PREDIZIONE DEL LIVELLO E DELLA PORTATA D'ACQUA IN UN BACINO IDRICO

Marco Braga, 829295
Andrea Maver, 828725

Progetto Industry Lab

Luglio 2022

Indice

Introduzione

Obiettivi

Lo scopo del progetto è stimare il Livello dell'Acqua e la Portata di Uscita di un bacino idrico artificiale.
Poter prevedere il Livello del Bacino e la Portata con una settimana di anticipo permette di non sprecare risorse e ciò porta valore e benefici all'azienda gestore del bacino.

Sono stati testati diversi metodi e algoritmi per poter trovare il più adatto a raggiungere gli scopi prefissati.

Water 4.0

Per poter relizzare i propri scopi, l'azienda ha installato un sistema di sensoristica, condizione essenziale per l'Industria 4.0, o meglio, Water 4.0

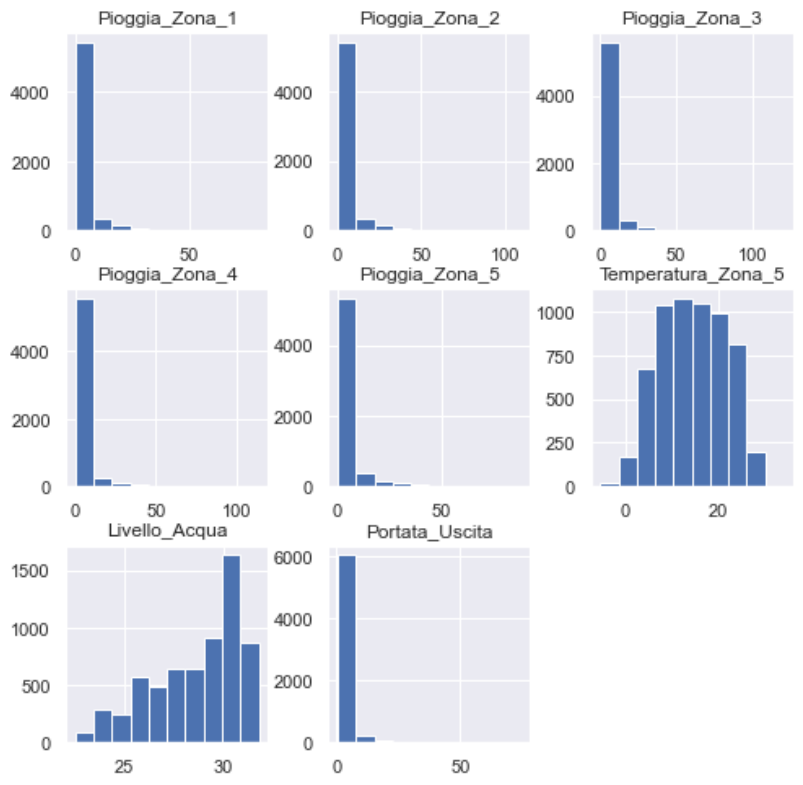
Dataset

Il dataset è composto da 6386 osservazioni e 9 colonne:

