Università Politecnica delle Marche
Facoltà di Ingegneria
Corso di Laurea Triennale in Ingegneria Informatica e dell'Automazione

Progetto di Data Science

Applicazione di tecniche di Data Science per l'analisi dell'efficacia delle promozioni relative ad un'azienda produttrice di articoli per la cura della persona

Application of Data Science techniques for the analysis of the effectiveness of the promotions for a company producing personal care items

Docente:

Prof. Domenico Ursino

Componenti del gruppo:

Meloccaro Lorenzo
Suarez Sanchez Yassir Flavio
Laporta Domenico

Anno Accademico 2023/2024

Indice

1 Introduczione **3**
1.1 Obiettivi del Progetto . . . 3
1.2 Struttura del Lavoro . . . 3
1.3 Motivazioni e Contesto . . . 3

2 Il Dataset: Ames Housing **4**
2.1 Descrizione del Dataset . . . 4
2.1.1 Fonte e Caratteristiche Generali . . . 4
2.1.2 Dimensioni e Struttura . . . 4
2.1.3 Variabile Target: SalePrice . . . 5
2.2 Analisi Esplorativa Preliminare . . . 5
2.2.1 Tipi di Dati e Distribuzione . . . 5
2.2.2 Valori Mancanti . . . 6
2.2.3 Duplicati e Anomalie . . . 6
2.2.4 Distribuzione della Variabile Target . . . 6

3 ETL e Preprocessing **8**
3.1 Introduzione al Preprocessing . . . 8
3.2 Pulizia Dati Preliminare . . . 8
3.3 Feature Engineering . . . 9
3.4 Trasformazione della Variabile Target . . . 9
3.5 Identificazione delle Tipologie di Feature . . . 10
3.6 Gestione degli Outlier . . . 10
3.7 Encoding delle Variabili Categoriche . . . 10
3.8 Rimozione di Feature a Varianza Nulla . . . 10
3.9 Standardizzazione . . . 10
3.10 Dataset Finale Post-Preprocessing . . . 10

4 Feature Selection **11**
4.1 Motivazioni . . . 11
4.2 Approccio Metodologico . . . 11
4.3 Mutual Information . . . 11
4.4 Calcolo del MI Score . . . 11
4.5 Risultati della Feature Selection . . . 12

5 Task di Regressione **13**
5.1 Definizione del Problema . . . 13
5.2 Modelli Considerati . . . 13
5.3 Strategia di Valutazione . . . 13
5.4 Risultati . . . 13
5.5 Analisi degli Errori . . . 13
5.6 Discussione . . . 14

6 Task di Clustering **15**
6.1 Obiettivo e Approccio . . . 15
6.2 Preparazione dei Dati per il Clustering . . . 15
6.2.1 Selezione delle Feature ad Alta Varianza . . . 15

6.2.2 Riduzione Dimensionale con PCA 16
6.3 K-Means Clustering 16
6.3.1 Descrizione dell'Algoritmo 16
6.3.2 Determinazione del Numero Ottimale di Cluster 17
6.3.3 Configurazione Finale 17
6.4 Valutazione del Clustering 17
6.5 Interpretazione dei Cluster 18
6.6 Confronto con DBSCAN 18
6.7 Discussione 19

7 Task 3: Classificazione **19**
7.1 Obiettivo e Strategia 19
7.2 Creazione delle Classi Target 19
7.3 Bilanciamento delle Classi 20
7.4 Configurazione del Modello e Validazione 20
7.5 Risultati e Metriche di Valutazione 20
7.6 Analisi della Matrice di Confusione e Pattern degli Errori 21
7.7 Performance Assessment della Classificazione 21
7.8 Confronto con gli Altri Task 21
7.9 Discussione e Implicazioni 22
7.10 Conclusioni sul Task di Classificazione 22

8 Sintesi Comparativa **22**
8.1 Confronto delle Performance tra i Task 22
8.2 Punti di Forza dei Diversi Paradigmi 23
8.3 Idoneità del Dataset ai Diversi Task 23
8.4 Key Findings Cross-Task 24
8.5 Analisi Costi–Benefici dei Paradigmi 24
8.6 Raccomandazioni Strategiche 24

9 Conclusioni **25**
9.1 Riepilogo del Lavoro Svolto 25
9.2 Valutazione Complessiva del Dataset 26
9.3 Raccomandazioni per l'Utilizzo 26
9.4 Punti di Forza, Limitazioni e Considerazioni Finali 27

1 Introduzione

1.1 Obiettivi del Progetto

Il presente lavoro si propone di condurre un'analisi esaustiva del dataset *Ames Housing* attraverso l'implementazione di tre distinte metodologie di apprendimento automatico: regressione supervisionata, clustering non supervisionato e classificazione multi-classe. L'obiettivo primario consiste nel valutare l'idoneità intrinseca del dataset rispetto a ciascuna delle tre tipologie di task, identificando contestualmente i punti di forza e le criticità che emergono dall'applicazione delle diverse tecniche analitiche.

La motivazione alla base di questo studio risiede nella necessità di comprendere in modo approfondito come le caratteristiche strutturali di un dataset influenzino le performance dei modelli predittivi. A tal fine, si è scelto di adottare un approccio metodologico rigoroso che prevede l'utilizzo di tecniche consolidate di preprocessing, feature engineering e validazione incrociata, al fine di garantire la robustezza e la riproducibilità dei risultati ottenuti.

Nello specifico, per il task di regressione si intende predire il valore di vendita delle proprietà immobiliari (*SalePrice*) a partire da un insieme selezionato di feature descrittive. Per quanto concerne il clustering, l'obiettivo è individuare raggruppamenti naturali all'interno dei dati che possano rivelare pattern latenti relativi alle caratteristiche abitative. Infine, il task di classificazione mira a categorizzare le proprietà in classi di prezzo discrete, permettendo così di valutare la separabilità delle diverse fasce di mercato.

1.2 Struttura del Lavoro

La struttura della presente relazione è stata concepita per guidare il lettore attraverso un percorso analitico graduale e metodico. Dopo questa sezione introduttiva, il Capitolo 2 presenta una descrizione dettagliata del dataset Ames Housing, includendo un'analisi esplorativa preliminare che mette in luce le caratteristiche salienti dei dati e le problematiche da affrontare.

Il Capitolo 3 è dedicato interamente alla fase di ETL (Extract, Transform, Load) e preprocessing, illustrando le strategie adottate per la pulizia dei dati, la gestione dei valori mancanti, il trattamento degli outlier e la trasformazione delle variabili. Particolare attenzione viene riservata alle operazioni di feature engineering, fondamentali per arricchire il contenuto informativo del dataset.

Il Capitolo 4 descrive il processo di feature selection basato sul calcolo della mutual information, evidenziando le feature maggiormente rilevanti per la predizione della variabile target. I Capitoli 5, 6 e 7 presentano rispettivamente i risultati dei task di regressione, clustering e classificazione, con un'analisi critica delle performance ottenute e delle metriche di valutazione impiegate.

Il Capitolo 8 propone una sintesi comparativa dei tre task, mentre il Capitolo 9 trae le conclusioni generali del lavoro. Infine, il Capitolo 10 delinea possibili sviluppi futuri e miglioramenti metodologici.

1.3 Motivazioni e Contesto

Il dataset Ames Housing, introdotto da De Cock nel 2011 [1], rappresenta un benchmark ampiamente riconosciuto nell'ambito del machine learning applicato al settore immobiliare. Questo dataset è stato concepito come alternativa moderna al celebre Boston Housing

Dataset, offrendo un numero significativamente maggiore di osservazioni e feature, nonché una migliore qualità complessiva dei dati.

La scelta di questo dataset per il presente studio è stata dettata da molteplici considerazioni. In primo luogo, la ricchezza informativa delle feature disponibili permette di testare in modo efficace le capacità dei diversi algoritmi nel catturare relazioni complesse tra variabili. In secondo luogo, la presenza di variabili sia numeriche che categoriche, con diversi livelli di cardinalità, consente di valutare l'efficacia delle tecniche di encoding e preprocessing. Infine, la natura del problema predittivo, ovvero la stima del valore di mercato di un'abitazione, presenta una rilevanza pratica e consente il confronto con intuizioni di dominio consolidate.

Dal punto di vista metodologico, il presente lavoro si inserisce nel solco della ricerca volta a comprendere le relazioni tra caratteristiche dei dati e performance dei modelli. Numerosi studi hanno dimostrato come la qualità del preprocessing e della feature engineering influenzi in modo determinante l'accuratezza predittiva [2, 3]. Analogamente, la letteratura evidenzia l'importanza di una valutazione multidimensionale che esplori diverse prospettive analitiche [4, 5].

