

# RegressorComparison

Daniele Russo

January 8, 2024

## Contents

<b>1</b>	<b>Introduzione</b>	<b>3</b>
1.1	Analisi e contestualizzazione del problema . . . . .	3
1.2	Proposta della soluzione . . . . .	3
1.3	Obiettivi e Criteri di Successo . . . . .	3
<b>2</b>	<b>Specifiche dell'ambiente</b>	<b>4</b>
2.1	Enviroment . . . . .	4
2.2	Actuators . . . . .	4
2.3	Sensors . . . . .	4
<b>3</b>	<b>Modello di sviluppo</b>	<b>4</b>
3.1	Team Member: . . . . .	5
<b>4</b>	<b>Comprensione dei Dati</b>	<b>5</b>
4.1	Raccolta, Origine e Formato dei Dati . . . . .	5
4.2	Analisi Preliminare . . . . .	6
<b>5</b>	<b>Preparazione dei Dati</b>	<b>6</b>
5.1	Pulizia dei Dati . . . . .	6
5.2	Gestione dei Dati Mancanti . . . . .	6
5.3	Trasformazione dei Dati . . . . .	6
5.4	Normalizzazione . . . . .	6
5.5	Codifica delle Variabili Categorie . . . . .	7
<b>6</b>	<b>Modellazione</b>	<b>7</b>
6.1	Selezione e analisi delle Caratteristiche . . . . .	7
6.1.1	Nuove Tabelle di Correlazione . . . . .	7
6.1.2	Riduzione Dimensionale . . . . .	7
6.2	Scelta del Modello . . . . .	7
6.2.1	Tipi di Modelli Considerati . . . . .	7
<b>7</b>	<b>Valutazione</b>	<b>9</b>
7.1	Misurazione delle Prestazioni . . . . .	9
7.1.1	Metriche di Valutazione Utilizzate . . . . .	9
7.1.1.1	Assunzioni . . . . .	10
7.2	Validazione incrociata: . . . . .	10
<b>8</b>	<b>Deploy e Implementazione</b>	<b>10</b>
8.1	Scelta del Linguaggio . . . . .	10
8.2	Framework e Strumenti Utilizzati . . . . .	11
8.2.1	Dipendenze necessarie . . . . .	11
8.2.2	Link utili: . . . . .	11
8.3	QuickStart progetto - Primo avvio . . . . .	12

8.4	Struttura del progetto . . . . .	12
8.5	Normalizer . . . . .	12
8.6	Agente . . . . .	14
8.7	AgentFarm . . . . .	15
8.8	Main . . . . .	16
<b>9</b>	<b>Conclusioni e Pianificazione Futura</b>	<b>16</b>
9.1	Risultati . . . . .	16
9.2	Successi e Sfide . . . . .	25
9.3	Sviluppi Futuri e Miglioramenti Possibili . . . . .	25

# 1 Introduzione

## 1.1 Analisi e contestualizzazione del problema

Oggi giorno, ci troviamo sempre di più circondati da tool che utilizzano intelligenza artificiale, ma questi necessitano di dati che, molte volte, sono manchevoli, parziali, o semplicemente si vuole un modo per stimare dei valori discreti, ma non sappiamo come. Per questo, si ricorre all'utilizzo dei regressori per completare i propri dati. Ma come capiamo quale tipo di regressore utilizzare? Di qui nasce l'idea di un comparatore generico che renda facile analizzare le prestazioni dei vari tipi di regressori, tale idea viene affiancata ad un caso d'uso pratico: partendo da dati riguardanti lavoratori, riuscire a stimare i "Benefit" ovvero i sussidi aggiuntivi allo stipendio.

La regressione è una tecnica statistica utilizzata per comprendere la relazione tra una variabile dipendente continua e una o più variabili indipendenti. In termini semplici, un regressore cerca di modellare la funzione che meglio si adatta ai dati osservati, consentendo di effettuare previsioni o stime su valori continui.

Ma come funziona un regressore?

1. **Input dei dati:** Il regressore richiede un insieme di dati di addestramento, composti da coppie di input e output. Gli input sono le variabili indipendenti, mentre gli output sono i corrispondenti valori della variabile dipendente.
2. **Addestramento del modello:** Durante la fase di addestramento, il regressore cerca di apprendere la relazione tra gli input e gli output del dataset. Utilizzando un algoritmo di apprendimento, il modello ottimizza i suoi parametri in modo da minimizzare l'errore tra le previsioni e i valori effettivi.
3. **Generazione della funzione di regressione:** Una volta addestrato, il regressore genera una funzione di regressione che rappresenta la relazione stimata tra le variabili indipendenti e la variabile dipendente.
4. **Previsioni:** Il regressore può quindi essere utilizzato per fare previsioni su nuovi dati o dati non visti in precedenza. Utilizzando la funzione di regressione, il modello stima i valori della variabile dipendente in base ai nuovi input.

## 1.2 Proposta della soluzione

La soluzione proposta dal candidato è affiancata a un caso d'uso pratico, che prende il nome di "RegressorComparator", dove viene applicata l'idea del comparatore di regressori ad un problema pratico, data la situazione di un lavoratore possiamo stimare i suoi "Benefit"? I problemi di regressione sono istanze di problemi di apprendimento supervisionato, quindi in maniera simile alla classificazione, un problema di regressione porta alla costruzione di un modello, ovvero di uno strumento che fa uso di un algoritmo di apprendimento, anche detto regressore, per predire i nuovi elementi sulla base del training set. I regressori sono essenzialmente delle funzioni matematiche che cercano di descrivere i dati. Diversi regressori si distinguono tra di loro per via delle assunzioni fatte sui dati così come delle specifiche proprietà che portano alla regressione, ma anche del numero di variabili indipendenti (predittori) di cui si dispone.

## 1.3 Obiettivi e Criteri di Successo

**Obiettivi del Progetto:** Il progetto mira a sviluppare un modo facile per confrontare i modelli di regressione in grado di predire in modo accurato i "Benefit" dei dipendenti. L'obiettivo è di sopprimere la mancanza di esperienza e rendere accessibile tale tecnologia.

**Risultati Attesi:** Ci aspettiamo che almeno uno dei modelli sviluppati raggiungano un adattamento del modello superiore al 90%, predicendo con successo i "Benefit".

## 2 Specifiche dell'ambiente

L'ambiente è un'istanza del problema per la quale l'agente rappresenta la soluzione. La descrizione di un ambiente viene generalmente redatta tramite l'utilizzo della formula PEAS, caratterizzata dai seguenti quattro fattori:

- **Performance** (prestazioni): Per la seguente sezione, era riduttivo dedicare un sottoparagrafo; per tale motivo, le è stato dedicato un approfondimento nel capitolo 6.
- **Environment** (ambiente)
- **Actuators** (attuatori)
- **Sensors** (sensori)

Analizziamo, quindi, punto per punto questi fattori relativamente al problema.

### 2.1 Enviorment

L'environment è la descrizione degli elementi che formano l'ambiente. Considerando, come già detto, che l'ambiente è un'istanza del problema per la quale l'agente rappresenta una soluzione, l'ambiente sarà composto da una entry del dataset (come sono strutturate le entry verrà specificato in seguito). Le seguenti specifiche ci permettono di definire l'environment in esame:

- L'ambiente è **completamente osservabile**, in quanto i sensori dell'agente (rappresentati in questo caso da una funzione, come specificato in seguito) hanno completo accesso all'ambiente in ogni momento;
- L'ambiente è **deterministico** poiché lo stato dell'ambiente dipende dall'azione eseguita dall'agente;
- L'ambiente è **sequenziale** perché l'azione dell'agente viene eseguita solamente alla fine dell'analisi dell'ambiente;
- L'ambiente è **statico** dal momento che rimane invariato mentre l'agente delibera;
- L'ambiente è **discreto** poiché fornisce un numero limitato di percezioni ben distinte;
- L'ambiente è **singolo** in quanto è presente un unico agente.

### 2.2 Actuators

Gli attuatori sono gli strumenti che ha a disposizione un agente per effettuare le azioni. Nell'ambito del progetto, l'azione che dovrà compiere l'agente (che ricordiamo essere un regressore), sarà predire il valore della variabile dipendente di una entry. In questo caso quindi, l'attuatore dell'agente sarà una funzione che permetterà di assegnare il valore al campo "Benefit" di una determinata persona.

### 2.3 Sensors

I sensori sono gli strumenti che l'agente utilizza per prendere in input la percezione dell'ambiente. Nel nostro caso è una entry del dataset, ovvero prenderà in input tutte le variabili indipendenti per valutare e predire la variabile dipendente.

## 3 Modello di sviluppo

Il progetto ha seguito il modello di sviluppo CRISP-DM, che è riuscito a fornire agilità e flessibilità per il corretto completamento del progetto. In generale, sono state affrontate tutte le fasi in modo iterativo, consentendo di portare facilmente a termine il progetto:

- **Comprendere del Business** (Business Understanding)
  - Definizione degli obiettivi del business.

- Traduzione degli obiettivi in obiettivi di data mining.
- Valutazione della situazione corrente.

- **Comprensione dei Dati** (Data Understanding):

- Raccolta dei dati necessari per il progetto.
- Caratterizzazione iniziale dei dati.
- Esplorazione preliminare dei dati per comprendere la loro natura.

- **Preparazione dei Dati** (Data Preparation):

- Pulizia dei dati, gestione dei valori mancanti e rimozione degli outlier.
- Selezione delle variabili rilevanti.
- Trasformazione dei dati per renderli adatti all’analisi.

- **Modellazione** (Modeling):

- Selezione delle tecniche di modellazione più adeguate.
- Creazione di modelli usando i dati di addestramento.
- Validazione e ottimizzazione dei modelli.

- **Valutazione** (Evaluation):

- Valutazione delle prestazioni dei modelli rispetto agli obiettivi del business.
- Se necessario, revisione e ripetizione delle fasi precedenti.

- **Deploy** (Deployment):

- Implementazione dei modelli nell’ambiente di produzione.
- Monitoraggio delle prestazioni dei modelli implementati.

- **Pianificazione del Monitoraggio** (Monitoring):

- Verifica costante delle prestazioni dei modelli implementati.
- Aggiornamento dei modelli o del processo di data mining se necessario.

### 3.1 Team Member:

Il progetto è stato interamente pensato, progettato e sviluppato da:

- Russo Daniele

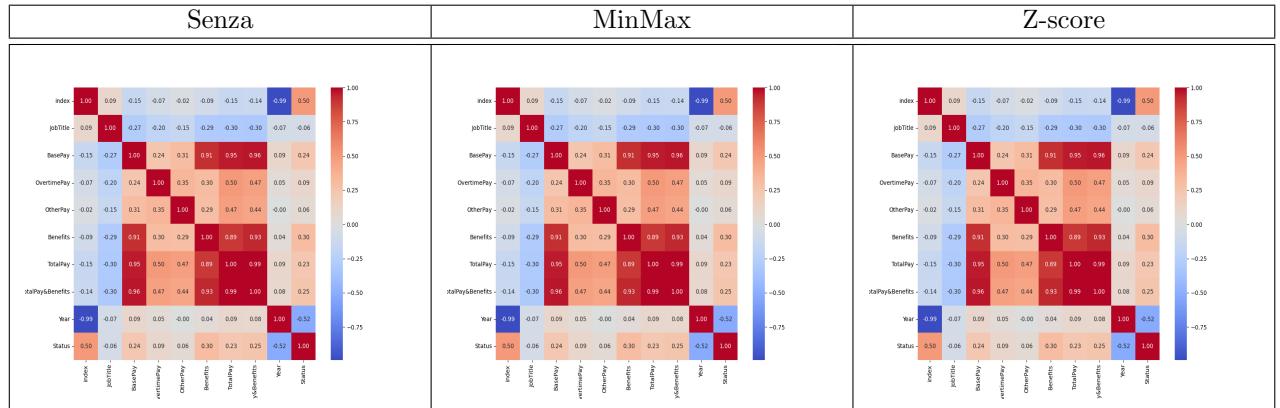
## 4 Comprensione dei Dati

### 4.1 Raccolta, Origine e Formato dei Dati

Un regressore di machine learning trae insegnamenti dai dati, per stimare nel nostro caso d’uso, i ”Benefit”. Però richiedono un considerevole numero di esempi per un addestramento efficace e per affinare la capacità predittiva del modello. Di qui si intuisce la fondamentale importanza della quantità di dati da avere a disposizione. Per stimare i ”Benefit”, è necessaria una raccolta di dati che comprenda informazioni specifiche su ciascun esempio, insieme alle relative etichette che indicano il valore effettivo dei ”Benefit”. La ricerca di un dataset appropriato, ha portato all’esplorazione della community Kaggle, dove sono disponibili diversi dataset contenenti informazioni sulla situazione stipendiaria di varie classi di lavoratori. Successivamente a un’analisi approfondita, è emerso che i dataset che permettano l’utilizzo di regressori, specialmente quelli di tipo lineare, risultano essere pochi. Di conseguenza, la scelta è ricaduta su un dataset composto da colonne che rappresentano i salari dei dipendenti pubblici della California dal 2011 al 2019. Con i seguenti campi: Employee Name, Job Title, Base Pay, Overtime Pay, Other Pay, Benefits, Total Pay, TotalPay&Benefits, Year, Status. Il dataset è composto da circa 3.216.663 righe e 10 colonne. Di conseguenza, è stato effettuato un lungo studio sui dati, che verrà approfondito nella fase di Data Understanding.