1. Data
2. Pioggia Zona i (i=1,..,5)
3. Temperatura Zona 5
4. Livello Acqua (target)
5. Portata Uscita (target)

Esplorazione Dataset

Esplorazione



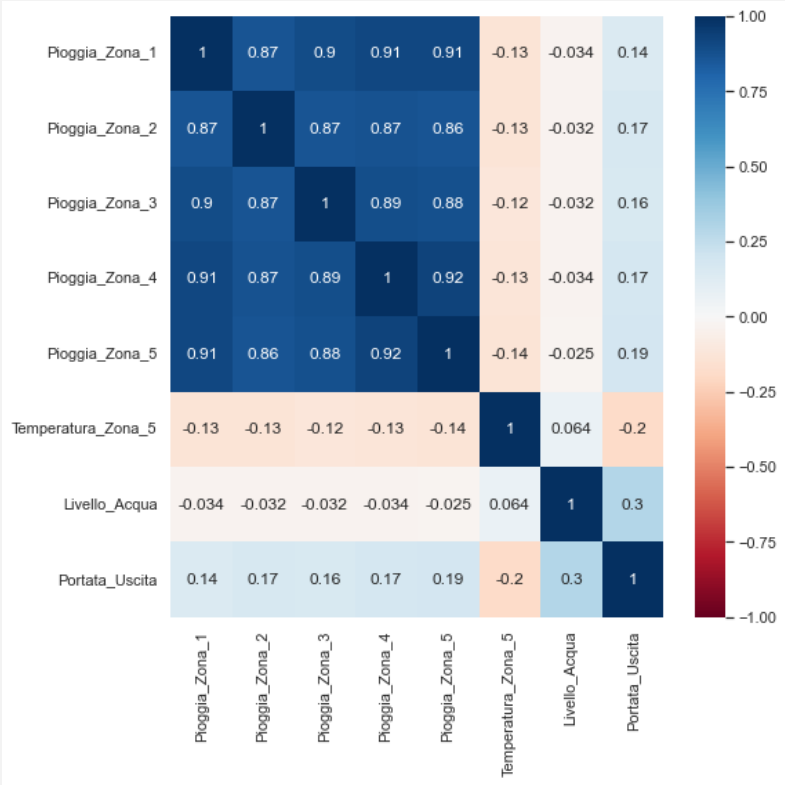


Figure: Distribuzioni delle variabili e correlazione di Pearson

Livello Acqua

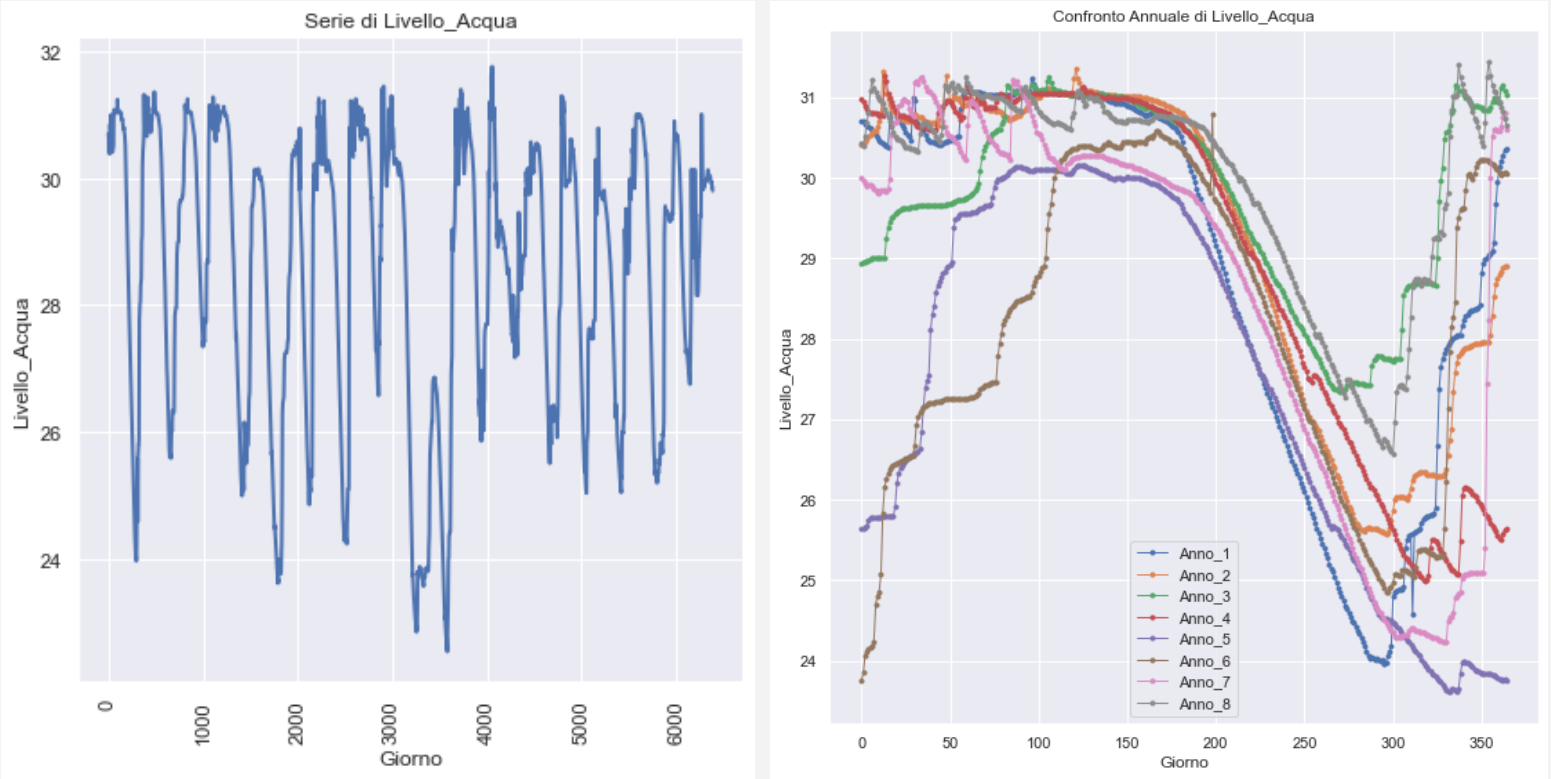
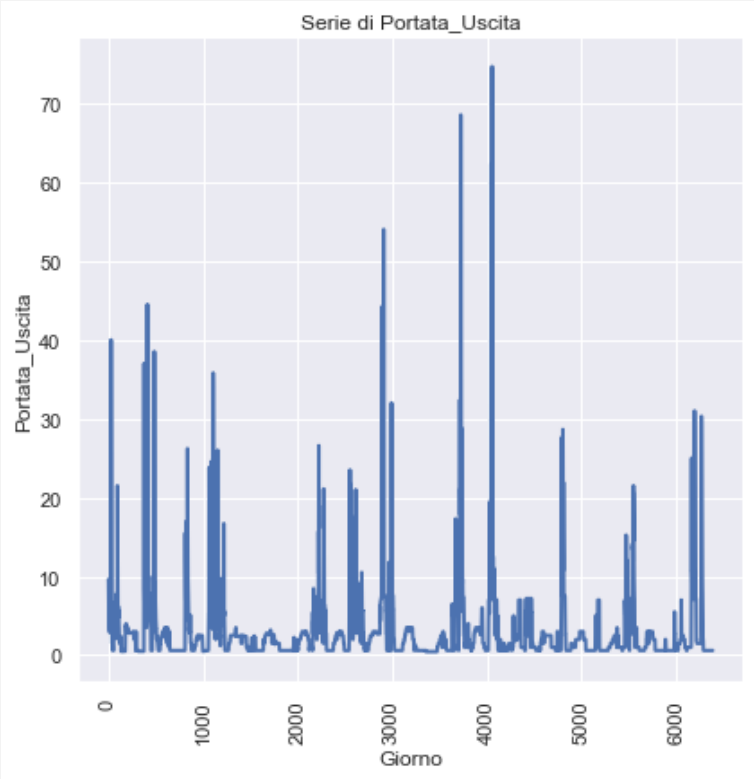