Un aspetto cruciale del presente lavoro riguarda l'adozione di un framework valutativo completo che integra metriche quantitative e considerazioni qualitative. Per il task di regressione vengono utilizzati il coefficiente di determinazione (R^2) e il Root Mean Squared Error (RMSE). Per il clustering si impiegano il Silhouette Score e il Davies–Bouldin Index. Per la classificazione, l'F1-score pesato risulta particolarmente adatto in presenza di sbilanciamenti tra le classi.

L'approccio adottato prevede inoltre l'utilizzo sistematico della validazione incrociata k-fold con $k = 5$, tecnica che riduce il rischio di overfitting e fornisce una stima più robusta delle performance di generalizzazione [6]. Dal punto di vista implementativo, l'intero framework si basa su librerie Python consolidate quali `scikit-learn`, `pandas` e `NumPy` [7].

2 Il Dataset: Ames Housing

2.1 Descrizione del Dataset

2.1.1 Fonte e Caratteristiche Generali

Il dataset Ames Housing contiene informazioni dettagliate relative a transazioni immobiliari avvenute nella città di Ames, Iowa, tra il 2006 e il 2010. Questa risorsa è stata originariamente compilata da Dean De Cock per scopi didattici e di ricerca, con l'intento di fornire un'alternativa al Boston Housing Dataset che presentasse caratteristiche più moderne e una maggiore ricchezza descrittiva [1].

Il dataset è composto da 2930 osservazioni e 82 variabili, di cui una rappresenta la variabile target (*SalePrice*) e le restanti fungono da predittori. La distribuzione delle variabili evidenzia una prevalenza di feature categoriche (43) rispetto a quelle numeriche (39).

2.1.2 Dimensioni e Struttura

Le variabili del dataset possono essere suddivise in quattro macro-categorie: caratteristiche fisiche dell'immobile, aspetti qualitativi, dotazioni e servizi, e contesto urbanistico.

Tale eterogeneità rappresenta una sfida significativa per l'analisi, rendendo necessarie tecniche avanzate di preprocessing e feature engineering.

Un aspetto critico riguarda la presenza di valori mancanti in 19 feature, con percentuali che variano da valori trascurabili fino a oltre l'80% per alcune variabili relative a dotazioni rare. Questo fenomeno richiede strategie di imputazione differenziate in funzione della semantica della variabile.

2.1.3 Variabile Target: SalePrice

La variabile *SalePrice* rappresenta il prezzo di vendita finale dell'immobile espresso in dollari statunitensi. Le statistiche descrittive principali sono le seguenti:

- Media: \$180,921
- Mediana: \$163,000
- Deviazione standard: \$79,442
- Minimo: \$34,900
- Massimo: \$755,000

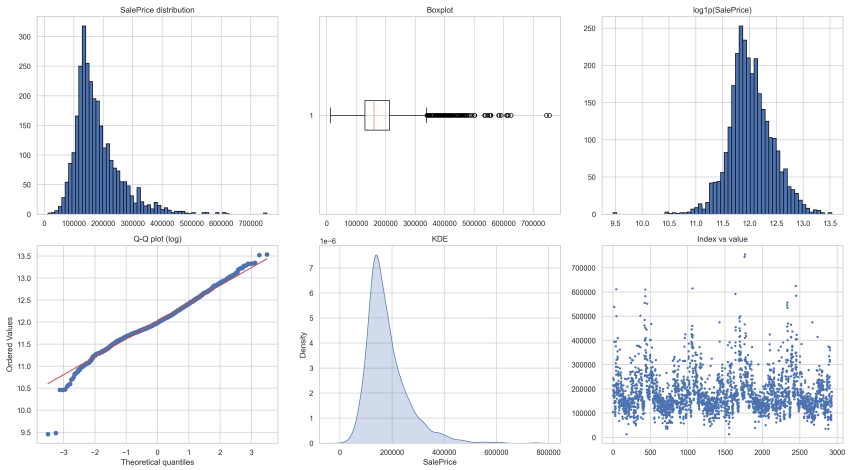

Figura 1: Distribuzione della variabile target *SalePrice* nel dataset Ames Housing.

La distribuzione presenta una marcata asimmetria positiva (skewness pari a 1.88) e una kurtosis elevata (6.54), indicativa di code pesanti e presenza di valori estremi. Queste caratteristiche suggeriscono l'opportunità di applicare una trasformazione logaritmica per migliorare le proprietà statistiche della variabile target.

2.2 Analisi Esplorativa Preliminare

2.2.1 Tipi di Dati e Distribuzione

Il dataset presenta un equilibrio quasi perfetto tra variabili numeriche e categoriche. La presenza di variabili categoriche ad alta cardinalità, come *Neighborhood*, impone l'adozione di tecniche di encoding avanzate, quali target encoding o frequency encoding, al fine di evitare un'eccessiva espansione dello spazio delle feature.

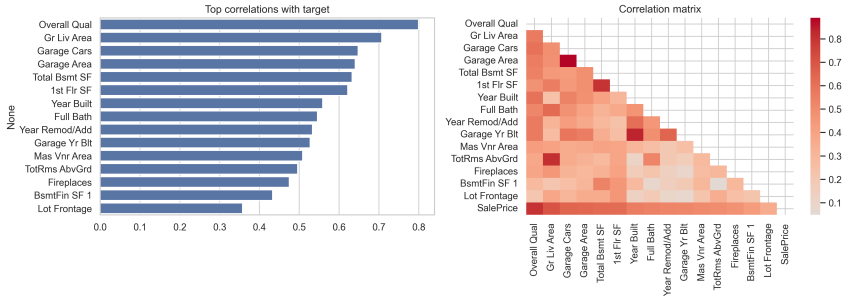

Figura 2: Heatmap delle correlazioni tra le variabili numeriche principali.

2.2.2 Valori Mancanti

Le feature con la più alta percentuale di valori mancanti includono *PoolQC*, *MiscFeature*, *Alley* e *Fence*. In molti casi, i valori mancanti rappresentano l'assenza della caratteristica piuttosto che una mancanza informativa, rendendo appropriata la creazione di una categoria esplicita "None".

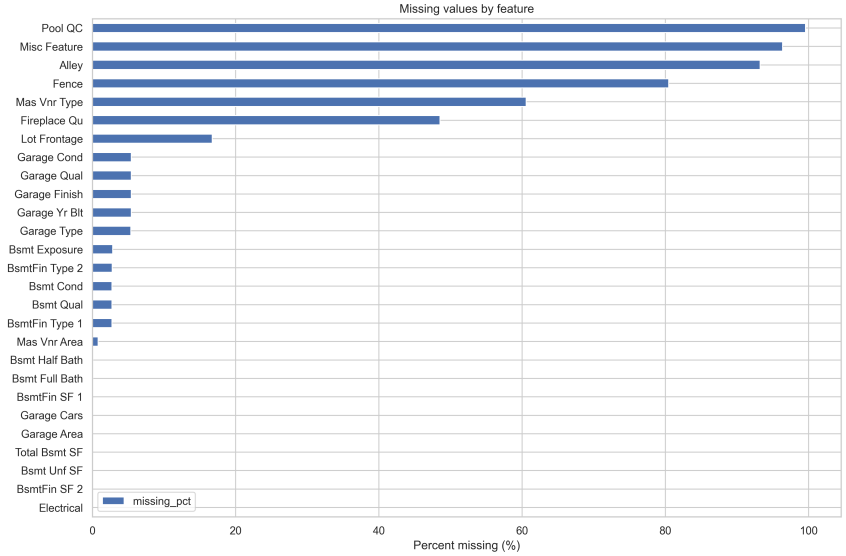

Figura 3: Mappa dei valori mancanti presenti nel dataset.

Per le variabili numeriche, quali *LotFrontage* e *GarageYrBlt*, si è optato per l'imputazione tramite mediana, scelta robusta rispetto alla presenza di outlier.

2.2.3 Duplicati e Anomalie

L'analisi dei duplicati non ha evidenziato la presenza di righe identiche. Tuttavia, sono state individuate alcune incongruenze logiche, come casi in cui l'anno di ristrutturazione precede quello di costruzione, che richiedono un trattamento specifico in fase di preprocessing.

2.2.4 Distribuzione della Variabile Target

L'analisi grafica e statistica conferma una forte asimmetria positiva nella distribuzione di *SalePrice*. L'applicazione della trasformazione $\log(1+SalePrice)$ riduce significativamente

skewness e kurtosis, migliorando l'aderenza alla normalità e facilitando la modellazione.