## 4.2 Analisi Preliminare

Durante il nostro corso, abbiamo studiato 2 tipologie di regressori: i regressori lineari e gli alberi decisionali regressivi. Per applicare la prima tipologia, abbiamo bisogno di determinate condizioni dei dati, che verranno approfondite successivamente. È importante verificare, affinché il regressore sia affidabile, che la variabile dipendente che vogliamo predire, ossia i "Benefit", sia esprimibile in funzione delle altre variabili, quelle indipendenti. In questa sezione, verificheremo l'esistenza di tale condizione, ovvero che i "Benefit" sia esprimibile in funzione delle altre Feature, attraverso la tabella di correlazione. Tutto il lavoro è stato svolto anche con differenti normalizzazioni dei dati, per verificare l'impatto di tale scelta.



Come possiamo osservare il campo "Benefit" ha una forte correlazione con diverse variabili tra le quali: BasePay, TotalPay e TotalPay&Benefit. Mentre con le altre ha una correlazione medio-alta come: OvertimePay, OtherPay e Status. Con le restanti la correlazione è pressoché inesistente.

## 5 Preparazione dei Dati

## 5.1 Pulizia dei Dati

In questa fase, andremo a sostituire tutti i campi nulli, indicati con Nan, con una stringa facilmente individuabile ovvero "Not provided". Inoltre rimuoviamo le colonne che non sono di nostro interesse tra cui: Employee Name, Total Pay, TotalPay&Benefits.

## 5.2 Gestione dei Dati Mancanti

Vista la grande quantità di dati, possiamo rimuovere tutti i campi nulli o vuoti. Nello specifico si andranno a rimuovere tutte le tuple con i campi "Benefit" e "BasePay" uguali a zero.

### 5.3 Trasformazione dei Dati

I regressori lavorano solo con dati numerici; per tale motivo è stata effettuata una sostituzione dei campi string con un numero per cui è stato creato un dizionario che ne permetta l'interpretazione. Tutto ciò è fornito da un altro modulo presente in `/util/normalizer.py` .

## 5.4 Normalizzazione

Siccome l'obiettivo è la fruizione di questi algoritmi a neofiti in questo campo, ogni test viene eseguito 3 volte, ognuna con un tipo di normalizzazione sui dati. Nel particolare sono utilizzati:

- Dati grezzi: dati così presenti nel dataset.

- MinMax:  $x' = a + \frac{(x - \min(x))(b - a)}{\max(a) - \min(x)}$

Dove  $a$  e  $b$  rappresentano i valori minimo e massimo che vogliamo ottenere dalla normalizzazione.

- Z-score:  $x' = \frac{x - \bar{x}}{\sigma}$

Dove  $x$  è il valore originale, e  $\bar{x}$  è la media della distribuzione e  $\sigma$  è la deviazione standard.

## 5.5 Codifica delle Variabili Categoriche

Nel dataset sono presenti due colonne contenenti variabili di categoria: JobTitle e Status. Quando verrà avviata l'util `normalizer`, questa creerà la cartella `/dataset`, al cui interno troveremo il dataset normalizzato e un file txt. All'interno di quest'ultimo, troviamo tutte le sostituzioni avvenute.

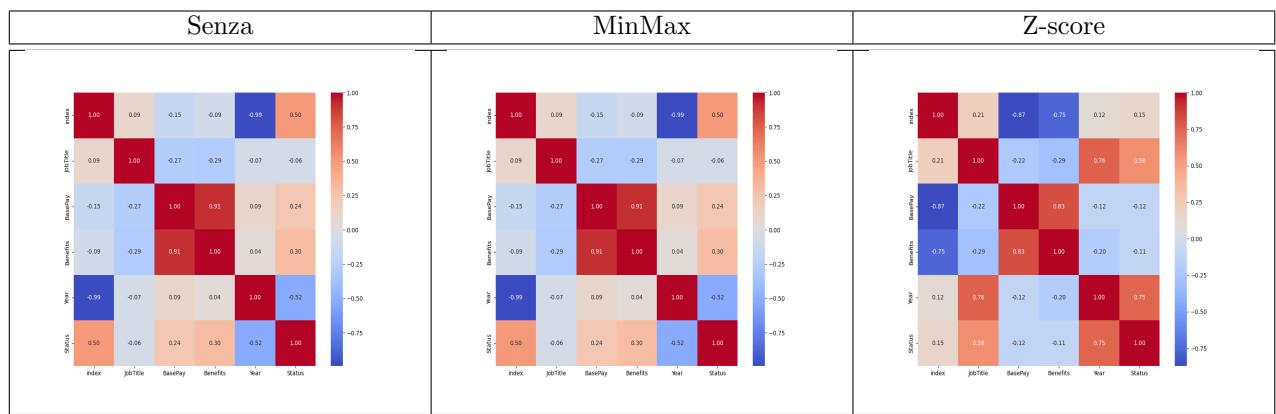
# 6 Modellazione

## 6.1 Selezione e analisi delle Caratteristiche

Ovviamente non manterremo tutte le colonne che il dataset dispone, ma manterremo solo quelle con un'alta correlazione con "Benefit" ovvero: JobTitle, BasePay, OvertimePay, OtherPay, Benefits, Year, Status.

### 6.1.1 Nuove Tabelle di Correlazione

Dopo la fase di pulizia, per una maggior accuratezza, sono state ri-create le tabelle di correlazione, per capire se si fossero estratte le feature giuste.



### 6.1.2 Riduzione Dimensionale

Vista la grande mole di dati presenti nel dataset, per poter comparare tutti i regressori, si è preso in considerazione il 2% del dataset, ovvero 64.332 istanze. Questa scelta è stata obbligata, in quanto l'utilizzo di più istante, portava la macchina sulla quale è stato eseguito il programma, ad un bottleneck, sia per la memoria RAM, dove si sono toccati picchi di 60GB richiesti, sia di CPU in quanto non tutti gli algoritmi sono di default parallelizzabili.

## 6.2 Scelta del Modello

Sono stati presi in esame tutti i modelli messi a disposizione dalla libreria scick-learn, di seguito ne è fornita una lista.

### 6.2.1 Tipi di Modelli Considerati

- **LinearRegression:**

È un modello di regressione lineare, che cerca di trovare la migliore retta di regressione per adattarsi ai dati

- **LARS:**  
Prova la soluzione del modello di regressione lineare in modo incrementale, aggiungendo le variabili più influenti ad ogni passo.
- **Ridge:**  
È una variante della regressione lineare, che introduce una regolarizzazione L2 per mitigare il rischio di overfitting.
- **LassoLars:**  
Utilizza il metodo LARS con regolarizzazione L1 (LASSO) per ottenere modelli di regressione sparsi.
- **ARDRegression - Automatic Relevance Determination:**  
Utilizza una combinazione di regolarizzazione L2 e L1 per determinare automaticamente la rilevanza delle variabili.
- **SGDRegressor:**  
Utilizza la discesa del gradiente stocastica per addestrare modelli di regressione in modo efficiente su grandi set di dati.
- **BayesianRidge:**  
È un regressore basato sulla teoria bayesiana, incorpora conoscenze a priori nel processo di apprendimento.
- **GaussianProcessRegressor:**  
Utilizza processi gaussiani per modellare la distribuzione delle previsioni. È adatto per problemi di regressione non lineari.
- **TweedieRegressor:**  
Modella i dati con distribuzione di Tweedie, adatto per problemi di regressione con dati a dispersione.
- **DecisionTreeRegressor:**  
È un modello basato su alberi decisionali, che suddivide iterativamente i dati in base alle caratteristiche, predice il valore medio di ciascun nodo foglia.
- **RandomForestRegressor:**  
Una foresta di alberi decisionali che aggredisce le previsioni di diversi alberi per ottenere una previsione più robusta.
- **KNeighborsRegressor:**  
Stima i valori target basandosi sui vicini più prossimi in uno spazio delle feature.
- **RadiusNeighborsRegressor:**  
È simile a KNeighborsRegressor, ma stima i valori target basandosi su un raggio di vicini invece di un numero fisso.
- **SVR - Support Vector Regression:**  
È una variante della regressione lineare che utilizza support vector machines per trovare la migliore retta di regressione.
- **LinearSVR:**  
È una versione lineare di SVR, utile per problemi di regressione lineare con grandi dataset.
- **NuSVR:**  
È una variante di SVR con un parametro aggiuntivo (nu) che controlla il numero di support vectors.

La scelta del regressore, dipende dalla natura del problema e dalle caratteristiche dei dati. Ogni tipo di regressore ha vantaggi e svantaggi specifici, e la scelta dovrebbe essere guidata dalla comprensione del contesto e degli obiettivi del problema di regressione che si sta affrontando.

## 7 Valutazione

### 7.1 Misurazione delle Prestazioni

#### 7.1.1 Metriche di Valutazione Utilizzate

Per valutare la bontà di un regressore prenderemo come metriche di riferimento:

- **MAE - Mean Absolute Error:**

$$\frac{\sum_{i=1}^n |\tilde{y} - y|}{n}$$

La MAE è una metrica che misura la media assoluta degli errori tra le previsioni di un modello e i valori effettivi. È calcolata sommando le differenze assolute tra le previsioni e i valori reali, e quindi dividendo per il numero totale di osservazioni.

**Obiettivo:** Minimizzare. Valori più bassi, indicano una maggiore precisione del modello.

- **MSE - Mean Squared Error:**

$$\frac{\sum_{i=1}^n (\tilde{y} - y)^2}{n}$$

L'MSE è una metrica che misura la media dei quadrati degli errori tra le previsioni di un modello e i valori effettivi. È calcolata sommando le differenze al quadrato tra le previsioni e i valori reali, e quindi dividendo per il numero totale di osservazioni.

**Obiettivo:** Minimizzare. Valori più bassi, indicano una maggiore precisione del modello.

- **RMSE - Root Mean Squared Error:**

$$\sqrt{\frac{\sum_{i=1}^n (\tilde{y} - y)^2}{n}}$$

Il RMSE è la radice quadrata dell'MSE. È spesso utilizzato per ottenere una misura dell'errore in unità originali dei dati, poiché è nell'unità della variabile dipendente.

**Obiettivo:** Minimizzare. Valori più bassi, indicano una maggiore precisione del modello.

- **$R^2$  - Coefficient of determination:**

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Dove:

$$SS_{res} = \sum_{i=1}^n (y_i - \tilde{y}_i)$$

$$\bar{y} = \frac{1}{n} - \sum_{i=1}^n y_i$$

$$SS_{tot} = \sum_i^n (y_i - \bar{y})^2$$

Misura la proporzione di varianza nella variabile dipendente che è spiegata dal modello. Può assumere valori compresi tra 0 e 1, dove 1 indica un perfetto adattamento del modello ai dati.

**Obiettivo:** Massimizzare. Un  $R^2$  più alto, indica un modello più adatto, mentre un  $R^2$  vicino a 0 indica che il modello non spiega bene la variabilità dei dati.

- **EVS - Explained variance score:**

$$1 - \frac{Var\{y - \bar{y}\}}{Var\{y\}}$$

L'EVS è una metrica che misura la proporzione della varianza totale della variabile dipendente che è spiegata dal modello. Come  $R^2$ , assume valori compresi tra 0 e 1.

**Obiettivo:** Massimizzare. Un punteggio più vicino a 1, indica una migliore capacità del modello di spiegare la varianza nei dati.

#### 7.1.1.1 Assunzioni

Sebbene sia semplice e talvolta efficace, la regressione lineare può essere utilizzata solo in determinati contesti.

- **Linearità dei dati.** La relazione tra la variabile indipendente X e la variabile dipendente Y deve essere lineare, ovvero può essere rappresentata tramite una funzione lineare.
- **Normalità dei residui.** Gli errori residui devono essere normalmente distribuiti.
- **Omoschedasticità.** Gli errori residui devono avere una varianza costante. Questa può essere verificata andando a plottare i residui standardizzati vs i valori predetti. Se la proprietà è soddisfatta, vedremo un trend orizzontale piuttosto che punti sparsi nello spazio.
- **Indipendenza degli errori.** Gli errori residui devono essere indipendenti per ogni valore di X. Un test statistico particolarmente utile è noto come Durbin-Watson: quando gli errori sono indipendenti, il valore del test sarà vicino a 2.

## 7.2 Validazione incrociata:

La scelta di utilizzare la validazione incrociata è stata motivata dalla necessità di ottenere una stima più affidabile delle prestazioni del modello. La validazione incrociata, fornisce una valutazione più robusta delle prestazioni del modello rispetto alla suddivisione tradizionale. Ho optato per una K-fold cross-validation con K=10, ripetuta 4 volte. Questa scelta è stata basata sulla dimensione complessiva del dataset e sulla necessità di bilanciare la varianza e la computazionalità.

## 8 Deploy e Implementazione

In questa sezione vedremo le scelte affrontate e le motivazioni dietro queste ultime, e come tutte le fasi precedenti siano poi state sviluppate in codice e come esso sia stato strutturato.

