Documentazione Progetto

Machine Learning

|  |  |
| --- | --- |
| Versione | 1.4 |
| Data | 16/01/2025 |
| Destinatario | Dipartimento di Informatica dell’Università degli studi di Salerno |
| Presentato da | Carla Stefanile |

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Data** | **Versione** | **Descrizione** | **Autori** |
| 15/01/2025 | 1.0 | Intestazioni del documento | Carla Stefanile |
| 16/01/2025 | 1.1 | Prima stesura | Carla Stefanile |
| 17/01/2025 | 1.2 | Aggiunta valori metriche di valutazione dei vari modelli | Carla Stefanile |
| 18/01/2025 | 1.3 | Prima revisione | Carla Stefanile |
| 25/01/2025 | 1.4 | Seconda revisione | Carla Stefanile |

Ruoli e Contatti

|  |  |  |
| --- | --- | --- |
| **Ruolo** | **Nome** | **Contatti** |
| Docente | Giuseppe Polese | gpolese@unisa.it |
| Docente | Loredana Caruccio | lcaruccio@unisa.it |
| Studente | Carla Stefanile | c.stefanile1@studenti.unisa.it |

Sommario

[1. Introduzione 4](#_Toc188715668)

[2. Data Collection 4](#_Toc188715669)

[2.1. Comprensione e Identificazione Dati Necessari 4](#_Toc188715670)

[2.2. Dataset 5](#_Toc188715671)

[2.3. Data Exploration 5](#_Toc188715672)

[3. Model Selection 5](#_Toc188715673)

[3.1. Random Forest Regressor 5](#_Toc188715674)

[3.2. Decision Tree Regressor 6](#_Toc188715675)

[3.3. Linear Regression 6](#_Toc188715676)

[4. Data Manipulation & Execution 6](#_Toc188715677)

[4.1. Preprocessing & Feature-Selection 6](#_Toc188715678)

[4.2. Pipeline 6](#_Toc188715679)

[5. Evaluation ed Analisi Comparativa 6](#_Toc188715680)

[5.1. Metriche di Valutazione 7](#_Toc188715681)

[5.2. Risultati 7](#_Toc188715682)

[6. Conclusioni 12](#_Toc188715683)

# Introduzione

L'obiettivo del progetto è quello di costruire un sistema di previsione accurato per i prezzi immobiliari nella città di Londra, utilizzando tecniche di Machine Learning. Londra rappresenta uno dei mercati immobiliari più complessi e costosi al mondo, caratterizzato da una grande variabilità nei prezzi dovuta a fattori geografici, dimensionali e strutturali.

Questo progetto implementa una pipeline completa che integra preprocessamento dei dati, selezione di modelli, ottimizzazione degli iperparametri e valutazione delle prestazioni.

Il lavoro segue le seguenti fasi principali:

* Raccolta e comprensione del dataset.
* Preprocessing e manipolazione dei dati.
* Scelta di modelli adatti e ottimizzazione.
* Valutazione delle metriche di performance.
* Interpretazione dei risultati per comprendere i punti di forza e di debolezza dei modelli impiegati.

# Data Collection

La raccolta dei dati è una fase cruciale in qualsiasi progetto di Machine Learning. Per costruire un sistema di previsione accurato, è essenziale disporre di dati rilevanti, completi e coerenti. In questa sezione vengono descritte le fonti dati utilizzate, i criteri di selezione e l'esplorazione preliminare. La comprensione delle caratteristiche principali del dataset consente di identificare le variabili più significative per il modello, garantendo la solidità delle analisi future.

## Comprensione e Identificazione Dati Necessari

L’analisi preliminare ha identificato i dati fondamentali per stimare i prezzi immobiliari:

* Dimensioni della proprietà.
* Numero di camere da letto e bagni.
* Codice postale e area geografica.
* Prezzo medio delle proprietà nell'area.

## Dataset

Il dataset utilizzato è stato pubblicato su Kaggle da Mehmet Emre Sezer il 25/12/2024 e contiene 29.537 record. Le caratteristiche includono:

* **Price**: Affitto mensile in GBP.
* **Property Type**: Classificazione della proprietà in cinque categorie principali.
* **Bedrooms**: Numero di camere da letto.
* **Bathrooms**: Numero di bagni.
* **Size**: Dimensioni in piedi quadrati.
* **Postcode**: Codice postale (216 categorie).
* **Area**: Area generale (17 categorie).
* **Area\_Avg\_Price**: Prezzo medio delle proprietà nell’area.

## Data Exploration

Sono stati generati grafici per analizzare le distribuzioni delle variabili principali come "Price" e "Size". Le distribuzioni sono risultate skewed, richiedendo trasformazioni logaritmiche.

L'analisi dei dataset ha confermato la coerenza dei dati e l'assenza di valori mancanti. I dati sono stati ulteriormente validati per identificare outlier.

L'analisi ha evidenziato la necessità di preprocessare i dati per migliorarne la qualità e ottimizzare le performance dei modelli.

# Model Selection

La scelta del modello rappresenta un passaggio fondamentale per il successo di un progetto di Machine Learning. In questa sezione vengono analizzati i modelli considerati, ognuno dei quali è stato scelto in base alla sua capacità di adattarsi alle caratteristiche del problema. Verranno descritti il Random Forest Regressor, il Decision Tree Regressor e la Linear Regression, con un focus sulle loro caratteristiche, punti di forza e motivazioni per il loro utilizzo.

## Random Forest Regressor

Un metodo di ensemble che utilizza alberi decisionali addestrati su sottoinsiemi casuali di dati e caratteristiche. Questo approccio riduce l'overfitting e migliora la capacità di generalizzazione.