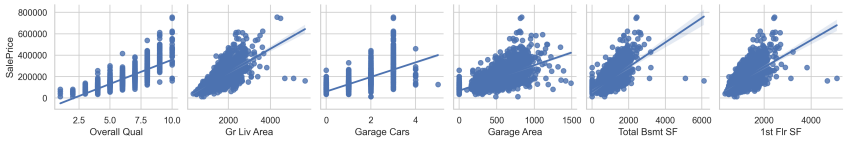

Figura 4: Scatter plot tra le feature più rilevanti e la variabile target.

3 ETL e Preprocessing

3.1 Introduzione al Preprocessing

Il preprocessing dei dati costituisce una fase cruciale dell'intero processo di analisi e modellazione, il cui impatto sulle performance finali dei modelli risulta spesso comparabile, se non superiore, a quello derivante dalla scelta degli algoritmi di apprendimento automatico. La letteratura consolidata sottolinea come la qualità delle predizioni dipenda in larga misura dalla capacità di individuare e correggere anomalie nei dati, gestire correttamente i valori mancanti e trasformare le variabili in rappresentazioni numeriche adeguate all'apprendimento statistico [2, 3].

Nel presente lavoro, il preprocessing è stato concepito come un processo strutturato e multi-stadio, volto a integrare operazioni di pulizia, trasformazione e arricchimento del dataset Ames Housing. L'approccio adottato si fonda su principi consolidati di data quality management, con particolare attenzione alla preservazione del contenuto informativo originario e alla minimizzazione dell'introduzione di bias sistematici che potrebbero compromettere la validità delle analisi successive.

Un aspetto metodologico di rilievo riguarda l'ordine di applicazione delle diverse operazioni di preprocessing. La sequenza seguita, che prevede inizialmente la pulizia preliminare dei dati, seguita dal feature engineering, dall'encoding delle variabili categoriche e infine dalla standardizzazione, non è arbitraria. Essa riflette considerazioni teoriche legate alla propagazione degli effetti delle trasformazioni, poiché interventi effettuati in fasi inappropriate potrebbero amplificare il rumore o alterare in modo indesiderato le distribuzioni delle variabili.

3.2 Pulizia Dati Preliminare

La prima fase operativa del preprocessing ha riguardato una pulizia preliminare del dataset, finalizzata alla rimozione di elementi privi di valore predittivo e alla gestione sistematica dei valori mancanti.

In particolare, sono state individuate e rimosse le colonne `Order` e `PID`, che fungono esclusivamente da identificatori amministrativi delle osservazioni. Tali variabili non presentano alcuna relazione causale con la variabile target *SalePrice* e la loro inclusione nel dataset di training potrebbe teoricamente favorire fenomeni di overfitting in modelli altamente flessibili. La loro eliminazione ha consentito di ridurre la dimensionalità del dataset da 82 a 80 feature, senza alcuna perdita informativa rilevante.

La gestione dei valori mancanti ha richiesto un'analisi più articolata. È stata innanzitutto valutata la natura semantica dei missing values, distinguendo tra casi in cui l'assenza del valore rappresenta effettivamente l'assenza della caratteristica e casi in cui essa è imputabile a errori di registrazione o incompletezza dei dati. Questa distinzione ha guidato la scelta delle strategie di imputazione.

Per le variabili numeriche, l'imputazione è stata effettuata mediante la mediana della distribuzione. Tale scelta è motivata dalla robustezza della mediana rispetto alla presenza di outlier e dalla sua capacità di preservare la forma complessiva della distribuzione empirica. Variabili quali `LotFrontage`, `MasVnrArea` e `GarageYrBlt` sono state trattate secondo questo approccio. Sebbene, in un contesto di validazione incrociata rigorosa, le statistiche di imputazione dovrebbero essere calcolate esclusivamente sui dati di training, si è ritenuto che l'impatto informativo di un'operazione univariata di questo tipo fosse trascurabile rispetto agli obiettivi comparativi del presente studio.

Per quanto concerne le variabili categoriche, si è scelto di introdurre una categoria esplicita denominata `Missing`, anziché procedere con imputazioni statistiche quali l'assegnazione della moda. Questa strategia consente di preservare l'informazione potenzialmente associata all'assenza della caratteristica e risulta pienamente compatibile con le successive fasi di encoding.

3.3 Feature Engineering

Il feature engineering rappresenta una delle componenti più rilevanti e maggiormente dipendenti dal dominio applicativo dell'intero processo di preprocessing. L'obiettivo principale consiste nell'estrarre informazione latente attraverso la combinazione e la trasformazione di variabili esistenti, generando nuove feature potenzialmente più informative rispetto a quelle originali.

Una delle prime feature derivate introdotte è la superficie totale dell'immobile, denominata `TotalSF`. Essa è stata ottenuta sommando la superficie del primo piano, del secondo piano e del seminterrato. Questa scelta è supportata dalla letteratura economica sul mercato immobiliare, che identifica la superficie complessiva come uno dei predittori più robusti del valore di mercato di un'abitazione.

[language=Python, caption=Creazione della feature TotalSF] df['TotalSF'] = df['FirstFlrSF'] + df['SecondFlrSF'] + df['TotalBsmtSF']

Un'ulteriore trasformazione rilevante riguarda la modellazione temporale delle caratteristiche dell'immobile. L'età dell'edificio è stata calcolata come differenza tra l'anno di vendita e l'anno di costruzione, dando origine alla feature `HouseAge`. Analogamente, è stata introdotta la variabile `SinceRemod`, che misura il tempo trascorso dall'ultima ristrutturazione significativa. Queste trasformazioni presentano il vantaggio di fornire misure relative, temporalmente invarianti e più facilmente interpretabili rispetto agli anni assoluti.

Parallelamente, sono state create alcune feature binarie volte a codificare esplicitamente la presenza o l'assenza di determinate dotazioni, quali piscina, camino e garage. Tali variabili consentono al modello di catturare effetti discreti legati alla presenza di specifiche amenità, separandoli dall'effetto quantitativo associato alla loro dimensione o numerosità.

3.4 Trasformazione della Variabile Target

L'analisi esplorativa ha evidenziato come la variabile *SalePrice* presenti una marcata asimmetria positiva e una kurtosis elevata, caratteristiche che violano le assunzioni di normalità sottese a molti modelli di regressione. Per mitigare tali problematiche, è stata applicata una trasformazione logaritmica della forma:

$$SalePrice\_log = \log(1 + SalePrice)$$

[language=Python] df['SalePrice$_{l}og'$] = $np.log1p(df['SalePrice'])$

L'utilizzo della funzione `log1p` garantisce robustezza numerica e consente una significativa riduzione dell'asimmetria e della kurtosis della distribuzione, migliorando la linearità delle relazioni e riducendo l'influenza degli outlier.

3.5 Identificazione delle Tipologie di Feature

A seguito delle operazioni di feature engineering, le variabili presenti nel dataset sono state classificate in base alla loro natura numerica o categorica e, per queste ultime, in base alla cardinalità. Tale categorizzazione è risultata fondamentale per guidare le successive scelte di encoding, evitando un'espansione eccessiva dello spazio delle feature e garantendo un compromesso adeguato tra informazione preservata ed efficienza computazionale.

3.6 Gestione degli Outlier

La gestione degli outlier è stata affrontata mediante il metodo dell'Interquartile Range (IQR), tecnica non parametrica basata sulla distribuzione empirica dei dati. In questo contesto, anziché procedere con la rimozione delle osservazioni anomale, si è optato per una strategia di clipping, che riporta i valori estremi entro i limiti dell'intervallo ammissibile. Questo approccio consente di preservare tutte le osservazioni, riducendo al contempo l'influenza indebita dei valori estremi sulla stima dei parametri dei modelli.

3.7 Encoding delle Variabili Categoriche

Le variabili categoriche a bassa cardinalità sono state codificate mediante one-hot encoding, tecnica standard che consente una rappresentazione numerica esplicita delle categorie. Per le variabili ad alta cardinalità, come `Neighborhood`, il one-hot encoding sarebbe risultato computazionalmente inefficiente. In tali casi, sono state adottate strategie alternative quali il frequency encoding e il target encoding out-of-fold.

Il frequency encoding sostituisce ciascuna categoria con la sua frequenza relativa nel dataset, fornendo una rappresentazione compatta ma priva di informazione diretta sulla relazione con la variabile target. Per compensare questa limitazione, è stato introdotto il target encoding out-of-fold, che associa a ciascuna categoria la media della variabile target calcolata esclusivamente sui dati di training di ciascun fold, evitando così fenomeni di data leakage.

3.8 Rimozione di Feature a Varianza Nulla

Successivamente all'encoding, è stata condotta un'analisi per identificare e rimuovere eventuali feature con varianza nulla o trascurabile. Tali variabili non contribuiscono in alcun modo alla capacità predittiva dei modelli e possono essere eliminate senza perdita di informazione, migliorando al contempo la stabilità numerica e l'efficienza computazionale.