### 8.1 Scelta del Linguaggio

La scelta di Python come linguaggio di programmazione per lo sviluppo di un regressore è motivata da diverse ragioni che ne fanno una delle opzioni più preferite nell'ambito del machine learning e dell'analisi dei dati. Ecco alcune motivazioni chiave:

- **Ricca Libreria di Machine Learning:** Python dispone di librerie ampie e mature per il machine learning, tra cui scikit-learn, TensorFlow e PyTorch. Queste librerie offrono implementazioni efficienti di diversi algoritmi di regressione e forniscono strumenti per valutare, ottimizzare e validare i modelli.
- **Versatilità e Integrazione:** Python è un linguaggio versatile che può essere utilizzato in diverse fasi di un progetto, dalla manipolazione dei dati all'implementazione dei modelli e alla creazione di interfacce utente. La sua capacità di integrarsi facilmente con altri linguaggi e tecnologie è un vantaggio significativo.

- **Ricchezza di Strumenti di Visualizzazione:** Python offre una vasta gamma di librerie di visualizzazione dei dati, come Matplotlib e Seaborn, che semplificano la rappresentazione grafica dei risultati del regressore. La visualizzazione è fondamentale per comprendere e comunicare efficacemente i risultati ottenuti.

In sintesi, la combinazione di una vasta libreria di machine learning, una comunità attiva, facilità di apprendimento e versatilità, fa di Python una scelta solida e popolare per lo sviluppo di regressori e modelli di machine learning in generale.

## 8.2 Framework e Strumenti Utilizzati

Tra le varie tecnologie e framework precedentemente citati, si è scelto per l'implementazione dell'agente la libreria **scikit-learn**. Per la data-visualization invece si è ricorso alla libreria **Matplotlib** e **Seaborn**. Per l'utilizzo del dataset sottoforma di dataframe in modo facile e intuitivo, si è utilizzato **pandas**. Per la creazione delle statistiche, si è ricorso alla libreria **statsmodels**.

### 8.2.1 Dipendenze necessarie

Appurate tutte le dipendenze del progetto, sarebbe utile in vista di utilizzare tale progetto, risolvere le seguenti dipendenze, di seguito sono riportate tutti gli snippet per installare le librerie necessarie:

- **Pandas:**

```
pip install pandas
```

- **Joblib:**

```
pip install joblib
```

- **Scikit-learn:**

```
pip install scikit-learn
```

- **Matplotlib:**

```
pip install matplotlib
```

- **Seaborn:**

```
pip install seaborn
```

- **Statsmodels:**

```
pip install statsmodels
```

### 8.2.2 Link utili:

Di seguito verrà riportata una breve lista di link che posso essere un ottimo spunto per approfondire le varie tematiche in modo singolo:

- **GitHub:**

[Link al progetto](#)

- **Kaggle:**

[Link al dataset](#)

- **Scikit-learn:**

[Link alla libreria Scikit-learn](#)

- **Pandas:**

[Link alla libreria Pandas](#)

- **Matplotlib:**

[Link alla libreria Matplotlib](#)

- **Statsmodels:**

[Link alla libreria Statsmodels](#)

- **Seaborn:**

[Link alla libreria seaborn](#)

- **Joblib:**

[Link alla libreria joblib](#)

### 8.3 QuickStart progetto - Primo avvio

In questa sezione verrà spiegato come il progetto funziona logicamente astraendo tutta la parte di programmazione che verrà sviluppata nel modulo successivo. Dopo aver scaricato il progetto e risolto tutte le dipendenze, si procede con l'avvio del `normalizer.py`, presente nella cartella `/util`. Il compito di tale classe, oltre a rimuovere la colonna "Employee Name" per una questione di ottimizzazione, è quello di effettuare la sostituzione per i campi di tipo stringa presenti nel dataset. Questo perché i regressori accettano solo numeri come input. Questo modulo come output produrrà due nuovi file "newDataset.csv" e "indexSostitution.txt" entrambi nella cartella `/dataset`. Dove:

- `newDataset`: contiene il nuovo dataset, privo di campi stringa.
- `indexSostitution`: è un dizionario contenente per ogni valore di ogni colonna, il campo sostituito con il relativo indice, così da poter ricostruire il dataset originale eventualmente.

Una volta termina l'esecuzione del `normalizer.py`. Qui possiamo avviare il `main.py` qui verranno fatte tutte le fasi di datacleaning, featurescaling, featureSelection per poi eseguire la comparazione tra i vari tipi di regressore; questo viene effettuato tre volte, una per ogni tipo di normalizzazione. Al termine dell'esecuzione, si sarà creata una nuova cartella `/analysis`, nella quale troveremo una cartella per ogni algoritmo, all'interno della quale una per ogni normalizzazione dove troveremo: il grafico della distribuzione dell'errore, il grafico per la variazione dell'errore e un report nel quale troviamo tutte le metriche al suo interno.

### 8.4 Struttura del progetto

Il progetto è stato creato affinché sia facilmente adattabile a più casistiche possibili. Ovviamente andranno modificati i nomi delle colonne presenti sia nel `normalizer.py` e sia `main.py`. Di seguito verrà approfondito singolarmente ogni modulo per consentire a chiunque un profonda comprensione degli stessi e l'aumento delle possibilità di riutilizzo e ampliamento del codice da me prodotto.

### 8.5 Normalizer

Suddividiamo il modulo in 4 blocchi principali:

```
## trasformo il vecchio dataset in una matrice
dataframe=pd.read_csv("./dataset/san-francisco-payroll_2011-2019.csv")
dataframe=dataframe.drop(columns=["Employee Name"])
#Sostituiamo i valori Nan con NotProvided per evitare di avere problemi con il dizionario
print("Number of instance Nan :" + str(dataframe.isna().sum().sum()))
dataframe=dataframe.replace(np.nan, value="Not Provided", regex=True)

data=[]
data.append([header for header in dataframe.columns])
for liste in dataframe.values.tolist():
    data.append(liste)
```

Figure 1: Primo blocco

In questa prima sezione, con l'aiuto di Pandas, viene caricato il dataframe e lo si converte in una matrice.

```

## inizializzo un dizionario che mi indicerà per ogni categoria l'ultimo indice inserito
## inizializzo il dizionario che conterrà un dizionario per ogni categoria dove c'è una stringa
for i in range(len(data[0])):
    if type(data[1][i]) is str:
        if not data[1][i].replace(".", "", 1).replace(".", "", 1).isnumeric() or not data[1][i].replace(".", "", 1).replace(".", "", 1).isalpha():
            last[data[0][i]] = 0
            listaDiSostituzioni[data[0][i]] = dict()

```

Figure 2: Secondo blocco

In questa seconda sezione, creiamo il dizionario delle sostituzioni e inizializziamo un dizionario dove, per ogni colonna che verrà sostituita, avremo l'ultimo indice utilizzato.

```

for i in range(1,len(data)):
    for j in range(0,len(data[i])):
        if data[0][j] in listaDiSostituzioni.keys():
            if str(data[i][j]).lower() not in listaDiSostituzioni[data[0][j]].keys():
                listaDiSostituzioni[data[0][j]][data[i][j].lower()] = last[data[0][j]]+1
                last[data[0][j]]+=1
                data[i][j]=listaDiSostituzioni[data[0][j]][data[i][j].lower()]
            else:
                if data[i][j]=="Not Provided":
                    data[i][j]=0
                else:
                    if "," in str(data[i][j]) or "." in str(data[i][j]):
                        # print("5")
                        data[i][j] = float(data[i][j])
                    else:
                        # print("6")
                        data[i][j] = int(data[i][j])

```

Figure 3: Terzo blocco

In questa sezione, per ogni riga della tabella, controlliamo se la colonna della cella in esame è presente nel dizionario delle sostituzioni, in caso positivo andiamo a verificare se è un valore sostituito in precedenza o invece è un nuovo valore da aggiungere; in caso negativo capiamo se il valore è un intero o un decimale e lo salviamo.

```

try:
    os.mkdir("../dataset")
except OSError as e:
    pass

newFile=open("../dataset/newDataset.csv","w")

for i in range(0,len(data)):
    newFile.write(str(data[i]).removesuffix("]").removeprefix("[").replace("old: \"", "new: \"").replace("old: '", "new: '"))
newFile.close()

indexSostitution = open("../dataset/indexSostitution.txt","w")

for key in listaDiSostituzioni.keys():
    indexSostitution.write(key+str(listaDiSostituzioni[key])+"\n")
indexSostitution.close()

```

Figure 4: Quarto blocco

In quest'ultimo blocco, creiamo la cartella dove andremo a salvare i due file: newDataset e indexSostitution.

## 8.6 Agente

Il modulo Agent.py implementa una serie di metodi dei quali verrà fornita la firma e una descrizione:

- `Agent(type:str,n_job:int)`

Questo metodo prende in input il tipo di regressore che l'agente deve utilizzare e il numero di thread che il sistema può mettere a disposizione per aumentare le prestazioni.

- **fit(X\_train,y\_train)**  
È un metodo con il quale l'agente effettua l'addestramento, prendendo in input le variabili indipendenti X e la variabile dipendente y.
- **predict(X\_test)**  
In questo metodo l'agente, sulla base di quanto ha imparato, predice il risultato della variabile dipendente y.
- **valuation(y\_test, pred)**  
In questo metodo, sulla base dei valori predetti, vengono restituite le metriche che sono state indicate in precedenza.
- **cross\_validation(X\_train, y\_train)**  
In questo metodo, effettuiamo una kcross validation 4 volte, per poi ritornare le metriche indicate in precedenza.

## 8.7 AgentFarm

Il modulo `AgentFarm.py` implementa una serie di metodi dei quali è fornita di seguito la firma e una descrizione:

- **AgentFarm(dataframe:pd.DataFrame,n\_job:int):**  
Al costruttore passiamo il dataset e il numero di thread che il modulo potrà utilizzare.
- **dataCleaning(listaRimossi:list):**  
In questo metodo eliminiamo, tutte le tuple che hanno i campi presenti in listaRimossi uguali a zero.
- **correlazioneVariabili(lable:str):**  
In questo metodo, creiamo il grafico di correlazione delle variabili nella cartella /analysis/TabellaDiCorrelazione. Il campo lable è formattato come normalizzazione\_value, dove value è "before" o "after" per indicare se è stato chiamato prima o dopo la featureSelection.
- **featureScaling(normalizzazione:str):**  
Questo metodo a seconda della stringa passata, effettuerà o meno la normalizzazione sui dati.
- **featureSelection(listaRimossi:list,percentageTest:float):**  
In questo metodo, eliminiamo le colonne presenti in listaRimossi, di solito con poca varianza e dividiamo il dataset in test e trainig secondo la percentuale passata.
- **startComparison(listaAgent:list):**  
Per ognuno dei regressori presenti in listaAgenti, eseguiamo la fase di fit, prediction,valuation e cross\_validation. In questa fase vengono prodotti, per ogni normalizzazione, i seguenti file:
  - **distribuzioneErroreResidio:** Dove vediamo graficamente come è distribuito l'errore residuo.
  - **varianzaErroreResiduo:** Dove vediamo graficamente la differenza tra l'errore residuo standardizzato e i valori previsti.
  - **Report.txt:** In questo file sono riportati: i valori per le metriche standard, quelle ottenute con la K cross validation e i test statistici per capire:
    - \* **Normalità del errore residuo:** vengono effettuati i test di: Shapiro-Wilk, Kolmogorov-Smirnov e di Anderson-Darling.
    - \* **Indipendenza degli errori residui:** viene eseguito il test di Durbin-Watson.

## 8.8 Main

Questa sezione è divisa in due parti:

```
#prepariamo il data frame da pandas
dataframe=pd.read_csv(path+"/"+file_name)
#print(dataframe.info(memory_usage='deep'))
print(dataframe.size)
dataframe = dataframe.sample(frac=0.02, random_state=42)
print(dataframe.size)

agentComparison = ["LinearRegression", "Ridge", "SGDRegressor", "LARS", "LassoLars",
                    "BayesianRidge", "ARDRegression", "TweedieRegressor",
                    "DecisionTreeRegressor", "RandomForestRegressor", "KNeighborsRegressor",
                    "RadiusNeighborsRegressor", "GaussianProcessRegressor", "SVR", "NuSVR", "LinearSVR"]
typeNormalization = ["StandardScaler", "MinMaxScaler", "None"]
listaRimossi=list({TotalPay, TotalPay&Benefits})
listaCleaning=[BasePay, Benefits]
n_job=8
```

Figure 5: Prima parte

In questa prima parte, preleviamo il dataset normalizzato e lo riduciamo di size per poi creare tutte le liste di cui abbiamo bisogno.

```
for norm in typeNormalization:
    print("Normalizzando: "+norm)
    farm=AgentFarm(dataframe,n_job)
    farm.dataCleaning(listaCleaning)
    farm.correlazioneVariabili(norm+"_before")
    farm.featureScaling(norm)
    farm.featureSelection(listaRimossi, percentageTest: 0.50)
    farm.correlazioneVariabili(norm+"_after")
    with parallel_backend('threading', n_jobs=n_job):
        farm.startComparison(agentComparison)
```