Figure: Livello Acqua: serie completa e divisa anno per anno

Portata Uscita



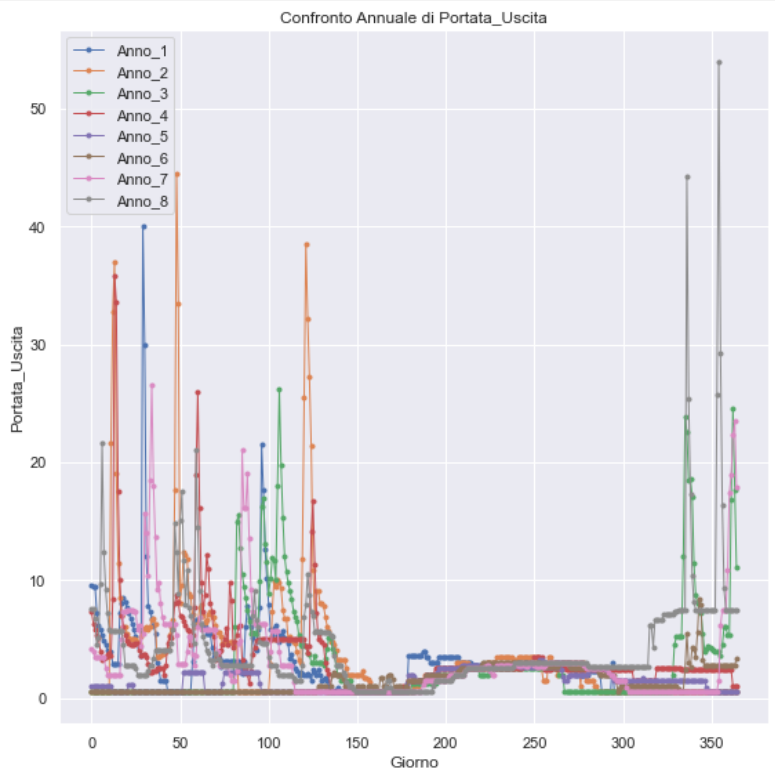
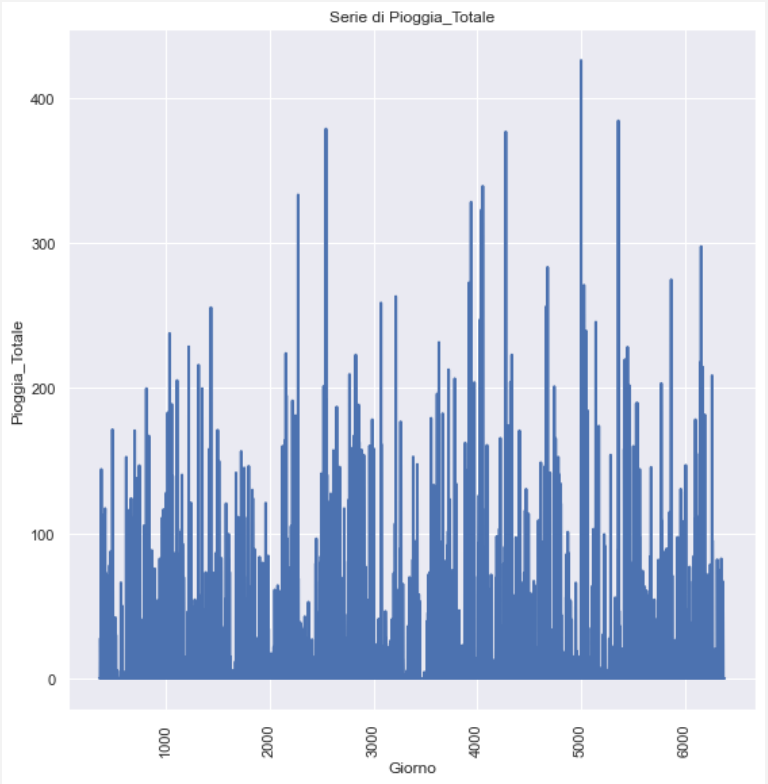