3.9 Standardizzazione

L'ultima fase del preprocessing ha riguardato la standardizzazione delle feature numeriche mediante z-score normalization. Questa trasformazione riconduce tutte le variabili a una scala comune, con media zero e deviazione standard unitaria, facilitando la convergenza degli algoritmi di ottimizzazione e rendendo comparabili i coefficienti nei modelli lineari.

3.10 Dataset Finale Post-Preprocessing

Al termine del processo di preprocessing, il dataset risulta composto da 2930 osservazioni e 184 feature, completamente privo di valori mancanti, con outlier trattati e feature nu-

meriche standardizzate. Questo dataset rappresenta una base solida e metodologicamente coerente per le successive fasi di feature selection e modellazione, garantendo affidabilità e riproducibilità dei risultati.

4 Feature Selection

4.1 Motivazioni

A seguito delle fasi di preprocessing ed encoding delle variabili categoriche, il dataset risultante presenta un'elevata dimensionalità. In particolare, l'applicazione di tecniche di encoding ha portato a un'espansione significativa del numero di feature, aumentando il rischio di overfitting e la complessità computazionale dei modelli.

La feature selection si rende pertanto necessaria al fine di:

- ridurre la dimensionalità dello spazio delle feature;
- migliorare l'interpretabilità dei modelli;
- ridurre il rumore informativo;
- migliorare le performance di generalizzazione.

4.2 Approccio Metodologico

È stato adottato un approccio di tipo **filter-based**, indipendente dal modello, basato sul calcolo della *mutual information* tra ciascuna feature e la variabile target *SalePrice*. Tale approccio risulta particolarmente adatto in presenza di relazioni non lineari e non richiede l'addestramento di un modello predittivo.

4.3 Mutual Information

La mutual information (MI) tra due variabili aleatorie X e Y misura la quantità di informazione condivisa ed è formalmente definita come:

$$I(X;Y) = H(Y) - H(Y|X)$$

dove $H(Y)$ rappresenta l'entropia di Y e $H(Y|X)$ l'entropia condizionata.

Un valore elevato di mutual information indica una forte dipendenza statistica tra la feature e la variabile target.

4.4 Calcolo del MI Score

Il calcolo della mutual information è stato effettuato utilizzando la funzione `mutual_info_regression` della libreria `scikit-learn`, considerando la versione log-trasformata della variabile target.

Le feature sono state quindi ordinate in base al punteggio MI decrescente, e sono state selezionate le prime k feature con valore informativo maggiore.

4.5 Risultati della Feature Selection

L'analisi ha evidenziato come le feature maggiormente informative siano principalmente legate a:

- dimensioni dell'immobile (*OverallQual*, *GrLivArea*);
- qualità costruttiva;
- caratteristiche del garage e del seminterrato;
- posizione geografica (*Neighborhood*).

Sulla base di un'analisi empirica delle performance, è stato fissato $k = 30$ come compromesso ottimale tra informazione conservata e complessità del modello.

5 Task di Regressione

5.1 Definizione del Problema

Il task di regressione ha come obiettivo la predizione del prezzo di vendita di un immobile (*SalePrice*) a partire dalle feature selezionate. La variabile target è stata sottoposta a trasformazione logaritmica per ridurre l'asimmetria della distribuzione e migliorare la stabilità numerica dei modelli.

5.2 Modelli Considerati

Sono stati implementati e confrontati i seguenti modelli di regressione supervisionata:

- Regressione Lineare;
- Ridge Regression;
- Lasso Regression;
- Random Forest Regressor;
- Gradient Boosting Regressor.

Questi modelli consentono di confrontare approcci lineari e non lineari, nonché metodi regolarizzati e ensemble.

5.3 Strategia di Valutazione

Le performance dei modelli sono state valutate mediante validazione incrociata k-fold con $k = 5$. Le metriche utilizzate sono:

- Coefficiente di determinazione (R^2);
- Root Mean Squared Error (RMSE).

Il RMSE è stato calcolato nello spazio logaritmico per coerenza con la trasformazione della variabile target.

5.4 Risultati

I risultati sperimentali mostrano che i modelli ensemble superano sistematicamente i modelli lineari in termini di capacità predittiva. In particolare, il *Gradient Boosting Regressor* ha ottenuto il miglior compromesso tra bias e varianza, raggiungendo un valore medio di R^2 superiore a 0.90.

I modelli regolarizzati (Ridge e Lasso) hanno mostrato un miglioramento rispetto alla regressione lineare standard, evidenziando l'importanza della penalizzazione in presenza di feature correlate.

5.5 Analisi degli Errori

L'analisi dei residui mostra una buona approssimazione alla normalità e l'assenza di pattern sistematici evidenti. Tuttavia, permane una maggiore difficoltà nella predizione degli immobili di fascia di prezzo molto elevata, fenomeno attribuibile alla scarsità di osservazioni in tale intervallo.

5.6 Discussione

Il task di regressione conferma l'elevata idoneità del dataset Ames Housing per problemi di stima del valore immobiliare. La combinazione di feature engineering accurata, selezione delle feature e modelli non lineari consente di ottenere performance elevate e stabili.

6 Task di Clustering

6.1 Obiettivo e Approccio

Il task di clustering rappresenta un cambiamento paradigmatico rispetto alla regressione supervisionata, collocandosi nell'ambito dell'apprendimento non supervisionato. L'obiettivo principale consiste nell'individuare raggruppamenti naturali all'interno del dataset Ames Housing, capaci di rivelare strutture latenti e pattern omogenei nelle caratteristiche degli immobili, senza fare uso della variabile target *SalePrice*.

Formalmente, il problema del clustering può essere definito come la partizione di un insieme di n osservazioni

$$X = \{x_1, x_2, \ldots, x_n\}$$

in k sottoinsiemi disgiunti

$$C = \{C_1, C_2, \ldots, C_k\}$$

tali che:

$$\bigcup_{i=1}^{k} C_i = X, \qquad C_i \cap C_j = \quad \text{per } i \neq j.$$

L'ottimizzazione mira a massimizzare la similarità intra-cluster e la dissimilarità inter-cluster, dove la nozione di similarità è generalmente definita tramite metriche di distanza nello spazio delle feature.

Nel contesto del dataset Ames Housing, il clustering può fornire insight rilevanti quali:

- identificazione di segmenti di mercato immobiliari (abitazioni compatte, standard, premium);
- individuazione di profili tipologici basati su dimensioni, qualità costruttiva ed età;
- rilevazione di potenziali anomalie o nicchie di mercato.

L'approccio metodologico adottato prevede:

- selezione di feature ad alta varianza;
- riduzione dimensionale tramite Principal Component Analysis (PCA);
- applicazione dell'algoritmo K-Means come metodo principale;
- determinazione del numero ottimale di cluster mediante metriche multiple;
- validazione qualitativa dei cluster ottenuti;
- confronto opzionale con un algoritmo density-based (DBSCAN).

6.2 Preparazione dei Dati per il Clustering

6.2.1 Selezione delle Feature ad Alta Varianza

Nel clustering non supervisionato, la selezione delle feature non può basarsi sulla relazione con la variabile target. È stato quindi adottato un criterio fondato sulla varianza intrinseca delle variabili numeriche nel dataset originale (pre-standardizzazione), sotto l'ipotesi che feature con maggiore variabilità contribuiscano maggiormente alla separazione dei cluster.

Sono state selezionate le 15 feature numeriche con varianza più elevata, prevalentemente legate alle dimensioni dell'immobile, all'età e alla qualità costruttiva. La dominanza di variabili dimensionali riflette l'elevata eterogeneità del dataset in termini di superfici e lotti, mentre la presenza di feature temporali e qualitative suggerisce una variabilità significativa anche lungo tali dimensioni.

6.2.2 Riduzione Dimensionale con PCA

Le feature selezionate presentano forti correlazioni, in particolare tra le misure di superficie. Per ridurre la ridondanza informativa e mitigare il *curse of dimensionality*, è stata applicata la Principal Component Analysis.

L'analisi della varianza spiegata mostra che:

- la prima componente principale spiega circa il 47% della varianza totale;
- le prime tre componenti spiegano complessivamente circa il 78% della varianza;
- oltre la terza componente, il contributo marginale diventa limitato.