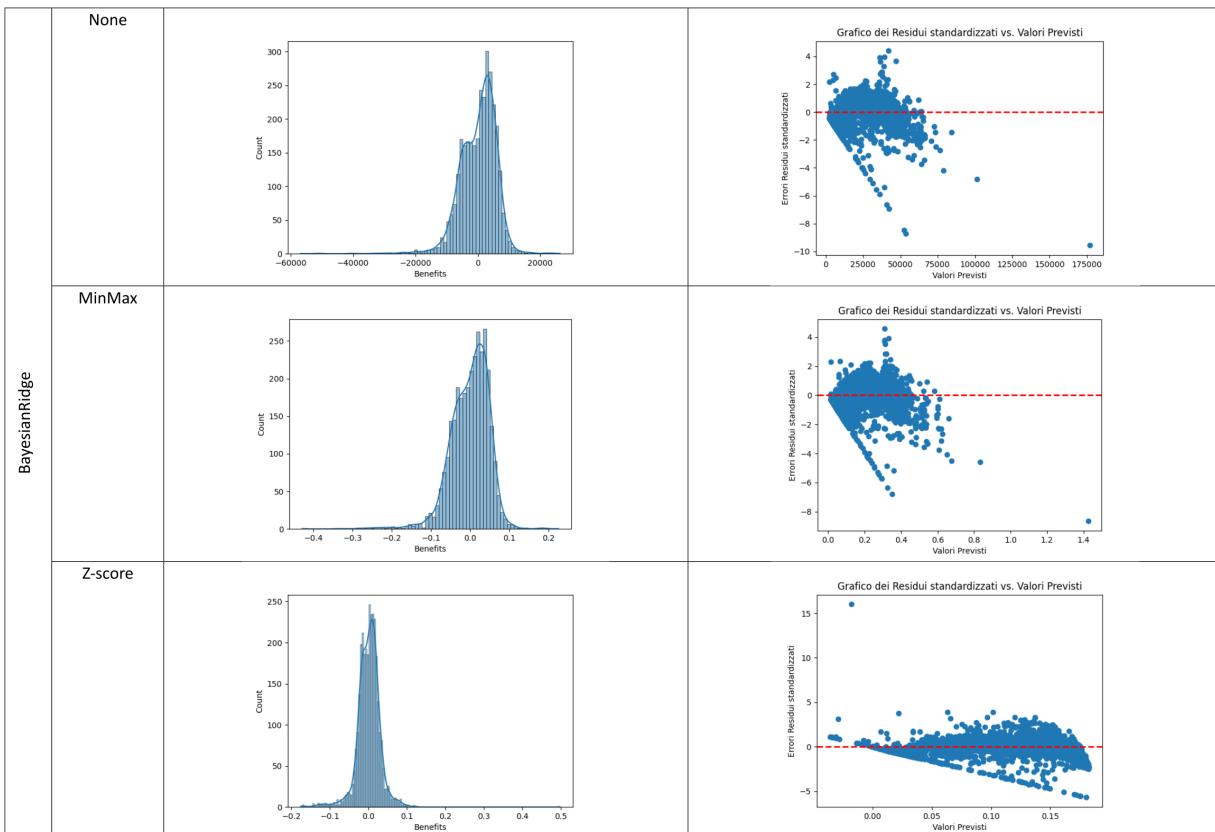
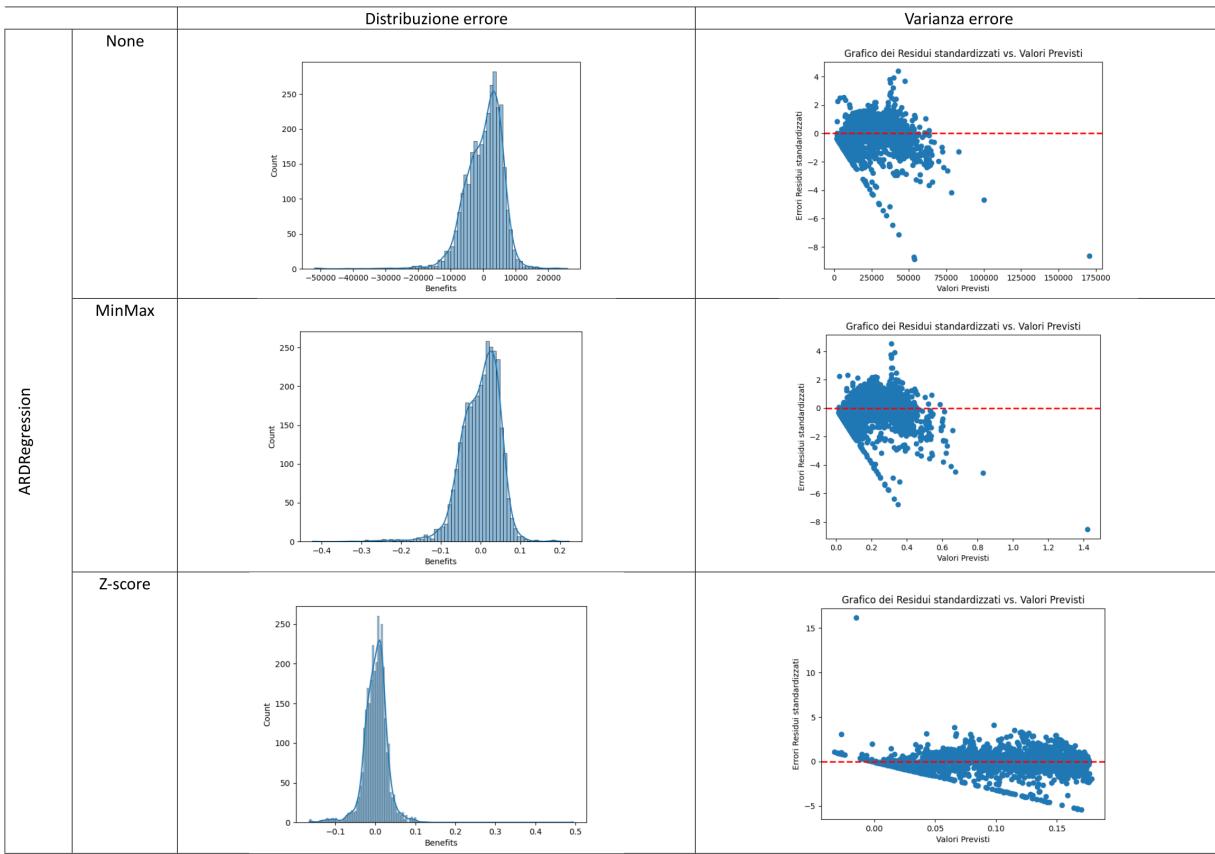
Figure 6: Seconda parte

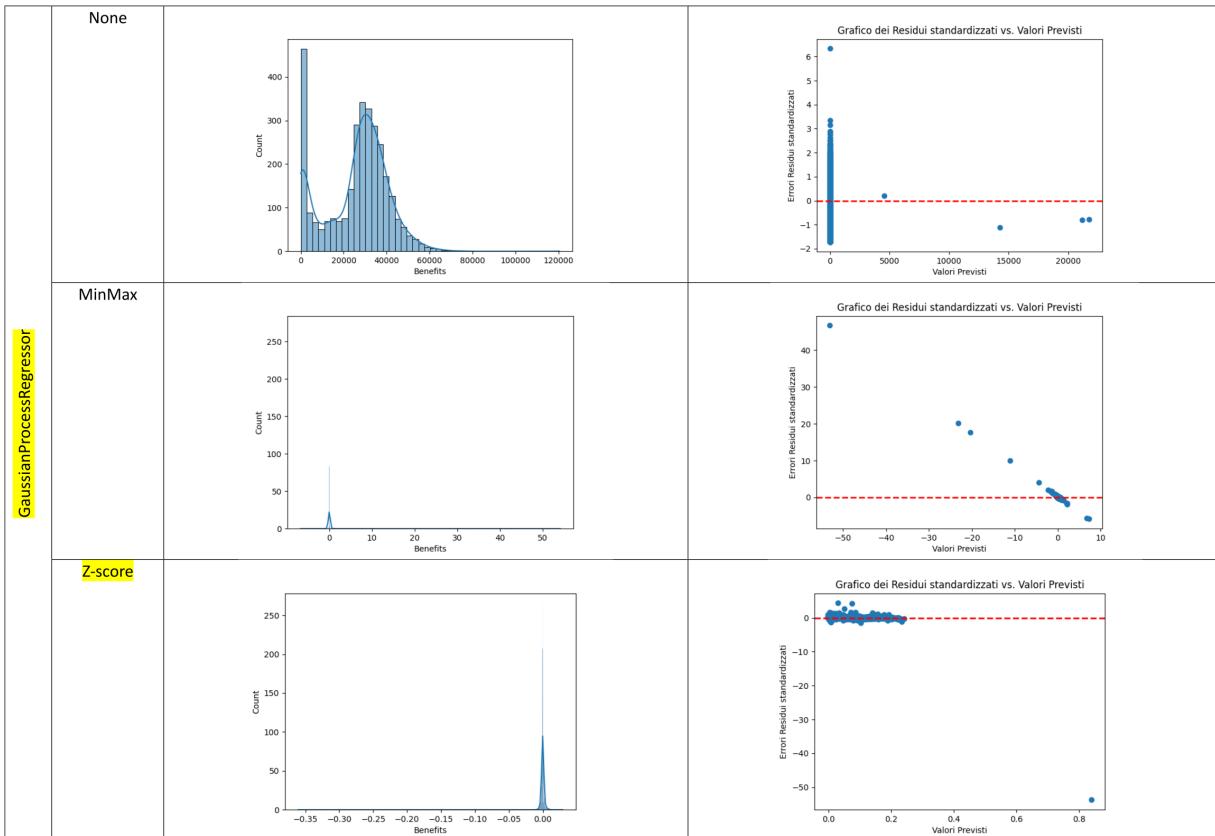
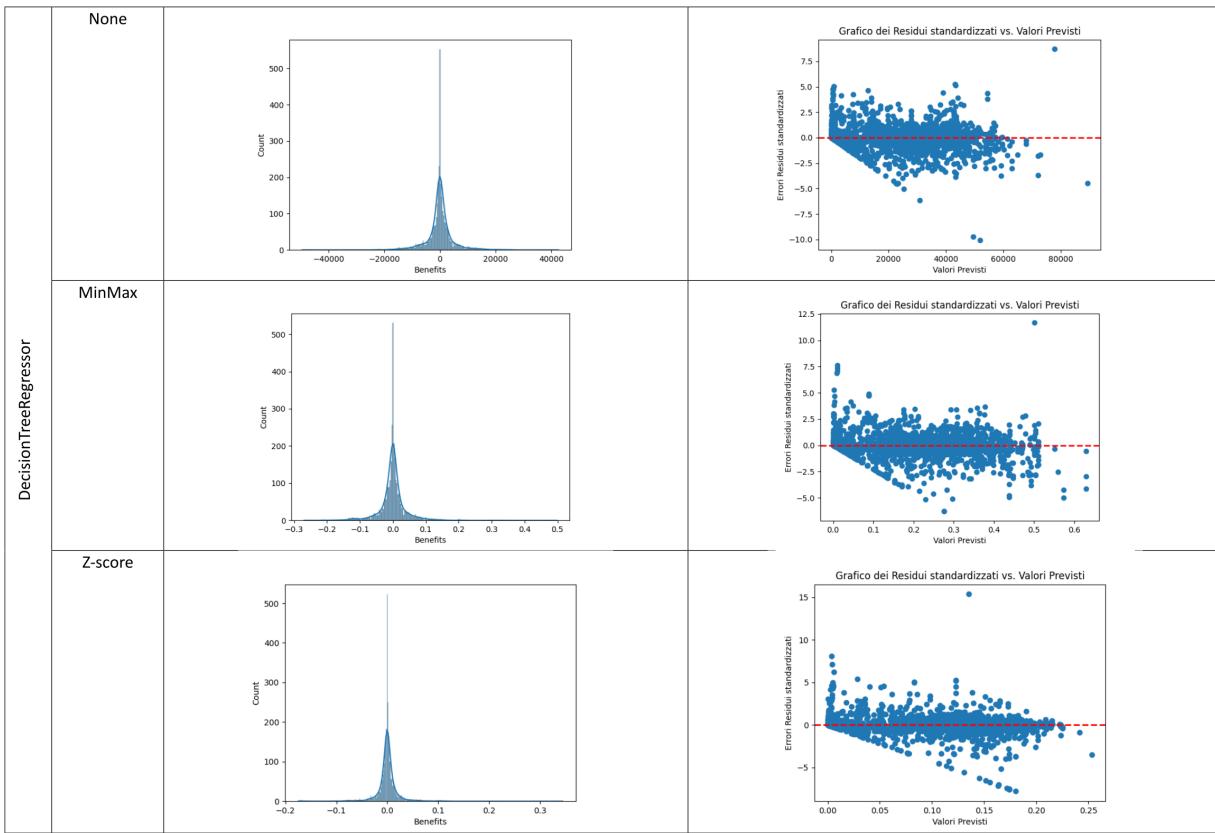
In quest'ultima, sezione per ogni normalizzazione facciamo eseguire alla farm tutte le fasi precedenti.