I principali iperparametri ottimizzati includono:  
 - n\_estimators (numero di alberi).  
 - max\_depth (profondità massima degli alberi).  
 - min\_samples\_split (numero minimo di campioni per uno split).  
 - min\_samples\_leaf (campioni minimi in una foglia).  
 - max\_features (numero massimo di caratteristiche per split).  
.

## Decision Tree Regressor

Un singolo albero decisionale che suddivide iterativamente i dati in base a condizioni sulle caratteristiche. Modello interpretabile, utile per analisi preliminari e confronto.

## Linear Regression

Un modello lineare che stima una relazione lineare tra le caratteristiche indipendenti e la variabile target. Utilizzato come baseline, non richiede ottimizzazione degli iperparametri.

# Data Manipulation & Execution

La manipolazione dei dati e l'esecuzione della pipeline sono elementi centrali per trasformare i dati grezzi in input ottimali per i modelli predittivi. In questa sezione vengono descritte le tecniche di preprocessing adottate, la selezione delle caratteristiche più rilevanti e lo sviluppo della pipeline per garantire efficienza e riproducibilità. Sarà fornita una panoramica sul flusso operativo, dalla preparazione dei dati fino all'addestramento e alla validazione dei modelli.

## Preprocessing & Feature-Selection

* Eliminazione di outlier.
* Raggruppamento di categorie rare in una singola classe "Others".
* Trasformazioni logaritmiche su "Size", "Bathrooms" e "Area\_Avg\_Price".
* Standardizzazione delle variabili numeriche.

## Pipeline

È stata sviluppata una pipeline, per automatizzare il preprocessamento e l’addestramento dei modelli, garantendo efficienza e riproducibilità, addestrata sul dataset suddividendolo in tre porzioni, 70% per l’addestramento, 15% per la validation, 15% per il testing.

# Evaluation ed Analisi Comparativa

La valutazione è una fase essenziale per determinare le prestazioni dei modelli e confrontarli in termini di accuratezza, errore e capacità predittiva. In questa sezione verranno illustrate le metriche utilizzate per l'analisi, come MAE e R², e saranno discussi i risultati ottenuti. L'analisi comparativa consente di identificare il modello migliore e di valutare l'efficacia delle tecniche di ottimizzazione applicate.

## Metriche di Valutazione

Per ciascuna esecuzione dei modelli verranno calcolati:

1. **MAE** (Mean Absolute Error): Media delle differenze assolute tra previsioni e valori reali.
2. **R²**: Percentuale di varianza spiegata dal modello.
3. **Cross-Validation MAE**: Media degli errori assoluti nelle iterazioni di validazione incrociata.

## Risultati

La valutazione delle performance dei modelli è stata effettuata confrontando tre metriche chiave: CV MAE, MAE, e R².

L'ottimizzazione ha avuto un impatto positivo soprattutto sul modello Random Forest e Decision Tree, con miglioramenti evidenti nelle metriche di errore (CV MAE e MAE) e nella capacità esplicativa (R²).

* Random Forest: Miglioramenti significativi in tutte le metriche.
* Decision Tree: Ottimizzazione mirata ha ridotto l'errore e incrementato l'R².
* Linear Regression: Le prestazioni sono rimaste invariate, poiché il modello non è stato ulteriormente ottimizzato.

Questi risultati dimostrano come l'ottimizzazione dei parametri influisca positivamente su modelli non lineari, migliorandone l'efficacia nel rappresentare i dati e riducendo gli errori di previsione.

Immagine che contiene schermata, linea, diagramma, Diagramma

Descrizione generata automaticamente

Immagine che contiene schermata, linea, diagramma, Diagramma

Descrizione generata automaticamente

Immagine che contiene schermata, linea, diagramma, Diagramma

Descrizione generata automaticamente

Immagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamente

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamenteImmagine che contiene testo, schermata, numero, diagramma

Descrizione generata automaticamente

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamenteImmagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

Immagine che contiene schermata, diagramma, testo

Descrizione generata automaticamenteImmagine che contiene testo, schermata, linea, diagramma

Descrizione generata automaticamente

Immagine che contiene testo, schermata, numero, diagramma

Descrizione generata automaticamente

# Conclusioni

Il progetto ha dimostrato come tecniche avanzate di Machine Learning possano essere applicate con successo per prevedere i prezzi immobiliari in un mercato complesso come quello di Londra. L'intero processo, dalla raccolta e preprocessamento dei dati alla selezione e valutazione dei modelli, ha evidenziato l'importanza di un approccio sistematico e ben pianificato.

I risultati hanno confermato il valore aggiunto dell'ottimizzazione degli iperparametri, soprattutto per modelli complessi come il Random Forest Regressor, che si è rivelato il più performante. Allo stesso tempo, il confronto tra modelli lineari e non lineari ha fornito utili spunti sull'efficacia di approcci diversi in base alla natura dei dati.

Da questa analisi, possiamo concludere che, sebbene il modello Random Forest sia efficace nel prevedere il prezzo della maggior parte delle proprietà, ci sono alcune aree di miglioramento. In particolare, il modello sottoperforma per proprietà di tipo Flat e Apartment, con dimensioni ridotte e situate in aree di fascia alta come 'Eastern' e 'Western and Paddington'. Miglioramenti potrebbero includere l'integrazione di variabili più dettagliate relative alla localizzazione o caratteristiche interne delle proprietà.

Le principali lezioni apprese includono:

* L'importanza del preprocessamento per migliorare la qualità dei dati e delle previsioni.
* Il valore della validazione incrociata per garantire la generalizzazione dei modelli.
* Il ruolo cruciale dell'ottimizzazione nel migliorare le prestazioni di modelli complessi.

Grazie infinite dell’attenzione.