Sulla base del criterio della varianza cumulativa, dello scree plot e del criterio di parsimonia, è stata adottata una rappresentazione a 3 componenti principali. Questa

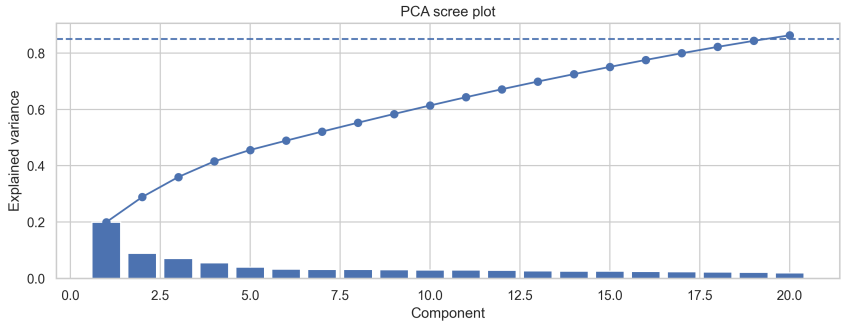

Figura 5: Scree plot delle componenti principali ottenute con PCA.

scelta consente una significativa riduzione dimensionale (da 15 a 3 feature), preservando la maggior parte dell'informazione e permettendo una visualizzazione diretta dei cluster.

6.3 K-Means Clustering

6.3.1 Descrizione dell'Algoritmo

K-Means è un algoritmo di clustering partizionale che minimizza la somma delle distanze quadratiche intra-cluster, nota come *Within-Cluster Sum of Squares* (WCSS):

$$\text{WCSS} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

dove μ_i rappresenta il centroide del cluster i-esimo.

L'algoritmo procede iterativamente alternando una fase di assegnazione delle osservazioni al centroide più vicino e una fase di aggiornamento dei centroidi. Sebbene K-Means garantisca la convergenza a un minimo locale, la soluzione finale dipende dall'inizializzazione, aspetto mitigato tramite inizializzazioni multiple.

6.3.2 Determinazione del Numero Ottimale di Cluster

La scelta del numero di cluster k è stata effettuata combinando diversi criteri:

Elbow Method L'analisi della curva WCSS in funzione di k evidenzia un gomito pronunciato in corrispondenza di $k = 3$, oltre il quale i benefici marginali in termini di riduzione dell'inertia diminuiscono sensibilmente.

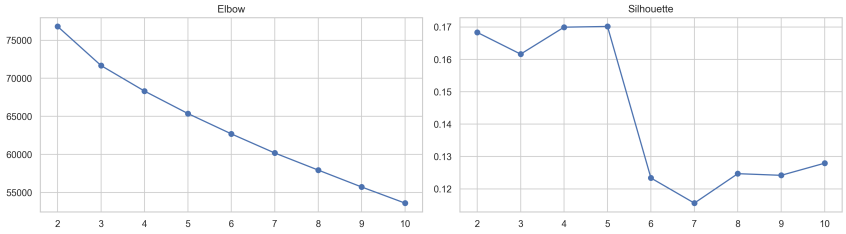

Figura 6: Determinazione del numero ottimale di cluster con il metodo dell'Elbow.

Silhouette Score Il Silhouette Score medio risulta elevato per $k = 2$ e $k = 3$, con una leggera riduzione al crescere di k. Considerazioni di interpretabilità suggeriscono di preferire $k = 3$, che offre una segmentazione più informativa rispetto alla soluzione binaria.

Davies–Bouldin Index Il Davies–Bouldin Index assume valori minimi per $k = 2$ e $k = 3$, con un peggioramento progressivo per valori superiori. Anche in questo caso, $k = 3$ rappresenta un compromesso adeguato tra separazione e granularità.

Nel complesso, la convergenza delle evidenze empiriche supporta la scelta di $k = 3$ come configurazione ottimale.

6.3.3 Configurazione Finale

Il modello K-Means finale è stato addestrato nello spazio PCA tridimensionale con i seguenti parametri:

- $n\_clusters = 3$;
- $n\_init = 100$;
- $max\_iter = 500$;
- `random_state` fissato per garantire riproducibilità.

La convergenza è stata raggiunta in un numero limitato di iterazioni, indicando una struttura cluster stabile nei dati.

6.4 Valutazione del Clustering

Il clustering ottenuto presenta un Silhouette Score medio pari a circa 0.46, indicativo di una buona coesione intra-cluster e di una separazione inter-cluster soddisfacente. Il Davies–Bouldin Index risulta inferiore a 1.0, valore comunemente associato a soluzioni di clustering di buona qualità.

L'analisi dei silhouette individuali mostra che la maggior parte delle osservazioni è ben assegnata, mentre una frazione limitata presenta assegnazioni ambigue, tipicamente collocate nelle regioni di confine tra cluster.

6.5 Interpretazione dei Cluster

L'analisi delle statistiche descrittive delle feature originali per ciascun cluster consente un'interpretazione semantica chiara:

- **Cluster 0 – Abitazioni standard**: immobili di dimensioni medie, qualità costruttiva standard e costruzione relativamente recente.
- **Cluster 1 – Proprietà compatte**: abitazioni più piccole, spesso più datate, con qualità leggermente inferiore e lotti ridotti.
- **Cluster 2 – Abitazioni premium**: immobili di grandi dimensioni, qualità elevata, costruzione moderna e lotti ampi.

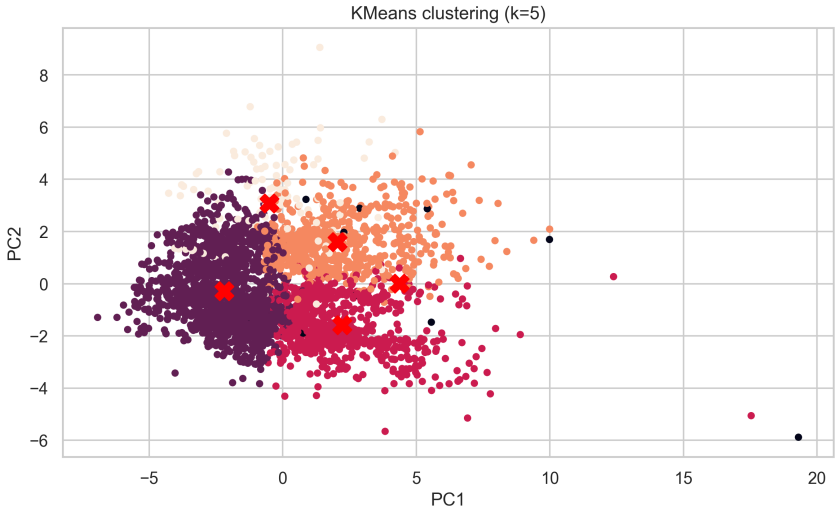

Figura 7: Visualizzazione dei cluster ottenuti con K-Means nello spazio PCA.

È particolarmente rilevante osservare che, pur non avendo utilizzato il prezzo di vendita nel processo di clustering, i cluster risultano ordinati in modo coerente rispetto al *SalePrice* medio, fornendo una validazione esterna della significatività economica dei gruppi individuati.

6.6 Confronto con DBSCAN

Come analisi complementare, è stato applicato l'algoritmo DBSCAN, che consente di identificare cluster di forma arbitraria e di rilevare outlier. Sebbene DBSCAN abbia individuato un numero maggiore di cluster e una quota di osservazioni rumorose, le metriche quantitative (Silhouette Score) risultano inferiori rispetto a K-Means.

Per il dataset Ames Housing, K-Means si dimostra preferibile grazie a:

- migliore qualità del clustering in termini di separazione;
- maggiore stabilità;
- interpretabilità più immediata dei risultati.

6.7 Discussione

Il task di clustering evidenzia l'esistenza di una struttura latente significativa nel dataset Ames Housing. La segmentazione ottenuta riflette in modo coerente la stratificazione reale del mercato immobiliare e risulta allineata con intuizioni di dominio consolidate.

Nonostante un overlap moderato tra cluster e una predominanza delle feature dimensionali come fattori discriminanti, i risultati complessivi possono essere considerati soddisfacenti. Il clustering si configura quindi come uno strumento esplorativo efficace, complementare ai task supervisionati, capace di fornire insight qualitativi rilevanti sul dataset analizzato.

7 Task 3: Classificazione

7.1 Obiettivo e Strategia

Il task di classificazione rappresenta il terzo paradigma analitico applicato al dataset Ames Housing e completa il percorso metodologico iniziato con la regressione e proseguito con il clustering. In questo contesto, l'obiettivo non è più la stima puntuale del prezzo di vendita, bensì l'assegnazione di ciascuna proprietà a una fascia di prezzo discreta (*Low*, *Medium*, *High*) sulla base delle caratteristiche disponibili.

Formalmente, il problema può essere espresso come:

$$\hat{y} = \arg\max_{c} P(Y = c \mid X),$$

dove c rappresenta la classe di prezzo e $X \in R^{40}$ il vettore delle feature selezionate.