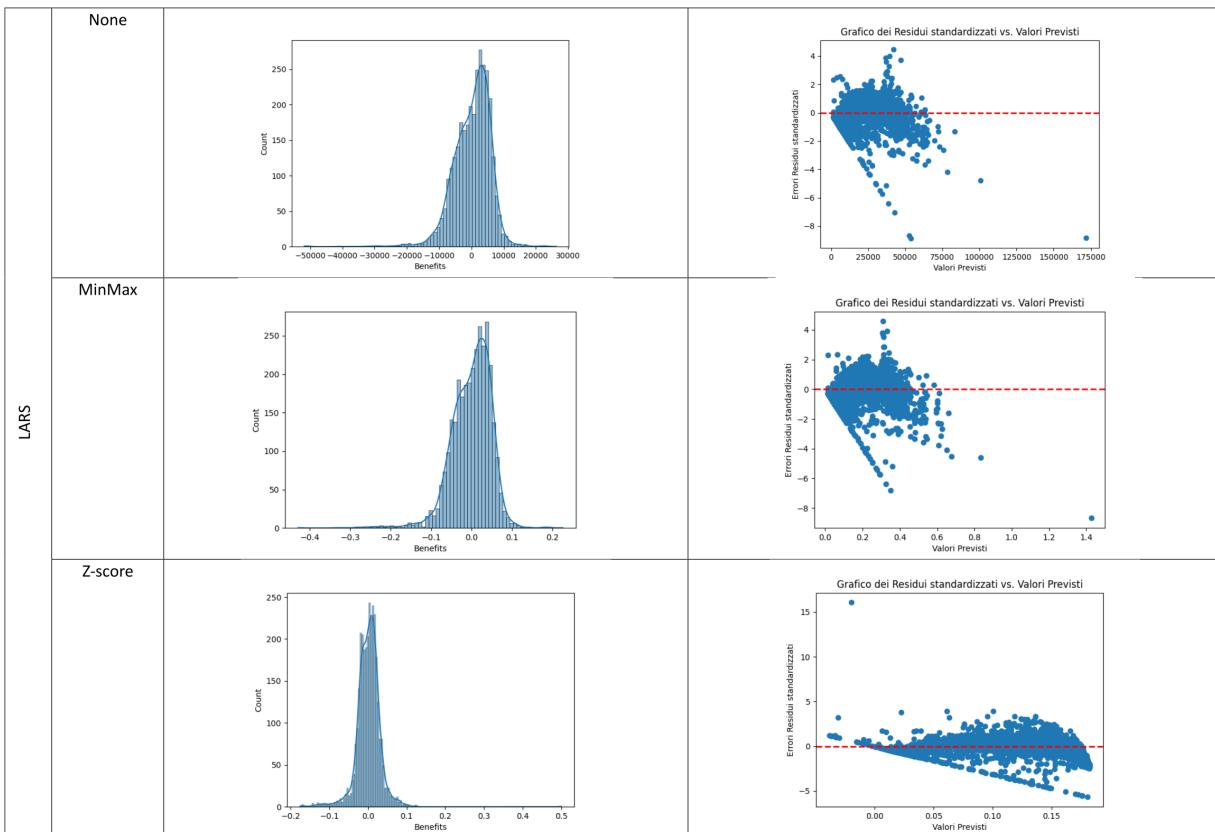
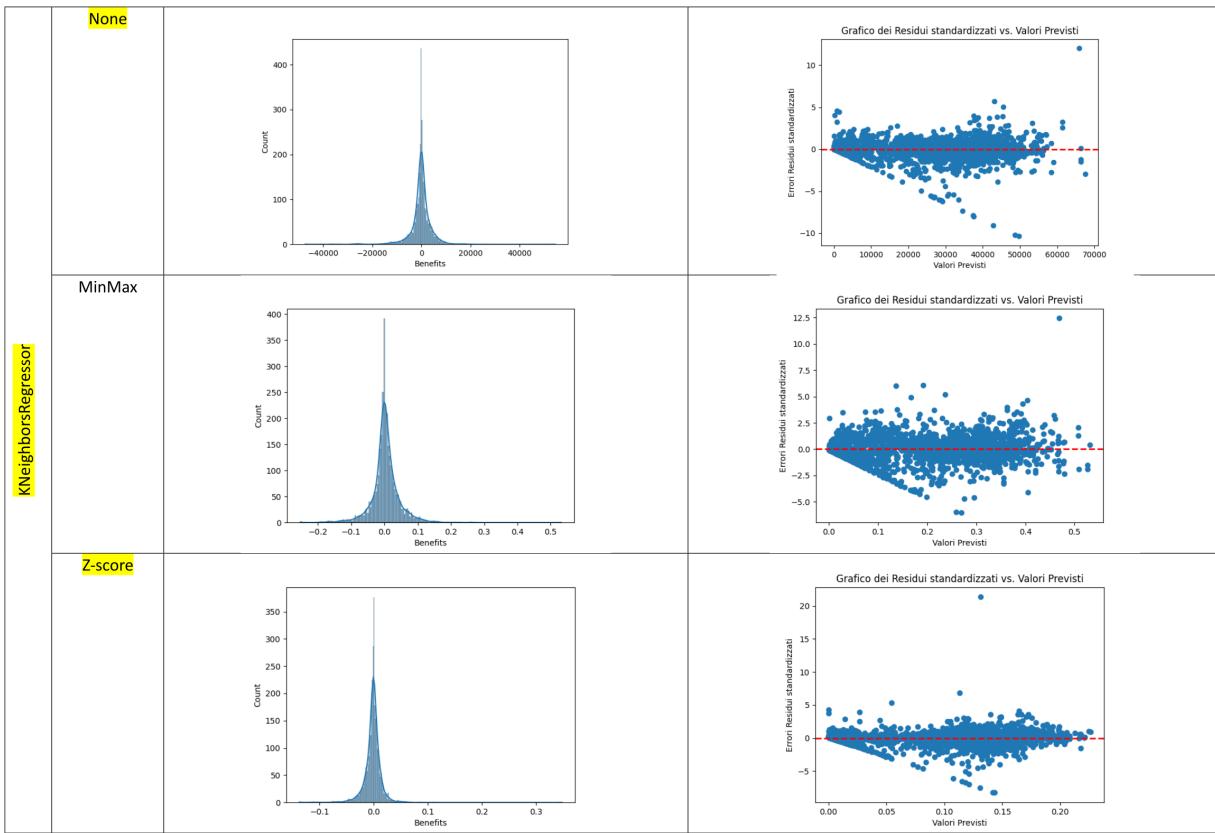
## 9 Conclusioni e Pianificazione Futura

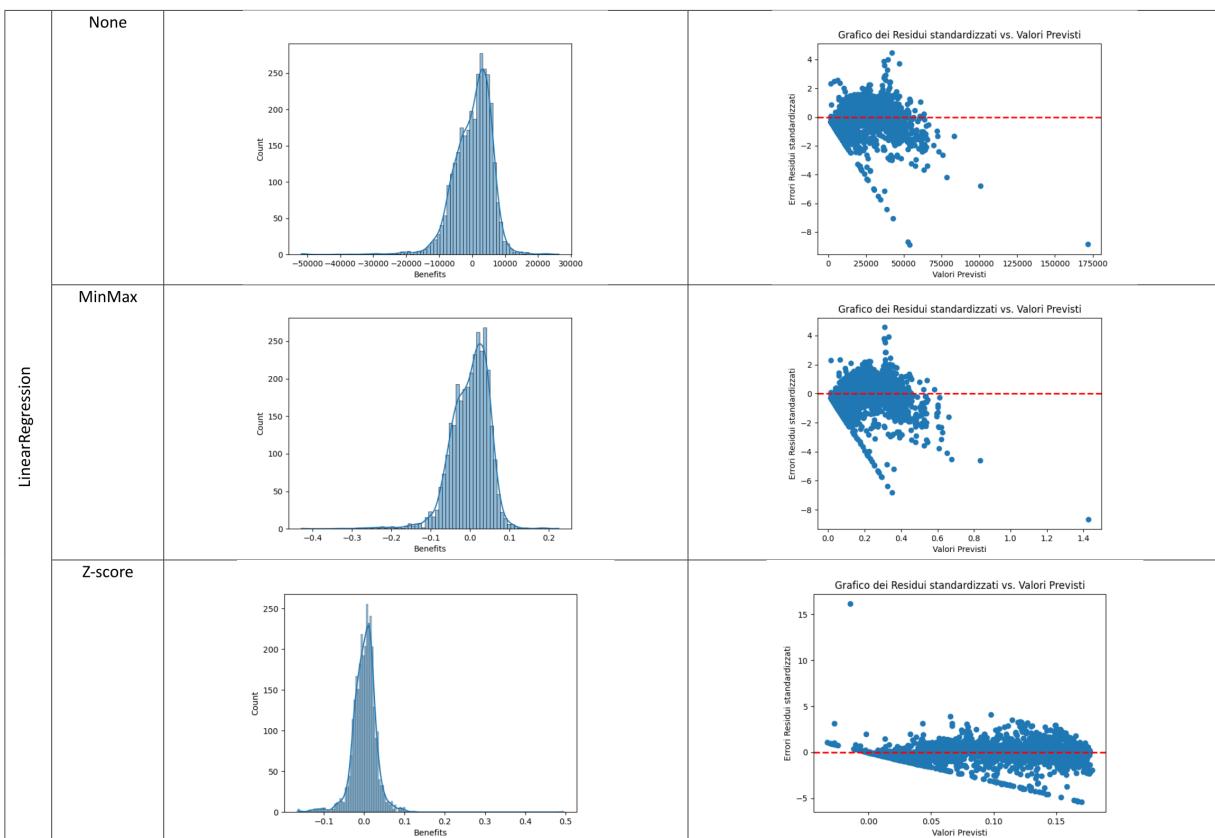
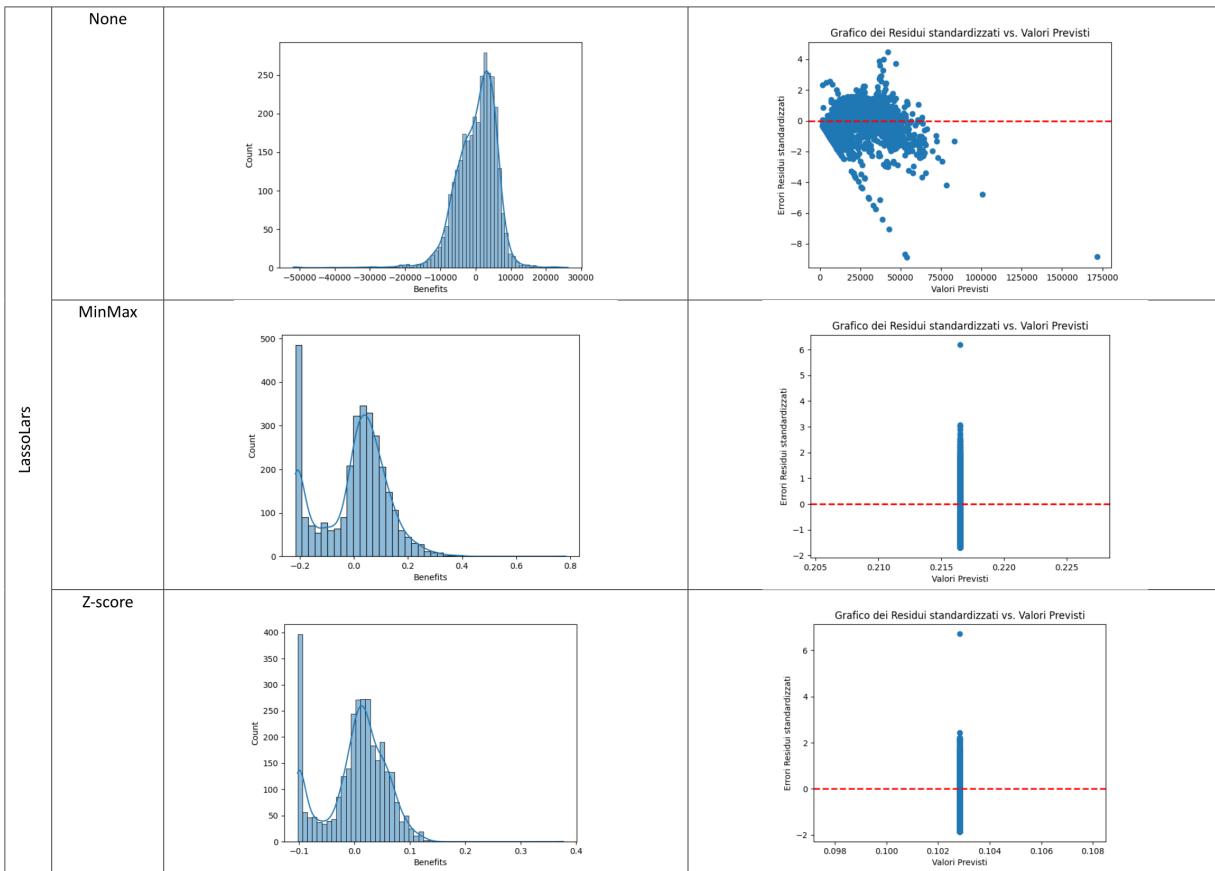
### 9.1 Risultati