Figure: Portata Uscita: serie completa e divisa anno per anno

Attributo Pioggia



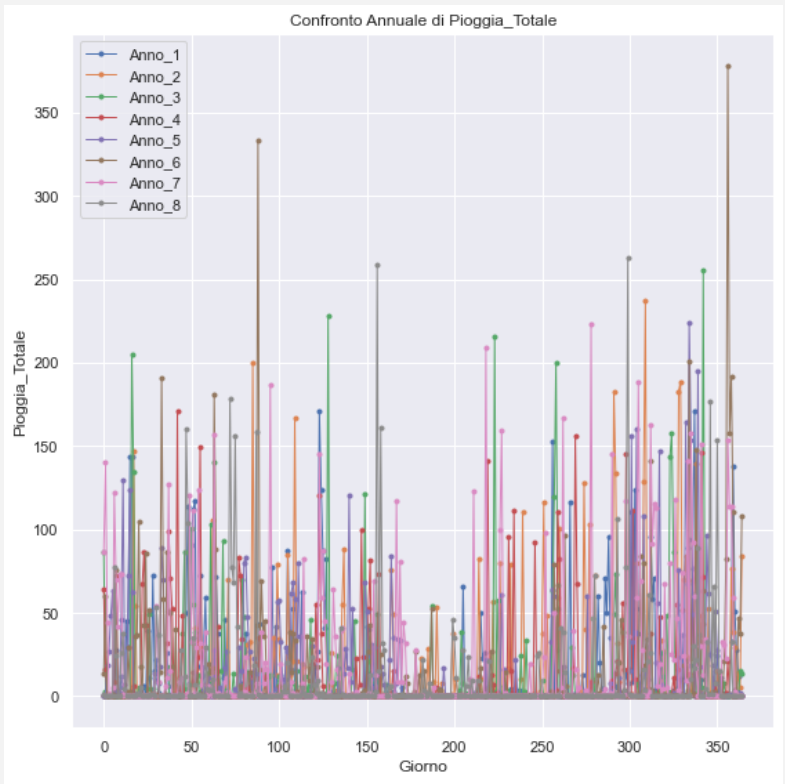
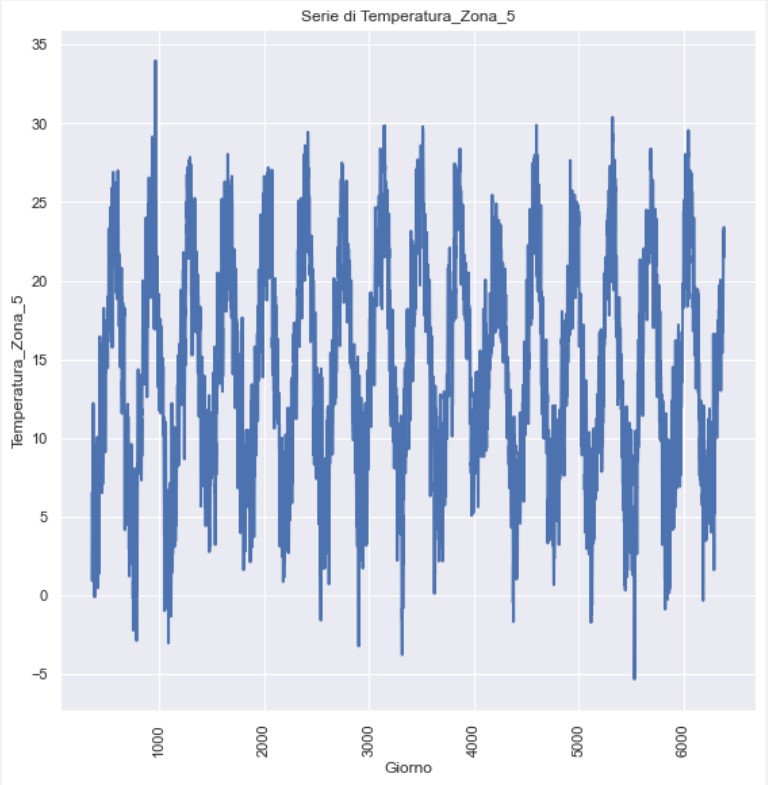