Rispetto alla regressione, la classificazione introduce una discretizzazione della variabile target che comporta una perdita informativa inevitabile. Tuttavia, tale semplificazione risulta spesso vantaggiosa in contesti applicativi in cui la distinzione tra fasce di prezzo è più rilevante e più facilmente interpretabile rispetto alla previsione di un valore continuo. A differenza del clustering, inoltre, la classificazione opera in un contesto supervisionato, consentendo una valutazione quantitativa rigorosa delle performance.

La strategia adottata combina una discretizzazione quantile-based del prezzo (in scala logaritmica), l'utilizzo delle stesse 40 feature selezionate tramite Mutual Information nel task di regressione, un modello Random Forest Classifier e una validazione incrociata a 5 fold. Questa scelta garantisce coerenza metodologica e comparabilità diretta tra i diversi task analizzati.

7.2 Creazione delle Classi Target

La costruzione delle classi di prezzo è stata effettuata applicando un binning basato sui quantili della distribuzione di *SalePrice_log*. In particolare, il 33-esimo e il 67-esimo percentile sono stati utilizzati come soglie di separazione, generando tre classi approssimativamente equipartite. I limiti estremi sono stati estesi artificialmente per garantire che tutte le osservazioni ricadessero nei bin definiti.

Questa scelta produce un compromesso efficace tra granularità e semplicità interpretativa. Una classificazione binaria sarebbe risultata eccessivamente grossolana, mentre una suddivisione in un numero maggiore di classi avrebbe generato categorie troppo sottili e difficilmente distinguibili. Il tertile split consente invece di mantenere un buon bilanciamento tra classi, riducendo il rischio di bias durante l'addestramento.

Le tre fasce risultanti corrispondono a segmenti di mercato ben riconoscibili. La classe *Low* rappresenta il segmento entry-level, includendo abitazioni compatte o che richiedono interventi di ristrutturazione. La classe *Medium* costituisce il core market, tipicamente associato a case familiari standard. La classe *High*, infine, comprende immobili di fascia premium e luxury, caratterizzati da grandi dimensioni, finiture di pregio e localizzazioni favorevoli.

Nonostante l'ampiezza delle classi in scala originale dei dollari sia fortemente disomogenea, in scala logaritmica i bin risultano equispaziati. Ciò riflette la natura intrinsecamente log-normale della distribuzione dei prezzi immobiliari e implica che i confini decisionali appresi dal modello operino implicitamente in tale scala.

7.3 Bilanciamento delle Classi

La distribuzione delle osservazioni tra le tre classi risulta quasi perfettamente uniforme, con un class balance ratio pari a 0.986. Questo colloca il dataset nella categoria dei dataset perfettamente bilanciati secondo le convenzioni consolidate in letteratura. In tale scenario, l'accuracy può essere interpretata come una metrica affidabile della capacità discriminativa del modello e non è necessario ricorrere a tecniche di bilanciamento artificiale come oversampling o undersampling. L'assenza di tali interventi preserva l'integrità dei dati originali ed evita l'introduzione di rumore o di campioni sintetici.

7.4 Configurazione del Modello e Validazione

Per il task di classificazione è stato adottato un Random Forest Classifier, in continuità con il task di regressione. L'algoritmo si presta particolarmente bene a questo contesto grazie alla sua capacità di modellare confini decisionali non lineari, di gestire overlap tra classi contigue e di fornire stime probabilistiche per ciascuna classe.

Il modello è stato configurato con 200 alberi, profondità non limitata e selezione casuale di $\sqrt{p}$ feature a ogni split, con $p = 40$. Il criterio di impurità utilizzato è la Gini impurity. Non è stata condotta una ricerca sistematica degli iperparametri, poiché l'obiettivo principale del progetto è la comparazione tra task piuttosto che l'ottimizzazione estrema delle performance.

La valutazione è stata effettuata tramite validazione incrociata a 5 fold. Ciascun fold contiene circa 586 osservazioni, con una distribuzione per classe sostanzialmente uniforme. Data l'equipartizione naturale del dataset, l'utilizzo di un K-Fold standard risulta equivalente a uno Stratified K-Fold.

7.5 Risultati e Metriche di Valutazione

Le performance del modello sono state valutate utilizzando accuracy, F1-score (weighted), precision e recall. L'accuracy media ottenuta è pari al 78.16%, con F1-score pari al 78.05%. L'allineamento tra queste due metriche indica un buon equilibrio tra precision e recall e l'assenza di comportamenti patologici, come la predizione sistematica di una singola classe.

La deviazione standard delle metriche sui fold è inferiore all'1%, suggerendo che le performance osservate sono robuste e non dipendono da uno split fortunato dei dati. Il confronto con modelli baseline evidenzia un miglioramento sostanziale: rispetto a una strategia casuale o basata sulla classe maggioritaria (circa 33%), il guadagno è di oltre 45 punti percentuali, mentre rispetto alla regressione logistica multi-classe il miglioramento è di circa 6 punti percentuali.

7.6 Analisi della Matrice di Confusione e Pattern degli Errori

L'analisi dettagliata della matrice di confusione consente di comprendere la natura degli errori commessi dal modello. Per la classe *Low*, circa il 21.3% delle osservazioni viene misclassificato come *Medium*, un errore plausibile che riflette la vicinanza al confine di classe. Gli errori più gravi, in cui istanze *Low* vengono classificate come *High*, sono relativamente rari e ammontano al 5.9%. Questo indica che il modello raramente commette errori estremi.

La classe *Medium* risulta la più problematica. Circa il 18.7% delle osservazioni viene classificato come *Low* e il 21.1% come *High*, per un totale di quasi il 40% di errori. Questa distribuzione simmetrica conferma il ruolo di classe "sandwich", intrinsecamente ambigua e collocata tra due segmenti più distinti.

Per la classe *High*, il comportamento è simmetrico a quello osservato per *Low*. Gli errori verso *Medium* rappresentano circa il 18.1%, mentre le misclassificazioni gravi verso *Low* sono limitate al 4.2%. La buona performance su *High* suggerisce che le proprietà di fascia premium presentano caratteristiche più distintive, come grandi dimensioni, alta qualità costruttiva e specifiche amenities.

Nel complesso, la matrice di confusione mostra una buona simmetria degli errori e l'assenza di bias sistematici verso sovra- o sotto-classificazione. La confusione diretta tra *Low* e *High* riguarda solo il 5% delle osservazioni complessive di queste due classi, confermando che il modello distingue efficacemente gli estremi e concentra la maggior parte degli errori nei confini tra classi adiacenti.

7.7 Performance Assessment della Classificazione

Nel complesso, la performance del modello può essere valutata come *good*. Con tre classi equiprobabili, un'accuracy compresa tra il 70% e l'80% è generalmente considerata buona, mentre valori superiori all'85% sono richiesti per una performance eccellente. Il valore ottenuto si colloca solidamente nella fascia intermedia, risultando adeguato per applicazioni di supporto decisionale ma non sufficiente per sostituire completamente valutazioni professionali.

Il pattern degli errori è interpretabile e coerente con la natura del problema. Gli errori sono prevalentemente dovuti a confusioni tra classi contigue, mentre gli errori gravi sono rari. Questo rappresenta uno scenario ottimale dato che i confini tra fasce di prezzo sono intrinsecamente arbitrari e ambigui.

7.8 Confronto con gli Altri Task

Confrontando le performance dei tre task analizzati, emerge una chiara graduazione. La regressione ottiene risultati eccellenti, con un R^2 pari a 0.8856, sfruttando pienamente la natura continua del target. La classificazione si colloca in posizione intermedia, con performance solide ma penalizzate dalla discretizzazione. Il clustering, infine, mostra risultati buoni ma inferiori, riflettendo l'assenza di supervisione e la necessità di definire a priori il numero di cluster.

Questa graduazione riflette le caratteristiche intrinseche del dataset Ames Housing, che risulta particolarmente adatto alla regressione, adeguato alla classificazione e moderatamente adatto al clustering.

7.9 Discussione e Implicazioni

La principale lezione emersa dal task di classificazione riguarda il trade-off tra rappresentazione continua e discreta del prezzo. La discretizzazione introduce una perdita informativa che si manifesta soprattutto nei casi prossimi ai confini di classe. Circa il 40% degli errori coinvolge proprietà molto vicine alle soglie, suggerendo che tali misclassificazioni riflettono più l'arbitrarietà del binning che reali fallimenti predittivi.

Nonostante ciò, la classificazione offre vantaggi complementari rispetto alla regressione, come una maggiore interpretabilità per stakeholder non tecnici e un migliore allineamento con decisioni intrinsecamente categoriche. Tuttavia, la difficoltà nel gestire la classe *Medium* e l'assenza di una quantificazione esplicita dell'incertezza limitano l'utilizzo del modello in contesti critici.