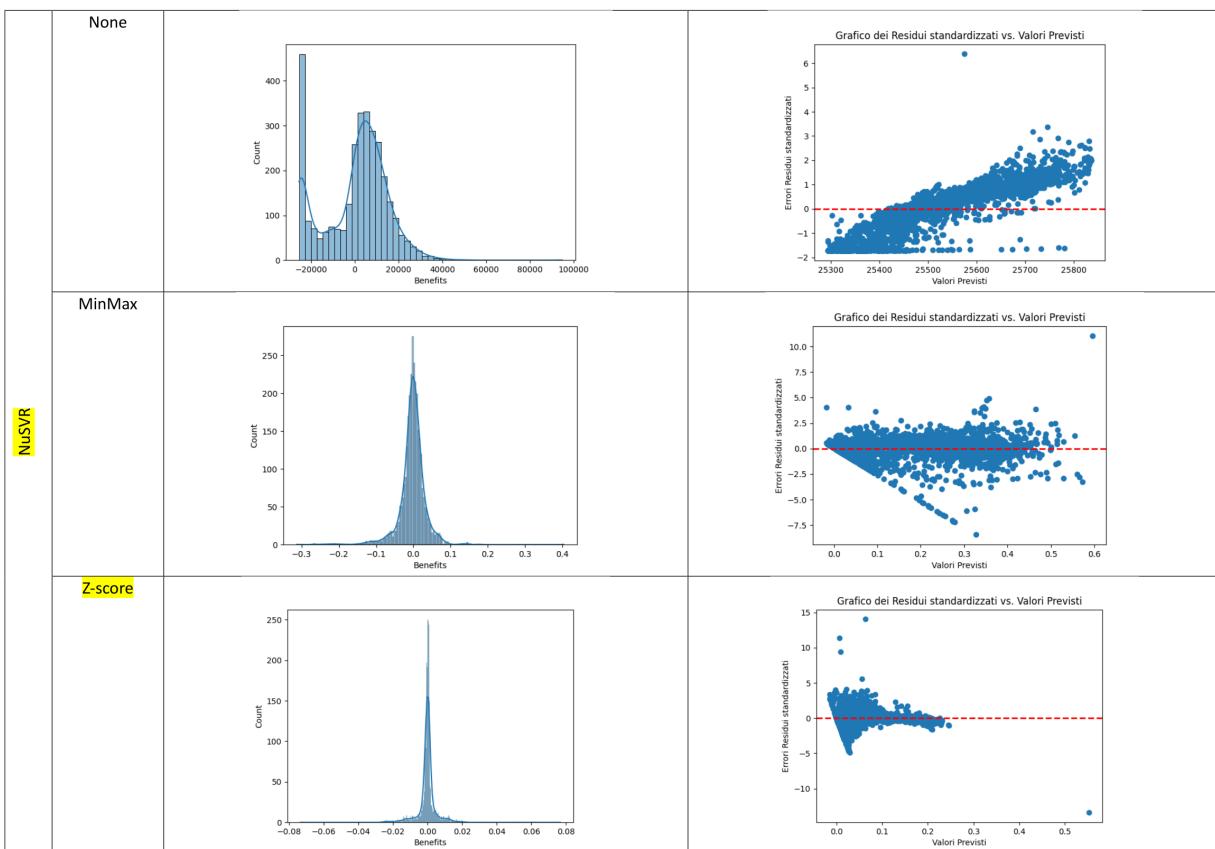
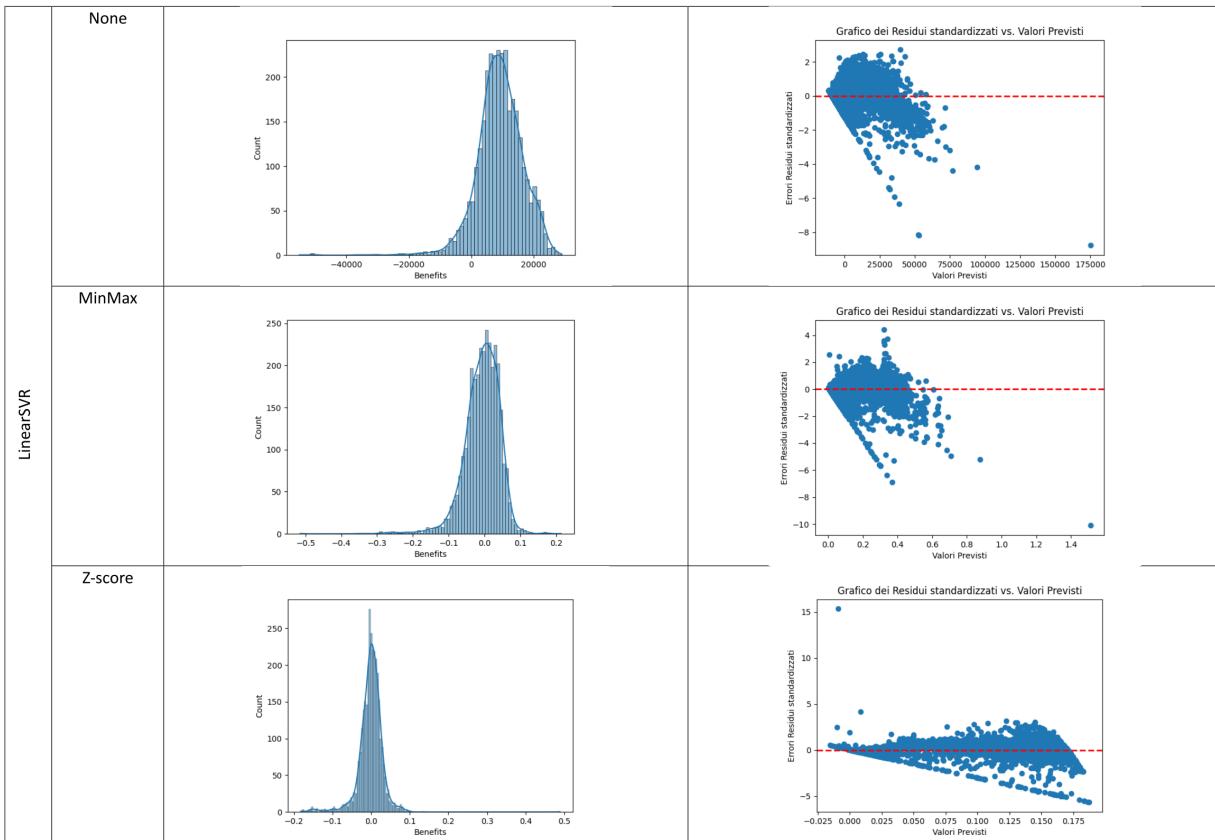
I questa sezione vedremo i dati ottenuti dall'esecuzione del file `main.py`. Per una facile visualizzazione i risultati sono stati organizzati in formato tabellare e sono stati evidenziati quelli che hanno superato i criteri di successo.

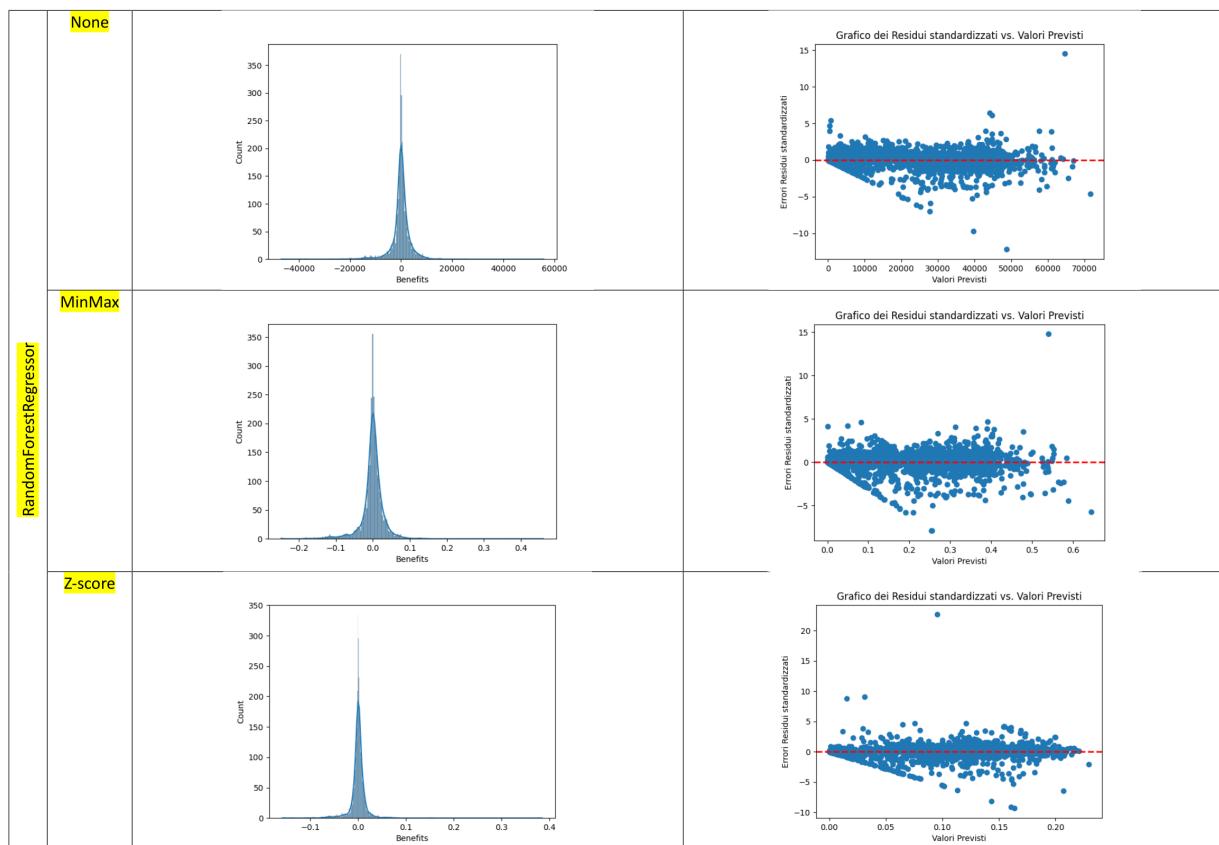
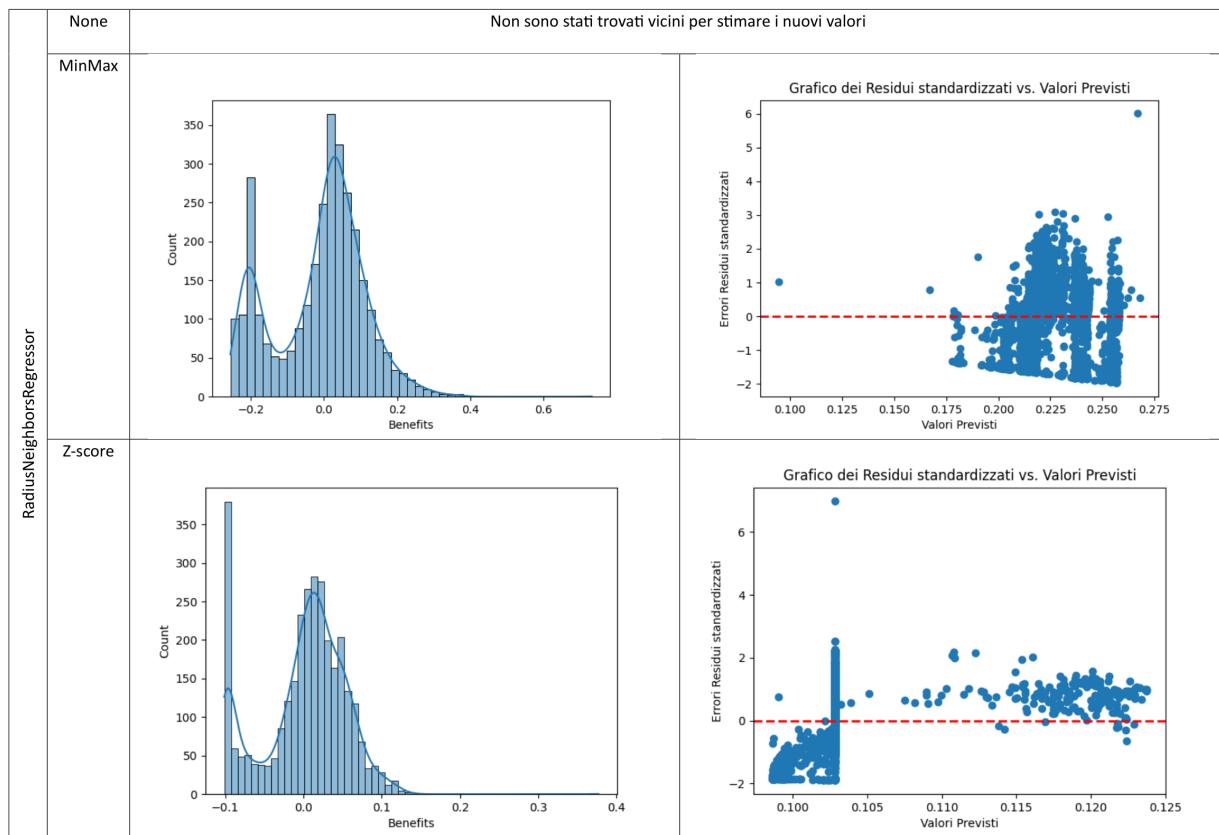