Figure: Pioggia Totale: serie completa e divisa anno per anno

Attributo Temperatura



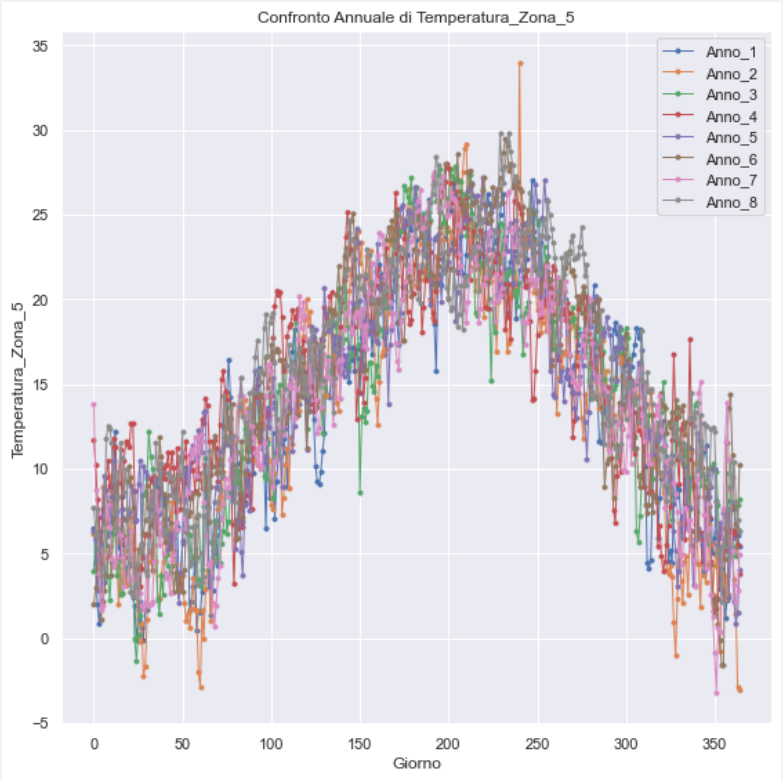


Figure: Temperatura: serie completa e divisa anno per anno

Aspetti metodologici

Modelli testati

Per lo sviluppo del progetto sono stati testati diversi modelli:

- Modelli statistici
  1. Regressione lineare
  2. ARIMAX
  3. Arch
  4. VAR
- Modelli machine learning
  1. Symbolic regression
  2. K-nearest neighbour
- Reti neurali

Pre-processing

Pre-processing 1/2

Le principali operazioni di manipolazione del dataset originale

- Eliminazione 360 valori nulli
- Creazione di 4 sinusoidi, utili per il modello ARIMA
- Divisione in 80% train e 20% test senza shuffle
- Lag avanti di sette giorni per le colonne target
- Standardizzazione degli attributi

Dopo questi passaggi vengono creati diversi dataset, per entrambi i target, sia con attributi normalizzati che non, ognuno utile per i vari modelli testati

Pre-processing 2/2

Vengono aggiunti degli attributi personalizzati per i due target.

Livello:

- Tempo e tempo al quadrato come trend deterministico quadratico
- Intercetta con tutti valori 1, necessaria per i modelli OLS

Portata:

- Tempo come seno del logaritmo del tempo standard, per simulare una serie con picchi
- Intercetta
- Livello al quadrato

Analisi dei Modelli

Regressione Lineare

La Regressione lineare viene considerata come modello baseline, di confronto. È il modello più semplice in quanto non considera dipendenze temporali.

	Time (ms)	MSE
Portata	206	17.292
Livello	270	5.172

Table: Risultati Regressione Lineare

Sarimax 1: Livello

Il modello scelto è $(3, 1, 1)$.

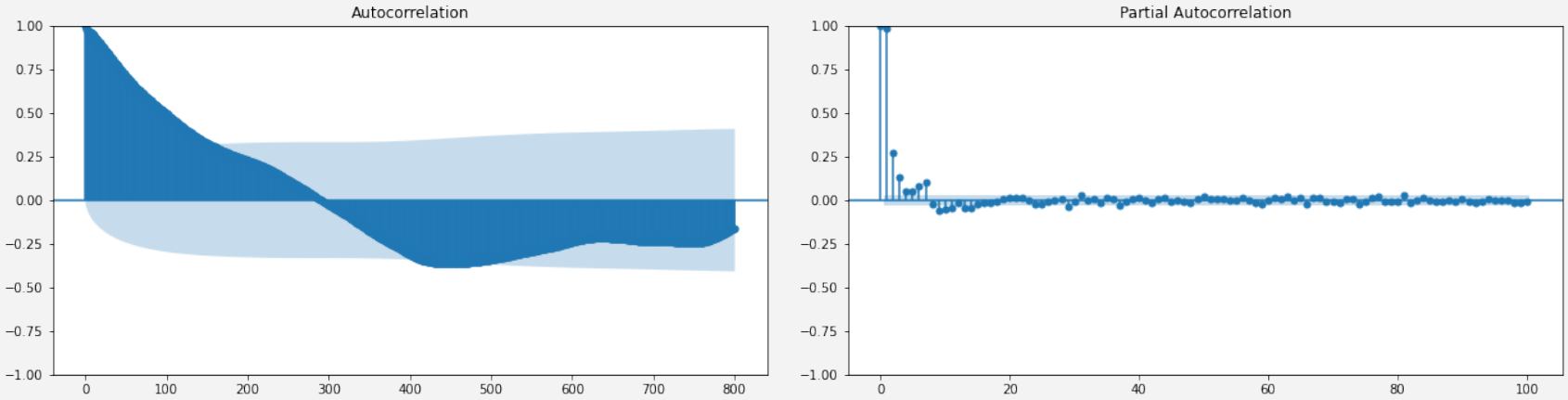


Figure: Acf e Pacf residui OLS

T (s)	MSE (cv)	AIC	BIC
1.56	0.173	-41806.40	-41741.6

Table: Risultati Sarimax Livello

Sarimax 2: Portata

Il modello scelto è $(1, 1, 1)$.