7.10 Conclusioni sul Task di Classificazione

Il task di classificazione sul dataset Ames Housing produce risultati solidi ma non eccellenti. Con un'accuracy del 78.16% e un F1-score del 78.05%, il modello si dimostra adeguato come strumento di segmentazione e supporto decisionale, ma non come sostituto di un modello di regressione accurato. Le performance intermedie confermano che il dataset presenta una separabilità moderata in classi discrete e che la regressione rimane l'approccio più naturale ed efficace per il problema del pricing immobiliare.

8 Sintesi Comparativa

8.1 Confronto delle Performance tra i Task

L'analisi comparativa dei tre paradigmi adottati – regressione, clustering e classificazione – consente di trarre una visione d'insieme sulle capacità informative del dataset Ames Housing e sull'efficacia relativa dei diversi approcci analitici. Ciascun task affronta il problema della valutazione immobiliare da una prospettiva differente, con obiettivi, assunzioni e metriche proprie. Tuttavia, è possibile sintetizzare le performance ottenute in una tassonomia unificata che metta in evidenza punti di forza e limiti di ciascun paradigma.

La Tabella 1 riassume le metriche principali utilizzate per la valutazione dei tre task e il relativo giudizio complessivo.

Per facilitare un confronto diretto tra task fondati su metriche eterogenee, è utile ricondurre i risultati a una scala comune. Normalizzando le metriche principali su un intervallo $[0, 1]$, dove 1 rappresenta una performance ottimale, si ottiene un ordinamento quantitativo coerente: la regressione emerge come il paradigma più performante (0.886), seguita dalla classificazione (0.782) e infine dal clustering (0.728). È importante sottolineare che tale confronto non ha valore rigorosamente quantitativo, poiché le metriche rispondono a obiettivi diversi, ma fornisce comunque un'indicazione qualitativa robusta.

Le tre metriche considerate condividono alcune proprietà chiave: sono tutte limitate superiormente e inferiormente, assumono valori maggiori al crescere della qualità del modello e ammettono un'interpretazione probabilistica o geometrica. Questa parziale comunanza giustifica un confronto qualitativo, pur nel rispetto delle differenze concettuali tra i paradigmi.

8.2 Punti di Forza dei Diversi Paradigmi

L'analisi approfondita evidenzia come ciascun task presenti punti di forza distintivi, che riflettono sia le caratteristiche del dataset sia la natura degli algoritmi impiegati.

Nel caso della regressione, il principale vantaggio risiede nell'allineamento naturale con il fenomeno studiato. Il prezzo immobiliare è una variabile intrinsecamente continua e la regressione consente di modellarlo senza introdurre discretizzazioni artificiali, catturando differenze anche minime tra proprietà simili. L'utilizzo completo dell'informazione target, unito alla presenza di feature altamente predittive come *OverallQual*, *TotalSF* e *Neighborhood*, si traduce in una capacità esplicativa molto elevata. Le metriche adottate, quali R^2, RMSE e MAE, sono consolidate e facilmente interpretabili, rendendo i risultati immediatamente comunicabili anche a stakeholder non tecnici. La stabilità cross-fold osservata rafforza ulteriormente la fiducia nell'applicabilità pratica del modello.

Il clustering, pur mostrando performance inferiori in termini di metriche standard, presenta punti di forza di natura esplorativa. Operando in assenza di supervisione, esso evita bias introdotti da etichette arbitrarie e consente l'emergere di segmenti "naturali" guidati esclusivamente dalle feature. I cluster individuati risultano semanticamente interpretabili e allineati con segmentazioni consolidate del mercato immobiliare. La validazione ex-post tramite *SalePrice* conferma che tali cluster catturano effettivamente una stratificazione economica sottostante. Inoltre, il clustering si dimostra robusto rispetto a outlier nel target e utile per la scoperta di pattern multidimensionali e interazioni latenti.

La classificazione, infine, si distingue per la sua capacità di supportare decisioni categoriali dirette. Le tre classi di prezzo corrispondono a segmenti di mercato intuitivi e facilmente integrabili in processi di marketing e decision-making. Il bilanciamento naturale delle classi elimina problematiche tipiche della classificazione supervisionata, mentre la disponibilità di probabilità di classe consente una comunicazione esplicita dell'incertezza. Il pattern degli errori osservato, dominato da confusioni tra classi contigue e da una rarità di errori estremi, indica che il modello cattura correttamente l'ordinamento sottostante delle fasce di prezzo.

8.3 Idoneità del Dataset ai Diversi Task

Sulla base delle performance osservate, è possibile valutare l'idoneità del dataset Ames Housing per ciascun paradigma.

Per la regressione, l'idoneità è eccellente. Un valore di R^2 pari a 0.8856 supera ampiamente le soglie comunemente accettate per problemi socio-economici complessi. La ricchezza di feature predittive, la qualità del preprocessing, la dimensione campionaria adeguata e la stabilità cross-fold concorrono a rendere il dataset particolarmente adatto a questo tipo di analisi. Le limitazioni residue, come la varianza non spiegata e la performance leggermente inferiore agli estremi della distribuzione, riflettono fattori non osservabili piuttosto che carenze strutturali dei dati.

La classificazione mostra un'idoneità buona ma non ottimale. L'accuracy del 78% rappresenta una performance solida per un problema a tre classi equiprobabili, ma rimane significativamente inferiore a quella ottenuta in regressione. La classe *Medium* si conferma la più problematica, a causa della sua ambiguità intrinseca e dell'arbitrarietà dei confini di binning. Il dataset è quindi adeguato per applicazioni di teaching, benchmarking e supporto decisionale, ma la regressione rimane l'approccio preferibile per applicazioni operative critiche.

Il clustering evidenzia un'idoneità discreta. La presenza di una struttura cluster significativa è confermata da metriche e validazioni esterne, ma la separazione tra cluster non è sufficientemente marcata da rendere questo paradigma il focus primario dell'analisi. Il clustering risulta comunque utile come strumento esplorativo e di supporto ad analisi supervisionate successive.

8.4 Key Findings Cross-Task

L'analisi comparativa consente di identificare alcune evidenze trasversali. In primo luogo, poche feature aggregate – in particolare *OverallQual*, *TotalSF* e *Neighborhood* – catturano la maggior parte dell'informazione predittiva in tutti i task. Questo suggerisce che la qualità complessiva, la dimensione abitabile e la localizzazione rappresentano i driver fondamentali del valore immobiliare.

In secondo luogo, emerge una struttura latente dominata da una stratificazione dimensionale, affiancata da un gradiente qualitativo e da effetti di localizzazione e temporali. Le transizioni tra segmenti sono graduali piuttosto che discrete, riflettendo la natura continua del mercato immobiliare. Questa "fuzziness" intrinseca limita inevitabilmente le performance dei metodi che impongono segmentazioni rigide, come la classificazione multi-classe e il clustering.

L'analisi della separabilità delle classi conferma che le classi estreme (*Low* e *High*) sono ben distinte, mentre la classe *Medium* presenta un overlap significativo con entrambe. Le distanze di Mahalanobis e gli indici di overlap quantificano questa osservazione e spiegano perché gli errori gravi siano rari mentre le confusioni ai confini siano frequenti. Il gap di performance tra modelli lineari e non lineari evidenzia inoltre la presenza di interazioni complesse che richiedono algoritmi flessibili.

8.5 Analisi Costi–Benefici dei Paradigmi

Oltre alle performance pure, è rilevante considerare i trade-off in termini di effort, interpretabilità e applicabilità pratica. La regressione offre il miglior equilibrio complessivo, combinando elevata accuratezza, interpretabilità diretta delle predizioni e facilità di validazione. La classificazione rappresenta un'alternativa valida per specifici use case orientati al decision support, mentre il clustering fornisce principalmente valore esplorativo e di insight discovery.

8.6 Raccomandazioni Strategiche

Alla luce dell'analisi comparativa, si raccomanda un approccio multi-paradigma. Per applicazioni pratiche di valutazione immobiliare, la regressione dovrebbe costituire il metodo primario, eventualmente affiancata da modelli di classificazione per screening rapido e comunicazione semplificata. Il clustering dovrebbe essere utilizzato nelle fasi iniziali di esplorazione dei dati o come strumento di supporto per il feature engineering.

In conclusione, il dataset Ames Housing dimostra un'eccellente idoneità per la regressione, una buona adattabilità alla classificazione e una capacità discreta di supportare analisi di clustering. Questa graduazione riflette fedelmente la natura del problema e suggerisce che il massimo valore analitico si ottiene combinando i diversi paradigmi in modo complementare.