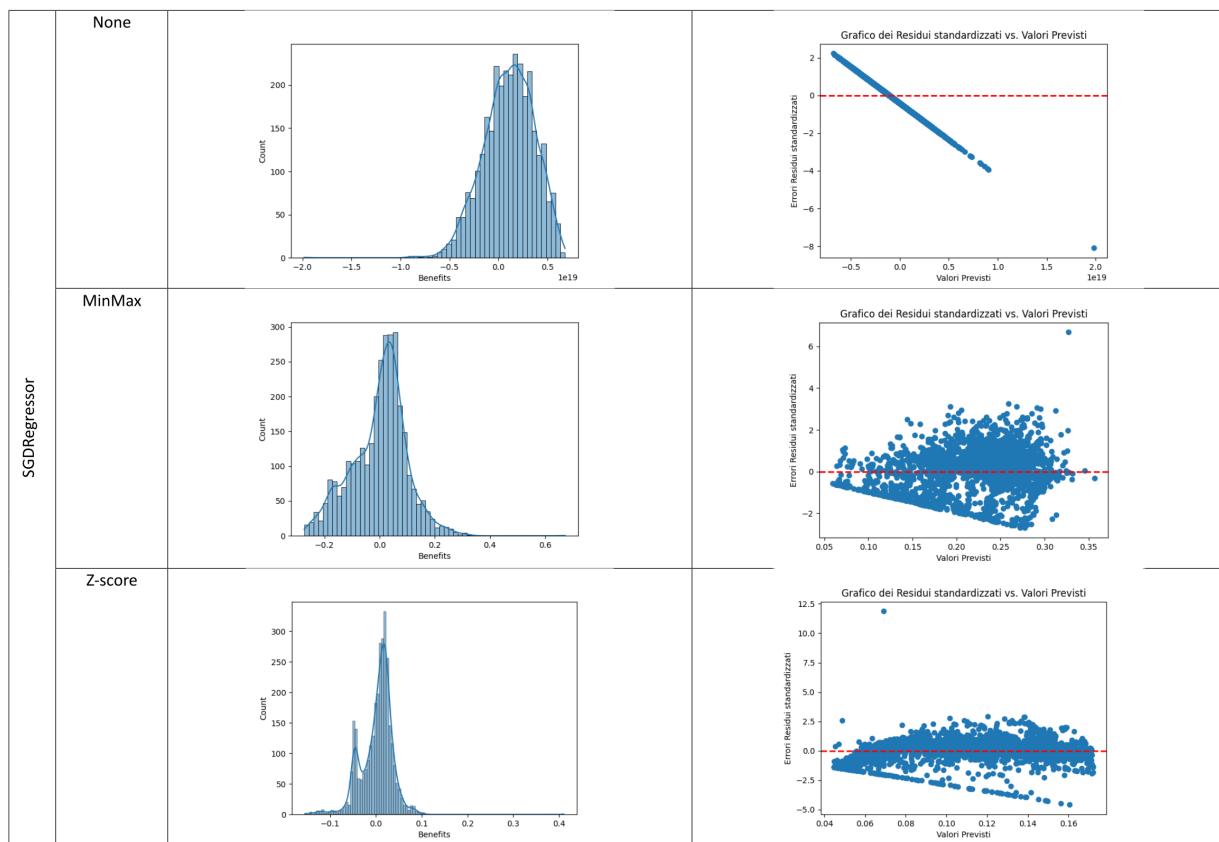
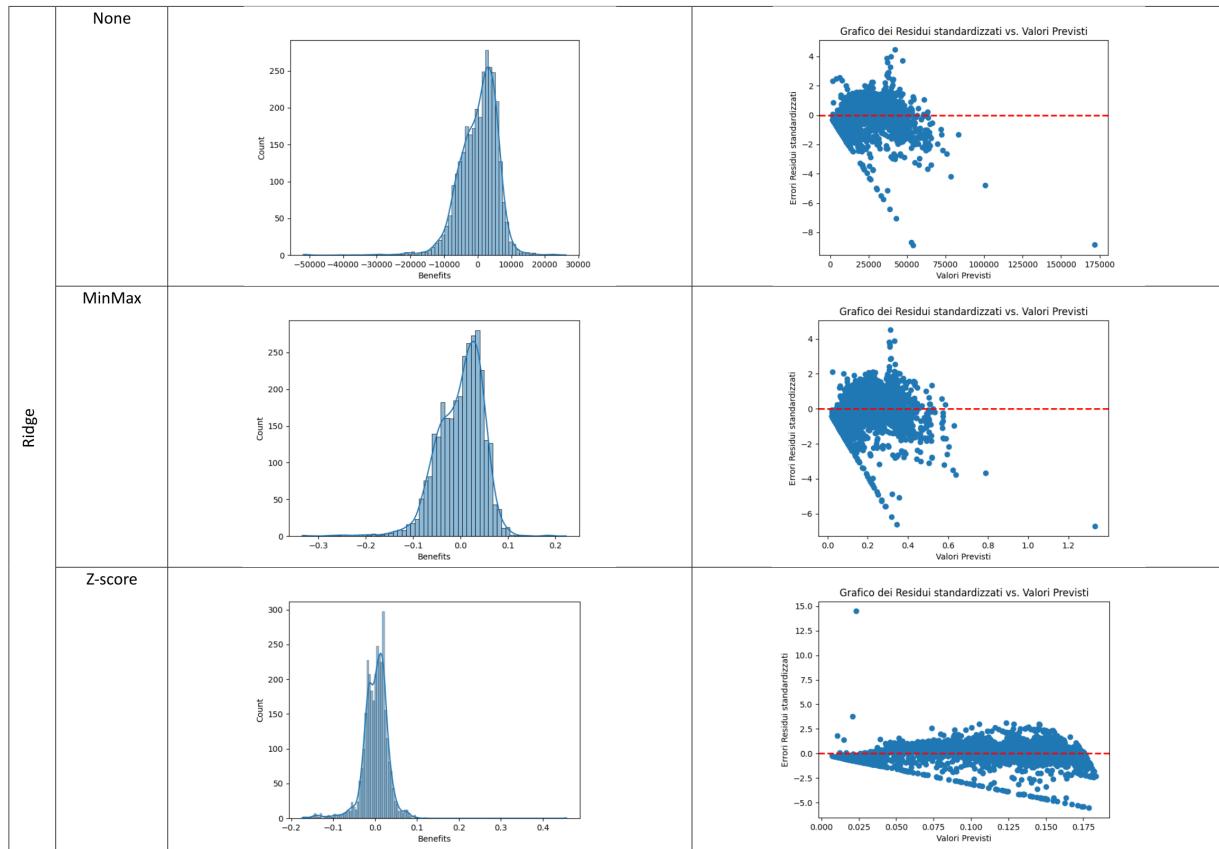


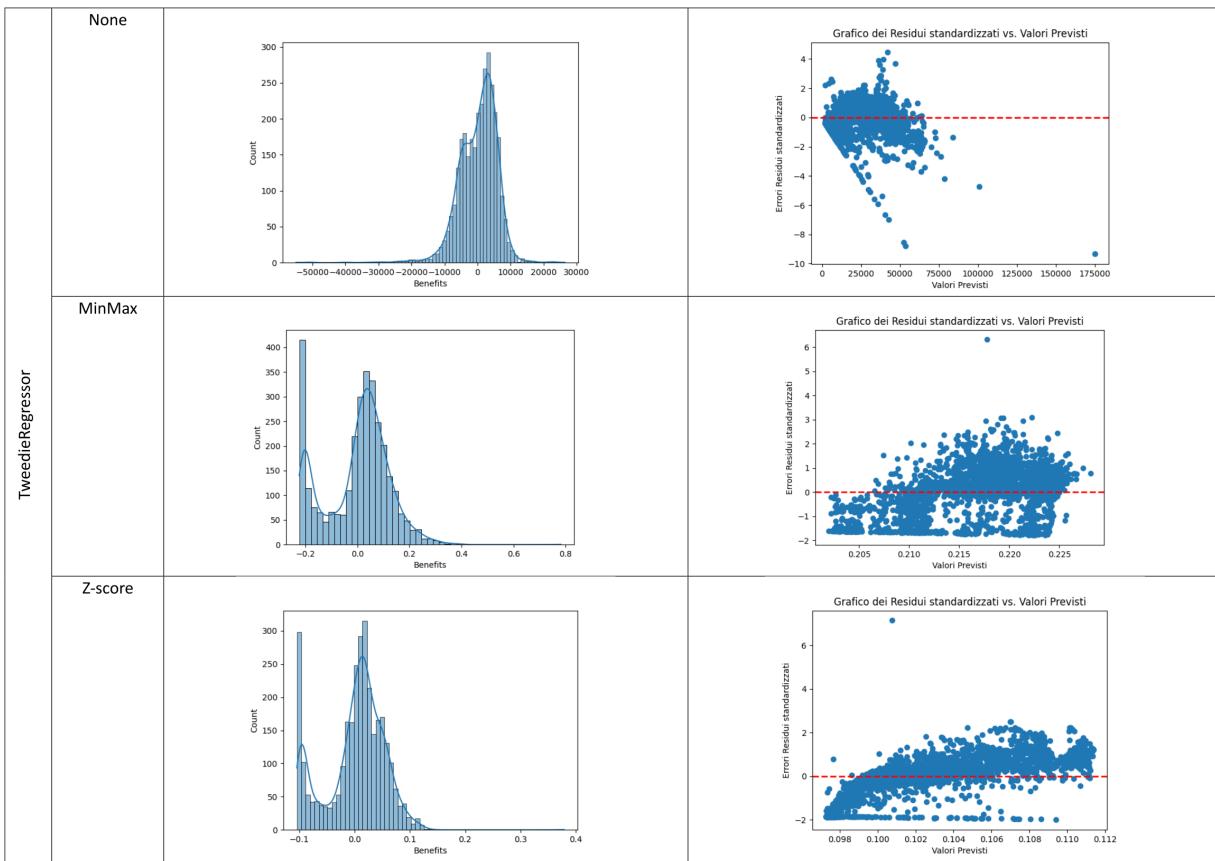
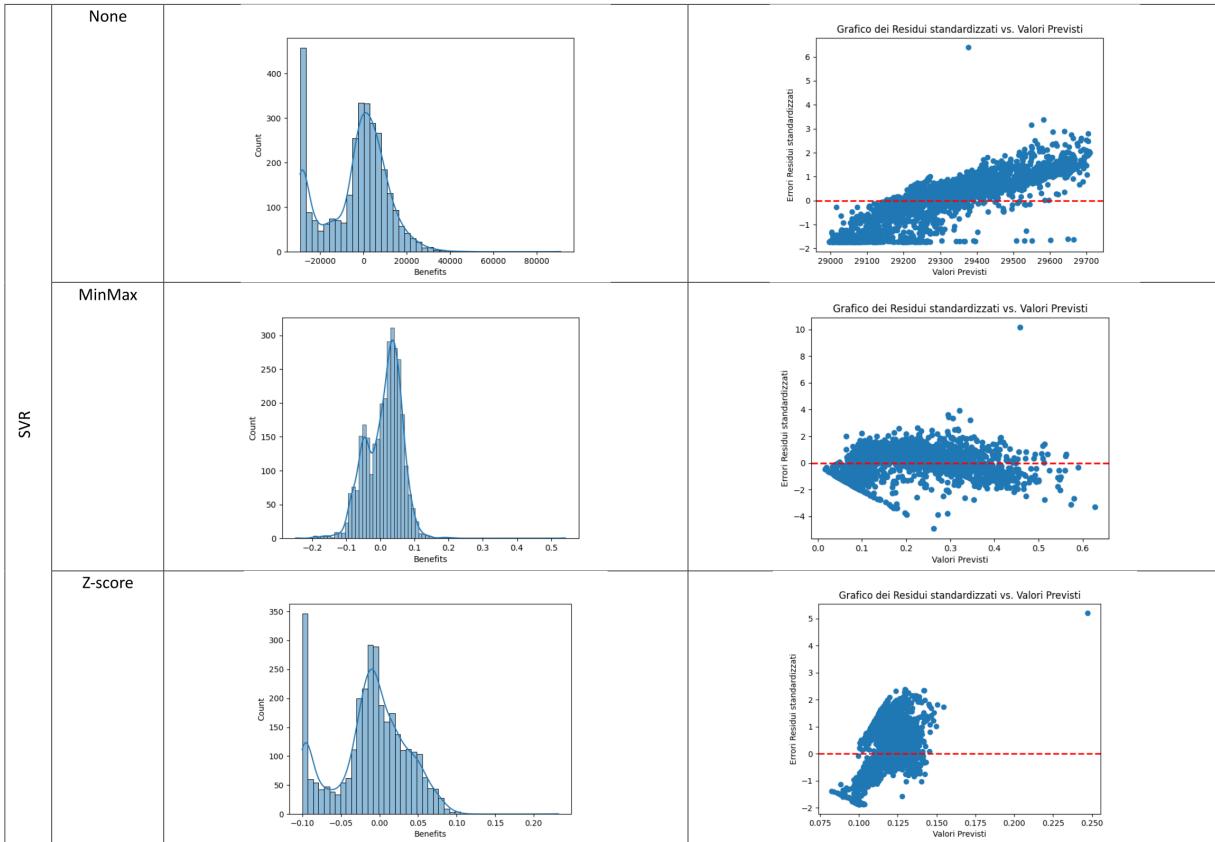












Come possiamo vedere, il tipo di normalizzazione influisce molto sia sulla distribuzione degli errori che sulla varianza, quindi può portare sia migliorie che peggioramenti al modello. Inoltre, dai grafici notiamo la presenza degli outlier, che in fase di analisi non erano stati rilevati, vista la grande mole del dataset. Possiamo anche notare come la normalizzazione MinMax non influenza la forma della distribuzione degli errori nella maggior parte dei casi. Va notato che non è stato possibile utilizzare l'algoritmo Radius Neighbors senza applicare alcun tipo di normalizzazione dei dati. Inoltre, possiamo notare come la normalizzazione tende a diminuire, se non a cancellare drasticamente, le metriche: MAE, MSE, RMSE. Quindi, per valutare complessivamente un algoritmo, possiamo prendere l' $R^2$  come metrica unica. Tale tabella è presente alla fine del documento, dove sono riportate le metriche per ogni algoritmo.

