Hackathon

Federico Molteni Damiano Orlandi Tommaso Rondani
Raffaele Sinani

1 Introduzione

Lo scopo di questo studio è prevedere la produzione oraria di un impianto per il mese di gennaio 2025 in base ai dati storici del 2024 e alle previsioni meteorologiche. Utilizziamo un approccio basato sull'intelligenza artificiale addestrato sui dati del 2024 (produzione effettiva e condizioni meteo) per stimare la produzione futura a partire da previsioni meteo. Di seguito descriviamo la struttura dei dati, i passi di preprocessing, il modello scelto e i risultati ottenuti.

2 Dati e Preprocessing

I dati a disposizione comprendono la produzione oraria dell'impianto nel 2024, i dati meteorologici orari osservati e previsti per il 2024, riferiti a cinque località: il comune in cui si trova l'impianto e 4 comuni limitrofi. Le variabili meteo disponibili per ciascuna località includono velocità del vento, direzione del vento, precipitazione e umidità relativa (tutte su base oraria).

In fase di preprocessing, i dati sono stati uniti e organizzati in un unico dataset orario. In particolare, a ogni ora del 2024 sono stati associati i valori delle variabili meteorologiche previsti delle cinque località.

Il risultato del preprocessing è un insieme di 8784 osservazioni orarie, ciascuna descritta da 96 caratteristiche e dal valore di produzione associato.

Successivamente, dato che si è notato un peggioramento dei risultati, è stata svolta un'operazione di *feature selection*, rimuovendo le variabili altamente correlate. Una tale operazione ha ridotto il numero di variabili a 16.

Per la previsione, è stato costruito in modo analogo il dataset con le osservazioni orarie per il mese di gennaio 2025, utilizzando solo le variabili meteo previste. Questo dataset di input è stato poi utilizzato per generare la stima di produzione.

3 Modello Predittivo

3.1 Multi-layer Perceptron

Il primo modello a essere testato è stato un semplice MLP. La performance sul dataset di base si dimostra inferiore ad altri modelli come XGBoost ed il Long Short Term Memory network.

La situazione cambia una volta che la *feature augmentation* viene applicata al dataset originario. Le nuove features migliorano molto la performance del MLP che arriva ad un MAPE di 0.36 sul validation set.

Testando varie configurazioni, la migliore risulta essere due semplici layer lineari con 32 e 1 neurone, applicando due layers di Layer Normalization tra loro.

3.2 LSTM

Poiché il problema in analisi è di natura sequenziale, le caratteristiche di un LSTM dovrebbero essere adatte alla situazione. Questo è confermato da un MAPE iniziale di 0.64 solo sulle 4 features del dataset di base.

La performance però non migliora notevolmente quando viene utilizzato il dataset augmentato, ma addirittura peggiora quando vengono utilizzate tutte le nuove variabili. Questo è probabilmente dovuto al fatto che molte delle variabili sono medie mobili che catturano già un aspetto sequenziale del problema, rendendo meno efficace il modello. Questo svantaggio, unito al

lungo tempo di training necessario, ha reso impossibile il fine tuning richiesto per selezionare le variabili più rilevanti e gli iperparametri migliori.

Durante la fase di testing, è stata verificata la capacità predittiva del modello sul periodo di gennaio 2025. A tal fine, il dataset *aumentato* è stato fornito al modello addestrato, ottenendo così la stima di produzione per ogni ora del mese. Queste stime sono state poi confrontate con i dati di produzione effettivamente osservati a gennaio 2025, allo scopo di valutare la performance.

4 Risultati e Discussione

Dai nostri test risulta particolarmente importante la fase di Data Augmentation, che porta numerosi benefici per quasi tutti i metodi che abbiamo testato.

Modelli più complicati come LSTM si sono dimostrati inadatti a causa del lungo tempo di training e il poco tempo a disposizione.

Modelli intermedi, seppur semplici, come XGBoost e il Multi-layer Perceptron si sono dimostrati una via di mezzo ideale per il problema.