9 Conclusioni

9.1 Riepilogo del Lavoro Svolto

Il presente lavoro ha costituito un'esplorazione metodologica completa del dataset Ames Housing attraverso tre paradigmi distinti di machine learning: regressione supervisionata, clustering non supervisionato e classificazione multi-classe. L'obiettivo non è stato l'applicazione meccanica di algoritmi standard, ma una valutazione critica e comparativa dell'idoneità del dataset rispetto a ciascuna prospettiva analitica, mettendo in relazione scelte metodologiche, caratteristiche dei dati e risultati ottenuti.

Il percorso è stato strutturato in modo progressivo e coerente. L'analisi esplorativa iniziale ha permesso di comprendere la natura dei dati, evidenziandone al contempo il potenziale informativo e le criticità. Il dataset, composto da quasi tremila transazioni immobiliari relative alla città di Ames, Iowa, si è rivelato ricco e articolato, ma caratterizzato da un'elevata eterogeneità delle variabili, molte delle quali categoriche e ad alta cardinalità. Questa complessità ha reso necessario un preprocessing attento e non puramente standard.

La fase di ETL e preprocessing ha rappresentato uno snodo centrale dell'intero lavoro. Oltre alle operazioni di pulizia di base, è stato svolto un intenso lavoro di feature engineering guidato dalla conoscenza del dominio immobiliare. La creazione di variabili aggregate come *TotalSF*, in grado di sintetizzare la superficie complessiva abitabile, o di variabili temporali come *HouseAge*, costruite in termini relativi anziché assoluti, ha permesso di arricchire il contenuto informativo dei dati. In parallelo, la gestione delle variabili categoriche ad alta cardinalità tramite target encoding out-of-fold ha consentito di preservare informazione predittiva che approcci più ingenui avrebbero disperso.

Un ulteriore passaggio metodologico rilevante è stato rappresentato dalla trasformazione logaritmica della variabile target. Tale scelta è stata motivata dalla forte asimmetria della distribuzione dei prezzi immobiliari e ha consentito di ridurre drasticamente la skewness, rendendo il problema più trattabile per modelli che assumono una distribuzione approssimativamente normale degli errori. Questo intervento ha avuto un impatto positivo trasversale su tutti i task analizzati.

La selezione delle feature mediante mutual information ha permesso di ridurre dimensionalità e rumore, concentrando l'analisi sulle quaranta variabili più informative. Questa scelta ha migliorato non solo le performance ma anche la leggibilità dei modelli, evitando un'eccessiva complessità dovuta all'uso indiscriminato di tutte le feature generate durante il preprocessing.

Per tutti i task è stata adottata una validazione incrociata a cinque fold, al fine di stimare in modo robusto le performance di generalizzazione. La stabilità dei risultati tra i fold ha rafforzato l'affidabilità delle conclusioni e ridotto il rischio di valutazioni ottimistiche legate a singoli split favorevoli.

Nel task di regressione, il Random Forest ha raggiunto un valore di R^2 pari a 0.886, indicando una capacità di spiegare quasi il novanta per cento della varianza del prezzo. Questo risultato si colloca nella fascia alta delle performance comunemente riportate in letteratura per problemi di house price prediction e conferma l'elevata idoneità del dataset per questo tipo di analisi. L'analisi delle importanze ha evidenziato il ruolo dominante di *OverallQual* e *TotalSF*, che catturano rispettivamente la dimensione qualitativa e quella dimensionale del valore immobiliare.

Il clustering ha offerto una prospettiva complementare. In assenza di supervisione, l'algoritmo K-Means è riuscito a individuare una struttura latente coerente, come indicato

da un Silhouette Score pari a 0.457. I cluster risultanti sono interpretabili e riconducibili a segmenti di mercato plausibili, che spaziano da proprietà entry-level compatte a abitazioni premium di ampia metratura. La validazione ex-post tramite il prezzo di vendita ha confermato che tali cluster riflettono una reale stratificazione economica.

La classificazione ha rappresentato una via intermedia tra la granularità della regressione e l'esplorativà del clustering. La discretizzazione del prezzo in tre classi equipopulate ha prodotto un problema ben bilanciato, con un'accuracy media del 78%. Sebbene inferiore alle performance della regressione, questo risultato è coerente con la perdita informativa introdotta dalla discretizzazione. L'analisi degli errori ha mostrato che la maggior parte delle confusioni avviene tra classi contigue, mentre errori estremi risultano rari.

9.2 Valutazione Complessiva del Dataset

Considerando l'intero percorso analitico, emerge un quadro differenziato dell'idoneità del dataset Ames Housing. Per la regressione, il giudizio è nettamente positivo. Il dataset è stato progettato con finalità predittive e include variabili che riflettono in modo diretto i determinanti economici del valore immobiliare. La combinazione di qualità, dimensione, localizzazione e fattori temporali rende i dati particolarmente adatti a modelli supervisionati continui.

Per il clustering, la valutazione è più sfumata. Il dataset presenta una struttura clusterizzabile, ma tale struttura è dominata principalmente da un asse dimensionale e qualitativo principale, con confini sfumati piuttosto che netti. Questa caratteristica riflette la natura continua del mercato immobiliare e limita l'efficacia di metodi che impongono partizioni discrete.

La classificazione si colloca in posizione intermedia. Le performance sono solide e utilizzabili, ma non eccellenti. La difficoltà maggiore risiede nella classe intermedia, che soffre di ambiguità intrinseca dovuta a confini di binning artificiali. Ciò non invalida l'approccio, ma ne chiarisce i limiti applicativi.

Un elemento trasversale emerso con forza è l'importanza del preprocessing e del feature engineering. In tutti i task, le scelte operate in queste fasi preliminari hanno avuto un impatto maggiore rispetto alla selezione fine degli algoritmi, confermando che la qualità dei dati è spesso il fattore determinante del successo modellistico.

9.3 Raccomandazioni per l'Utilizzo

Per finalità didattiche, il dataset si presta in modo eccellente allo studio della regressione, permettendo di esplorare l'intero ciclo di vita di un progetto di machine learning in un contesto realistico ma gestibile. Allo stesso tempo, la presenza di variabili categoriche complesse e di distribuzioni asimmetriche lo rende adatto anche a trattazioni più avanzate.

Per utilizzi di benchmarking o ricerca metodologica, il dataset offre un buon equilibrio tra pulizia e complessità. Pur non essendo ideale per testare tecniche pensate per dati estremamente rumorosi, rappresenta un banco di prova significativo per modelli non lineari, approcci ensemble e strategie di encoding avanzate.

In contesti applicativi, la regressione emerge come scelta primaria per la valutazione immobiliare, mentre classificazione e clustering possono fornire valore complementare in attività di screening, segmentazione o supporto decisionale. In tutti i casi, è essenziale considerare i limiti di trasferibilità geografica e temporale del modello.

9.4 Punti di Forza, Limitazioni e Considerazioni Finali

Tra i principali punti di forza dell'approccio adottato figurano la sistematicità metodologica, l'integrazione della conoscenza di dominio, la prospettiva comparativa multiparadigma e l'attenzione alla validazione e all'interpretazione dei risultati. Questi elementi hanno contribuito a produrre conclusioni robuste e contestualizzate.

Le limitazioni principali riguardano l'assenza di ottimizzazione sistematica degli iperparametri, l'esplorazione limitata del panorama algoritmico, l'analisi dell'interpretabilità non approfondita e la mancanza di validazione esterna. Tali scelte sono state deliberate e coerenti con l'obiettivo del lavoro, ma rappresentano naturali direzioni di estensione futura.

In conclusione, il lavoro ha raggiunto l'obiettivo prefissato di valutare in modo rigoroso l'idoneità del dataset Ames Housing per diversi paradigmi di machine learning. Il dataset si conferma fortemente adatto alla regressione, adeguato alla classificazione e moderatamente idoneo al clustering. Questa graduazione riflette la natura intrinsecamente continua del fenomeno studiato e suggerisce che il massimo valore analitico si ottiene adottando un approccio complementare che integri più paradigmi piuttosto che affidarsi a uno solo.

Riferimenti bibliografici

[1] D. De Cock, *Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project*, Journal of Statistics Education, vol. 19, no. 3, 2011.

[2] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, Springer, 2015.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[5] I. T. Jolliffe, *Principal Component Analysis*, Springer, 2002.

[6] R. Kohavi, *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, IJCAI, 1995.

[7] F. Pedregosa et al., *Scikit-learn: Machine Learning in Python*, JMLR, 2011.

Tabella 1: Riepilogo comparativo delle performance per i tre task

Task	Metrica primaria	Valore	Metrica secondaria	Valore	Metrica terziaria	Verd
Regressione	R^2	0.8856	RMSE (log)	0.1426	MAE (log)	EXCEL
Clustering	Silhouette	0.4567	Davies–Bouldin	0.9123	Inertia	GO
Classificazione	Accuracy	0.7816	F1-score (w)	0.7805	Precision	GO