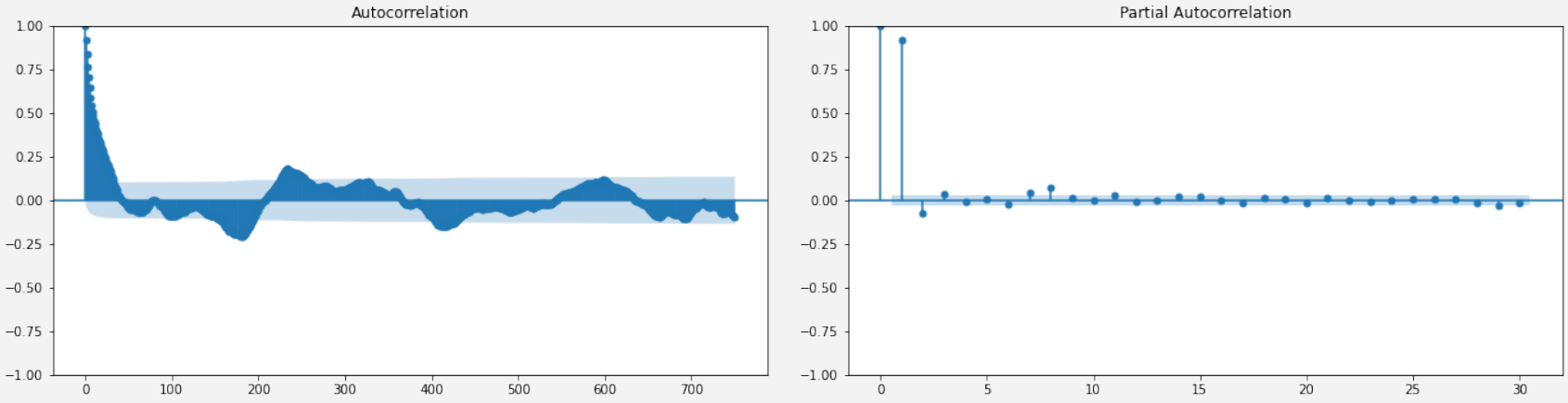


Figure: Acf e Pacf residui OLS

T (s)	MSE (cv)	AIC	BIC
28.6	2.67	-173.55	-95.81

Table: Risultati Sarimax Portata

Arch

Il modello Arch non produce dei risultati in termini di MSE, ma di intervalli di confidenza all'interno dei quali è ragionevole aspettarsi che la variabile target cada.

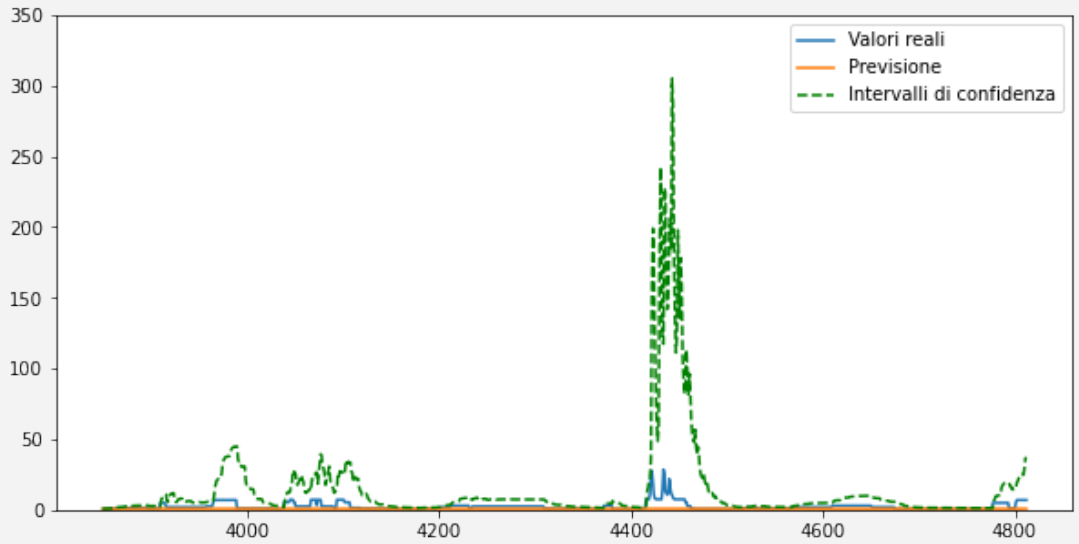


Figure: Intervallo di confidenza Arch

VAR

Il modello VAR consente di prevedere sia i target che gli attributi (Pioggia e Temperatura), basando la stima sulle interazioni tra feature anche nel futuro, in assenza dei dati reali.