## 9.2 Successi e Sfide

Indubbiamente, la parte difficile è stata progettare un sistema che si adatti a tutti i possibili regressori presenti, e organizzare i risultati di tali comparazioni affinché siano fruibili e comprensibili a più persone possibili. Nel caso d'uso portato, vediamo come diversi regressori abbiano superato quelli che erano i criteri di successo, come: GaussianProcessRegressor con normalizzazione z-score, RandomForestRegressor, KNeighborsRegressor senza normalizzazione e con normalizzazione z\_score e NuSVR con normalizzazione MinMax e z\_score.

## 9.3 Sviluppi Futuri e Miglioramenti Possibili

Premettendo che è un'ottima base per analisi di regressori, sarebbe sicuramente possibile comprendere tra le metriche anche il "Fit time mean" e il "Score time mean", che nel report sono stati indicati ma non riportati nelle tabelle. Sarebbe ottimale che il programma fornisse un algoritmo migliore, senza lasciare interpretare i dati all'utente. Idealmente potremmo restituire il migliore per ogni metrica. Inoltre, potremmo effettuare il grafico della varianza dell'errore residuo, valutando diverse normalizzazioni sui dati. Dai grafici notiamo alcuni outlier che potremmo rimuovere nella fase di pulizia dei dati. Inoltre potremmo valutare l'uso di librerie per rendere interattivi i grafici o più facilmente interpretabili e quindi migliorare la data-visualization.

		valuation	k-cross valuation
LinearRegression	non	MAE:4414.02 MSE:34240294.66 RMSE:5851.52 R2:0.85	MAE:-4584.90 MSE:-36363118.70 RMSE:-6012.39 R2:0.84
	minmax	MAE:0.04 MSE:0.00 RMSE:0.05 R2:0.85	MAE:-0.03 MSE:-0.00 RMSE:-0.04 R2:0.83
	z_score	MAE:0.02 MSE:0.00 RMSE:0.03 R2:0.70	MAE:-0.02 MSE:-0.00 RMSE:-0.029 R2:0.71
LARS	non	MAE:4414.02 MSE:34240294.66 RMSE:5851.52 R2:0.85	MAE:-4585.85 MSE:-36388949.97 RMSE:-6019.91 R2:0.84
	minmax	MAE:0.04 MSE:0.00 RMSE:0.05 R2:0.85	MAE:-0.03 MSE:-0.00 RMSE:-0.04 R2:0.83
	z_score	MAE:0.02 MSE:0.00 RMSE:0.03 R2:0.69	MAE:-0.02 MSE:-0.00 RMSE:-0.03 R2:0.70
Rige	non	MAE:4414.06 MSE:34240082.66 RMSE:5851.50 R2:0.85	MAE:-4588.17 MSE:-36421552.08 RMSE:-6013.55 R2:0.84
	minmax	MAE:0.04 MSE:0.00 RMSE:0.05 R2:0.85	MAE:-0.03 MSE:-0.00 RMSE:-0.04 R2:0.83
	z_score	MAE:0.02 MSE:0.00 RMSE:0.03 R2:0.68	MAE:-0.02 MSE:-0.00 RMSE:-0.03 R2:0.68

		valuation	k-cross valutation
LassoLars	non	MAE: 4414.19 MSE:34239402.57 RMSE:5851.44 R2:0.85	MAE:-4585.45 MSE:-36386719.94 RMSE:-6010.52 R2:0.84
	minmax	MAE:0.10 MSE:0.02 RMSE:0.13 R2:0.00	MAE:-0.09 MSE:-0.01 RMSE:-0.12 R2:0.00
	z_score	MAE:0.04 MSE:0.00 RMSE:0.06 R2:0.00	MAE:-0.04 MSE:-0.00 RMSE:-0.05 R2:-0.00
ARDRegression	non	MAE:4433.12 MSE:34265342.43 RMSE:5853.66 R2:0.85	MAE:-4600.41 MSE: -36423938.17 RMSE:-6020.70 R2:0.84
	minmax	MAE:0.04 MSE:0.00 RMSE:0.05 R2:0.85	MAE:-0.03 MSE:-0.00 RMSE:-0.04 R2:0.84
	z_score	MAE:0.02 MSE:0.00 RMSE:0.03 R2:0.70	MAE:-0.02 MSE:-0.00 RMSE:-0.02 R2:0.71
SGDRegressor	non	MAE:2276259820429790720 MSE:7806400250711500091555503270269026304 RMSE:2793993602482206720 R2:-34985771752266058653400825856	MAE:-7.56 MSE:-1.15 RMSE:-8.86 R2:-4.89
	minmax	MAE:0.08 MSE:0.01 RMSE:0.10 R2:0.37	MAE:-0.07 MSE:-0.00 RMSE:-0.09 R2:0.37
	z_score	MAE:0.03 MSE:0.00 RMSE:0.03 R2:0.62	MAE:-0.02 MSE:-0.00 RMSE:-0.03 R2:0.61

		valuation	k-cross valutation
BayesianRidge	non	MAE:4501.46 MSE:35469603.31 RMSE:5955.64 R2:0.84	MAE:-4653.39 MSE:-37494271.93 RMSE:-6106.87 R2:0.83
	minmax	MAE:0.04 MSE:0.00 RMSE:0.05 R2:0.85	MAE:-0.03 MSE:-0.00 RMSE:-0.04 R2:0.83
	z_score	MAE:0.02 MSE:0.00 RMSE:0.03 R2:0.69	MAE:-0.02 MSE:-0.00 RMSE:-0.03 R2:0.69
GaussianProcessRegressor	non	MAE:25742.20 MSE:885890338.39 RMSE:29763.91 R2:-2.97	MAE:-26203.35 MSE:-918970542.13 RMSE:-30305.83 R2:-2.98
	minmax	MAE:0.08 MSE:1.34 RMSE:1.16 R2:-82.24	MAE:-0.08 MSE:-1.14 RMSE:-0.65 R2:-73.65
	z_score	MAE:0.00 MSE:0.00 RMSE:0.01 R2:0.99	MAE:-0.00 MSE:-4.62 RMSE:-0.00 R2:0.98
TweedieRegressor	non	MAE:4456.86 MSE:34855324.49 RMSE:5903.84 R2:0.84	MAE:-4617.30 MSE:-36976462.38 RMSE:-6064.61 R2:0.83
	minmax	MAE:0.10 MSE:0.02 RMSE:0.12 R2:0.05	MAE:-0.09 MSE:-0.01 RMSE:-0.12 R2:0.04
	z_score	MAE:0.04 MSE:0.00 RMSE:0.05 R2:0.11	MAE:-0.04 MSE:-0.00 RMSE:-0.05 R2:0.10

		valuation	k-cross valuation
DecisionTreeRegressor	non	MAE:2767.09 MSE:24070002.09 RMSE:4906.12 R2:0.89	MAE:-2821.83 MSE:-25113669.21 RMSE:-4990.13 R2:0.89
		MAE:0.02 MSE:0.00 RMSE:0.04 R2:0.89	MAE:-0.02 MSE:-0.00 RMSE:-0.04 R2:0.87
		MAE:0.01 MSE:0.00 RMSE:0.02 R2:0.84	MAE:-0.01 MSE:-0.00 RMSE:-0.02 R2:0.82
	minmax	MAE:2107.78 MSE:14793332.88 RMSE:3846.21 R2:0.93	MAE:-2179.3 MSE:-14564856.99 RMSE:-3793.43 R2:0.93
		MAE:0.02 MSE:0.00 RMSE:0.03 R2:0.94	MAE:-0.01 MSE:-0.00 RMSE:-0.03 R2:0.93
		MAE:0.01 MSE:0.00 RMSE:0.02 R2:0.91	MAE:-0.00 MSE:-0.00 RMSE:-0.01 R2:0.90
	z_score	MAE:2481.89 MSE:20743583.79 RMSE:4554.51 R2:0.91	MAE:-2600.17 MSE:-21903678.75 RMSE:-4654.92 R2:0.90
		MAE:0.03 MSE:0.00 RMSE:0.04 R2:0.89	MAE:-0.02 MSE:-0.00 RMSE:-0.04 R2:0.86
		MAE:0.01 MSE:0.00 RMSE:0.02 R2:0.91	MAE:-0.01 MSE:-0.00 RMSE:-0.01 R2:0.89
RandomForestRegressor	non	MAE:2107.78 MSE:14793332.88 RMSE:3846.21 R2:0.93	MAE:-2179.3 MSE:-14564856.99 RMSE:-3793.43 R2:0.93
		MAE:0.02 MSE:0.00 RMSE:0.03 R2:0.94	MAE:-0.01 MSE:-0.00 RMSE:-0.03 R2:0.93
		MAE:0.01 MSE:0.00 RMSE:0.02 R2:0.91	MAE:-0.00 MSE:-0.00 RMSE:-0.01 R2:0.90
	minmax	MAE:2481.89 MSE:20743583.79 RMSE:4554.51 R2:0.91	MAE:-2600.17 MSE:-21903678.75 RMSE:-4654.92 R2:0.90
		MAE:0.03 MSE:0.00 RMSE:0.04 R2:0.89	MAE:-0.02 MSE:-0.00 RMSE:-0.04 R2:0.86
		MAE:0.01 MSE:0.00 RMSE:0.02 R2:0.91	MAE:-0.01 MSE:-0.00 RMSE:-0.01 R2:0.89
	z_score	MAE:2481.89 MSE:20743583.79 RMSE:4554.51 R2:0.91	MAE:-2600.17 MSE:-21903678.75 RMSE:-4654.92 R2:0.90
		MAE:0.03 MSE:0.00 RMSE:0.04 R2:0.89	MAE:-0.02 MSE:-0.00 RMSE:-0.04 R2:0.86
		MAE:0.01 MSE:0.00 RMSE:0.02 R2:0.91	MAE:-0.01 MSE:-0.00 RMSE:-0.01 R2:0.89
KNeighborsRegressor	non	MAE:2481.89 MSE:20743583.79 RMSE:4554.51 R2:0.91	MAE:-2600.17 MSE:-21903678.75 RMSE:-4654.92 R2:0.90
		MAE:0.03 MSE:0.00 RMSE:0.04 R2:0.89	MAE:-0.02 MSE:-0.00 RMSE:-0.04 R2:0.86
		MAE:0.01 MSE:0.00 RMSE:0.02 R2:0.91	MAE:-0.01 MSE:-0.00 RMSE:-0.01 R2:0.89
	minmax	MAE:2481.89 MSE:20743583.79 RMSE:4554.51 R2:0.91	MAE:-2600.17 MSE:-21903678.75 RMSE:-4654.92 R2:0.90
		MAE:0.03 MSE:0.00 RMSE:0.04 R2:0.89	MAE:-0.02 MSE:-0.00 RMSE:-0.04 R2:0.86
		MAE:0.01 MSE:0.00 RMSE:0.02 R2:0.91	MAE:-0.01 MSE:-0.00 RMSE:-0.01 R2:0.89
	z_score	MAE:2481.89 MSE:20743583.79 RMSE:4554.51 R2:0.91	MAE:-2600.17 MSE:-21903678.75 RMSE:-4654.92 R2:0.90
		MAE:0.03 MSE:0.00 RMSE:0.04 R2:0.89	MAE:-0.02 MSE:-0.00 RMSE:-0.04 R2:0.86
		MAE:0.01 MSE:0.00 RMSE:0.02 R2:0.91	MAE:-0.01 MSE:-0.00 RMSE:-0.01 R2:0.89

		valuation	k-cross valutation
		None	None
RadiusNeighborsRegressor	non	MAE:0.10 MSE:0.02 RMSE:0.12 R2:0.04	MAE:-0.09 MSE:-0.01 RMSE:-0.12 R2:0.05
	minmax	MAE:0.04 MSE:0.00 RMSE:0.05 R2:0.07	MAE:-0.04 MSE:-0.00 RMSE:-0.05 R2:0.06
	z_score	MAE:11467.26 MSE:231431412.85 RMSE:15212.87 R2:-0.04	MAE:-11703.46 MSE:-238188649.62 RMSE:-15422.92 R2:-0.02
	non	MAE:0.04 MSE:0.00 RMSE:0.05 R2:0.82	MAE:-0.04 MSE:-0.00 RMSE:-0.05 R2:0.81
	minmax	MAE:0.04 MSE:0.00 RMSE:0.05 R2:0.25	MAE:-0.03 MSE:-0.00 RMSE:-0.04 R2:0.23
	z_score	MAE:10021.96 MSE:136470337.52 RMSE:11682.05 R2:0.39	MAE:-7500.49 MSE:-107987723.58 RMSE:-9635.48 R2:0.53
SVR	non	MAE:0.04 MSE:0.00 RMSE:0.05 R2:0.84	MAE:-0.03 MSE:-0.00 RMSE:-0.05 R2:0.83
	minmax	MAE:0.02 MSE:0.00 RMSE:0.03 R2:0.67	MAE:-0.02 MSE:-0.00 RMSE:-0.03 R2:0.68
	z_score		
LinearSVR	non		
	minmax		
	z_score		

		valuation	k-cross valuation
NuSVR	non	MAE:11912.50 MSE:219871264.12 RMSE:14828.06 R2:0.01	MAE:-12296.28 MSE:-229898093.02 RMSE:-15153.89 R2:0.00
	minmax	MAE:0.02 MSE:0.00 RMSE:0.04 R2:0.92	MAE:-0.02 MSE: -0.00 RMSE:-0.03 R2:0.89
	z_score	MAE:0.00 MSE:0.00 RMSE:0.01 R2:0.99	MAE:-0.00 MSE:-3.70 RMSE:-0.00 R2:0.98