata tramite la dimensione dei punti (fig. 14).

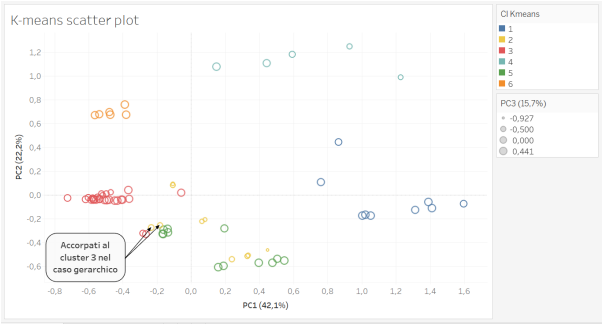

Figure 14. Scatter Plot

Il cluster numero 1 ad esempio, contiene le abitazioni con i prezzi più elevati, ciò probabilmente è dovuto anche all'elevata metratura e al numero di locali al loro interno. Possiamo notare anche che questo gruppo corrisponde ad un'area della contea (fig. 15) situata a nord e questo ci porta ad indurre che essa sia più ricca. Il cluster numero 3 al contrario, contiene le abitazioni con i prezzi più bassi e anche con il minor numero di locali al loro interno, esso corrisponde anche ad un quartiere definito di Seattle localizzato più nel centro sud. In generale dunque, un esito molto interessante della nostra ricerca, è dato dalla non dispersione a livello geografico dei vari zipcode appartenenti allo stesso gruppo come si può notare dalla mappa in fig. 15, segno che i nostri risultati rispecchiano l'esistenza di macro-aree con caratteristiche simili che compongono la città stessa. Non a caso in entrambi i metodi di clustering, ai vari raggruppamenti appartengono gli stessi zipcode ad eccezione di due. In figura è rappresentata il clustering ottenuto

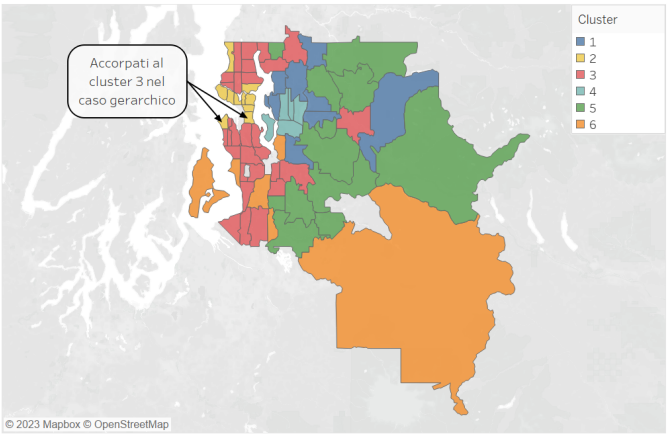

Figure 15. Visualizzazione geografica dei cluster

Per quanto concerne gli sviluppi futuri di questa ricerca, sarebbe interessante unire i due principali risultati ottenuti ovvero il modello predittivo e il cluster così da migliorarne le performance. Un altro obiettivo potrebbe essere quello di costruire un algoritmo più raffinato, partendo dalle tecniche di machine learning utilizzate in quest'analisi, che consente di aiutare l'acquirente a ridurre il suo campo di ricerca della casa o che in generale lo aiuti nella scelta in base ai suoi gusti e alle sue esigenze. Infine si potrebbe pensare di ottenere un dataset aggiornato con i prezzi degli immobili venduti nella King County Area nel 2022 per confrontare i risultati e analizzare se ci sono stati dei cambiamenti importanti negli ultimi anni.

References

[1] Office of financial Management. Median home price in washington. 2022.

[2] House sales in king county, usa. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction.

[3] Levi J. Wolf Sergio J. Rey, Dani Arribas-Bel. Clustering regionalization. 2020.

[4] Intelligenza Artificiale Italia. Analisi dei componenti principali (pca). 2020.