	T (s)	MSE
Portata	35.6	10.602
Livello	35.6	0.126

Table: Risultati VAR

Symbolic Regression

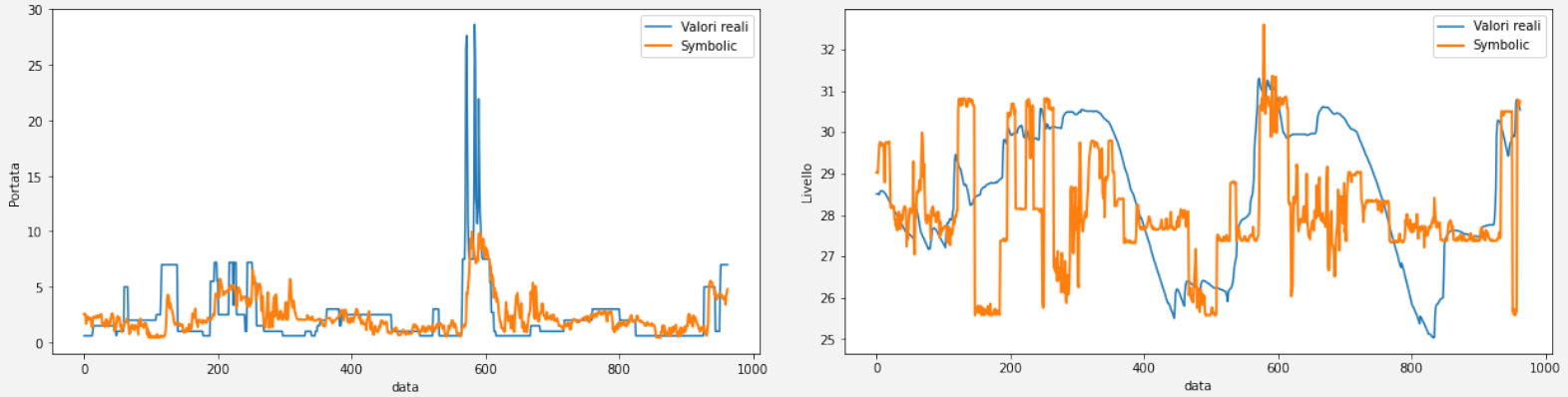


Figure: Previsione e dati reali per Portata e Livello

	T (min)	MSE	MSE (CV)
Portata	14,51	5.566	5.505
Livello	10,54	2.916	1.454

Table: Risultati Symbolic Regression

KNN

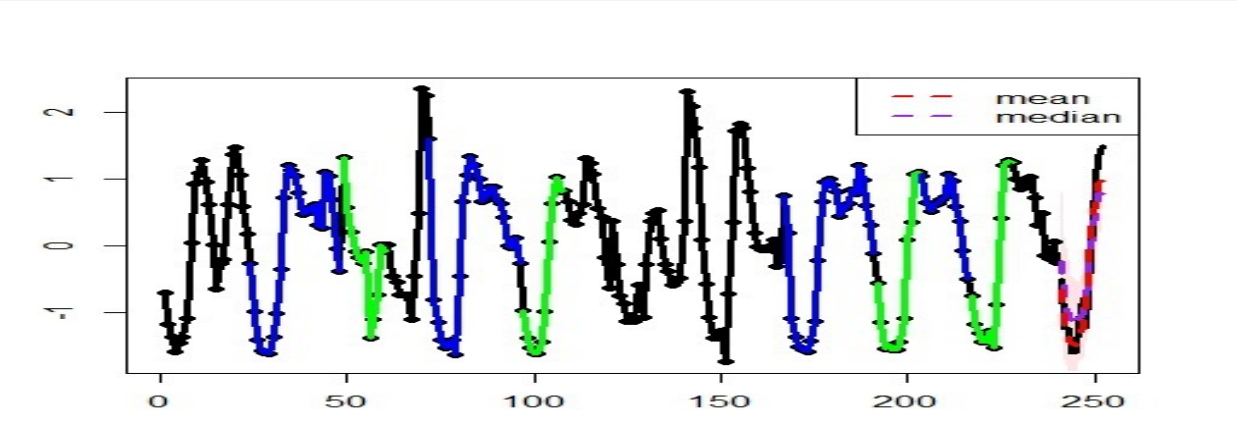


Figure: Esempio di applicazione del metodo

	T (min)	MSE
Portata	0,422	9.785
Livello	1,37	1.102

Table: Risultati KNN

Reti Neurali: Struttura

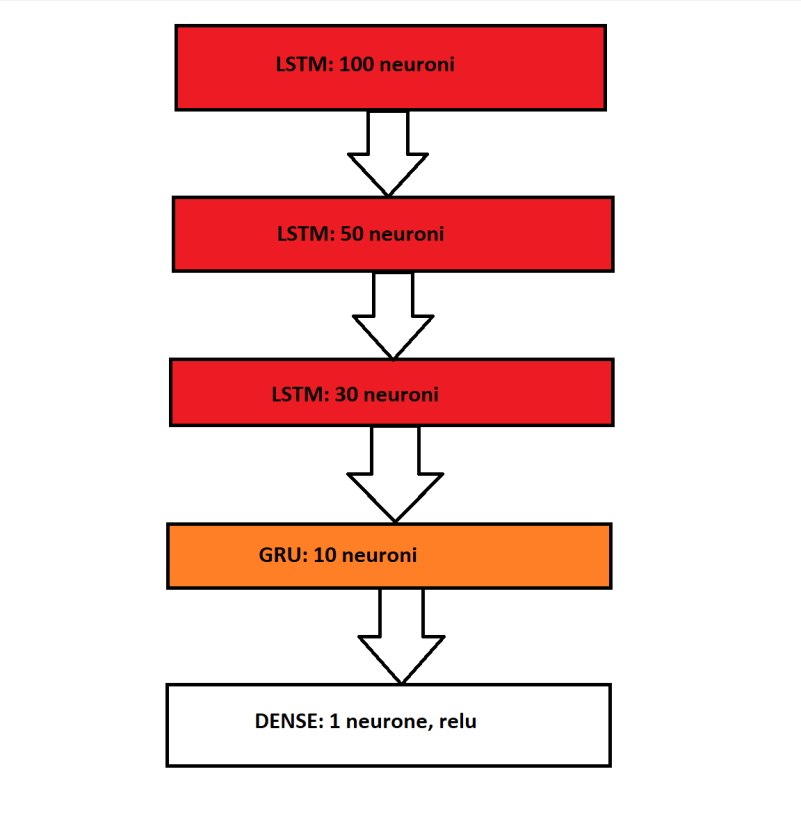


Figure: Struttura della rete

È stata definita una rete per ogni variabile target in quanto una rete a doppia uscita non portava vantaggi nei risultati

Reti Neurali: Risultati

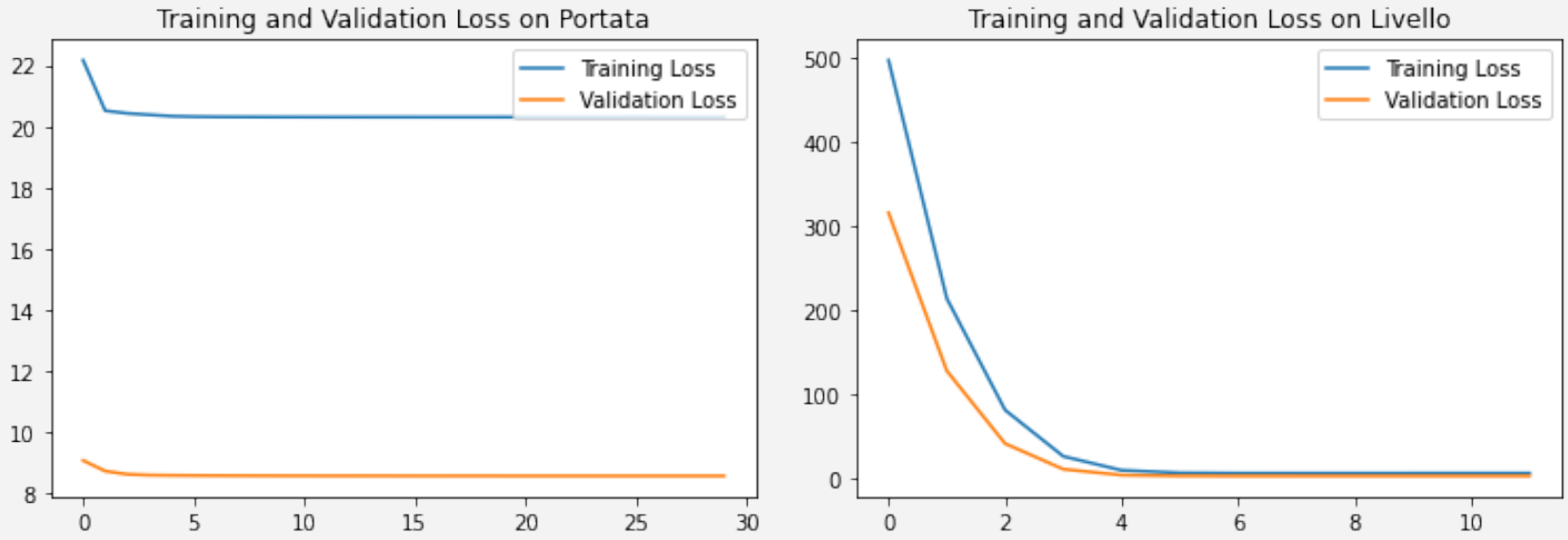


Figure: Funzioni di loss per Portata e Livello

	Epoche	T (sec)	MSE
Portata	18	42,2	8.568
Livello	12	30,8	2.765

Table: Risultati reti neurali

Modello Migliore

Confrontando i risultati dei modelli in termini di errore, di complessità e di tempo di addestramento, viene scelto il modello SARIMAX come migliore per entrambi i target

	T (s)	MSE
Portata	2.12	28.35
Livello	1.34	0.055

Table: Risultati testing modelli SARIMAX

Conclusioni

Conclusioni

La previsione del Livello si è dimostrata accurata a causa della periodicità e della stagionalità della serie, mentre, al contrario, la Portata è stata difficoltosa a causa dei numerosi e casuali picchi.

Il costo computazionale, per alcuni modelli, si è rivelato essere un forte impedimento.

Sviluppi futuri potrebbero riguardare l'utilizzo di macchine più potenti per poter allenare meglio e più velocemente alcuni modelli computazionalmente costosi.

Grazie!