Università degli studi di Milano - Bicocca

Department of Informatics, System and Communication (DISCo)

Master's Degree in Data Science

Streaming Data Management and Time Series Analysis Project

Davide Fabio Loreti - 865309

9 dicembre 2025

1 Esplorazione dei Dati e Preprocessing

1.1 Descrizione del Dataset

Il dataset oggetto di studio consiste in una serie temporale oraria che registra il numero di pedoni che transitano davanti a un sensore installato in una strada pubblica in Australia. I dati coprono un periodo di circa 5 anni, dal 15 aprile 2015 al 29 febbraio 2020, per un totale di 42.767 osservazioni orarie.

La struttura del dataset è composta da due variabili:

- `time`: variabile temporale con formato `yyyy-mm-dd hh:mm:ss` espressa in UTC
- `value`: variabile numerica rappresentante il conteggio di pedoni

L'obiettivo del progetto è prevedere le ultime 1.439 osservazioni mancanti, corrispondenti a 60 giorni di dati orari.

1.2 Visualizzazione della Serie Temporale

Figura 1: Serie temporale completa del conteggio pedonale e dettaglio degli ultimi 30 giorni osservati

Dal grafico emergono le seguenti osservazioni:

- La serie presenta una forte variabilità intra-giornaliera, con picchi e valli che si ripetono ciclicamente
- Si osserva una stagionalità evidente a livello giornaliero, con pattern che si ripetono ogni 24 ore
- Il dettaglio degli ultimi 30 giorni rivela chiaramente la presenza di cicli settimanali, con differenze visibili tra giorni feriali e weekend
- Non si osservano trend di lungo periodo evidenti, suggerendo una serie relativamente stazionaria attorno a un livello medio

1.3 Analisi dei Pattern Stagionali

1.3.1 Stagionalità Giornaliera

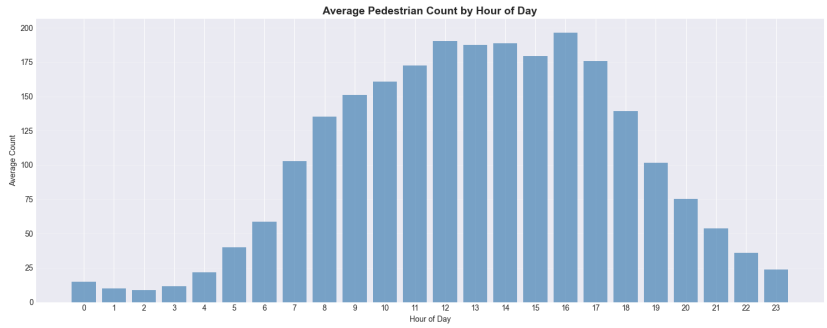

Figura 2: Conteggio medio di pedoni per ora del giorno

- **Ore di picco**: Il traffico pedonale raggiunge il suo massimo alle ore 16:00 con una media di circa 196 pedoni/ora, seguito da un plateau tra le 12:00 e le 17:00 con valori compresi tra 170 e 195 pedoni/ora
- **Ore notturne**: I valori minimi si registrano nelle ore notturne, in particolare tra l'1:00 e le 3:00 del mattino, con una media di circa 10 pedoni/ora
- **Crescita mattutina**: Si osserva un aumento graduale del traffico a partire dalle 6:00 (circa 60 pedoni/ora) fino a raggiungere il primo picco intorno alle 9:00 (circa 150 pedoni/ora)
- **Decrescita serale**: Dopo il picco pomeridiano, si registra una diminuzione progressiva del traffico a partire dalle 18:00 (circa 140 pedoni/ora) fino alle ore notturne

1.3.2 Stagionalità Settimanale

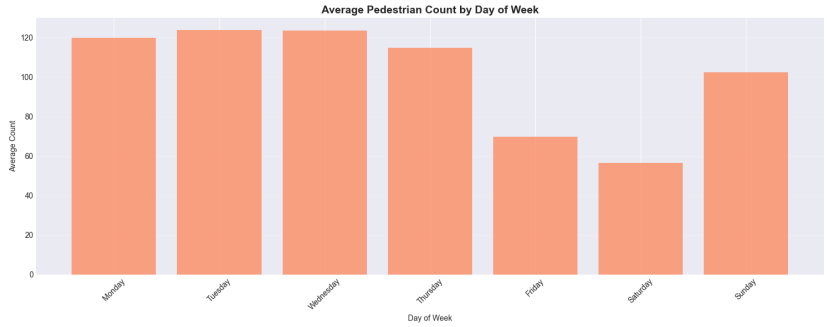

Figura 3: Conteggio medio di pedoni per giorno della settimana

L'analisi settimanale rivela un pattern distintivo tra giorni lavorativi e fine settimana. I primi tre giorni della settimana (lunedì, martedì e mercoledì) mostrano i valori più elevati, seguiti da un lieve calo il giovedì. Il venerdì segna una diminuzione più marcata, mentre il sabato registra il minimo settimanale. La domenica presenta una ripresa moderata, confermando una dinamica legata alle attività lavorative e scolastiche nei giorni feriali e a comportamenti ricreativi diversificati durante il weekend.

1.3.3 Stagionalità Annuale

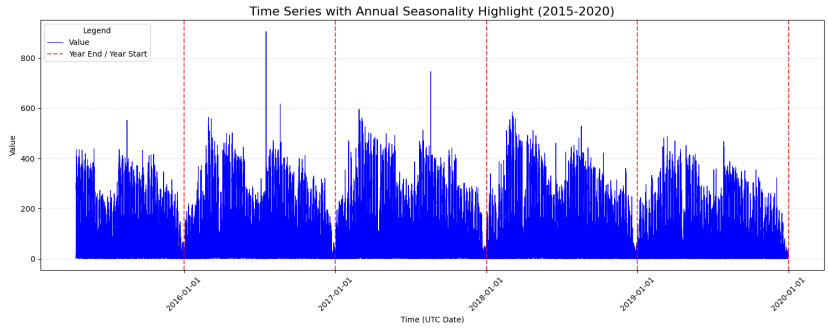

Figura 4: Conteggio medio di pedoni per mese dell'anno

La stagionalità annuale rivela variazioni nel comportamento pedonale su scala mensile. Si osservano mesi caratterizzati da maggiore affluenza, tipicamente associati a periodi turistici o condizioni climatiche favorevoli, e mesi con un calo significativo nei flussi pedonali, spesso in coincidenza con stagioni fredde o periodi festivi

1.4 Analisi di Autocorrelazione

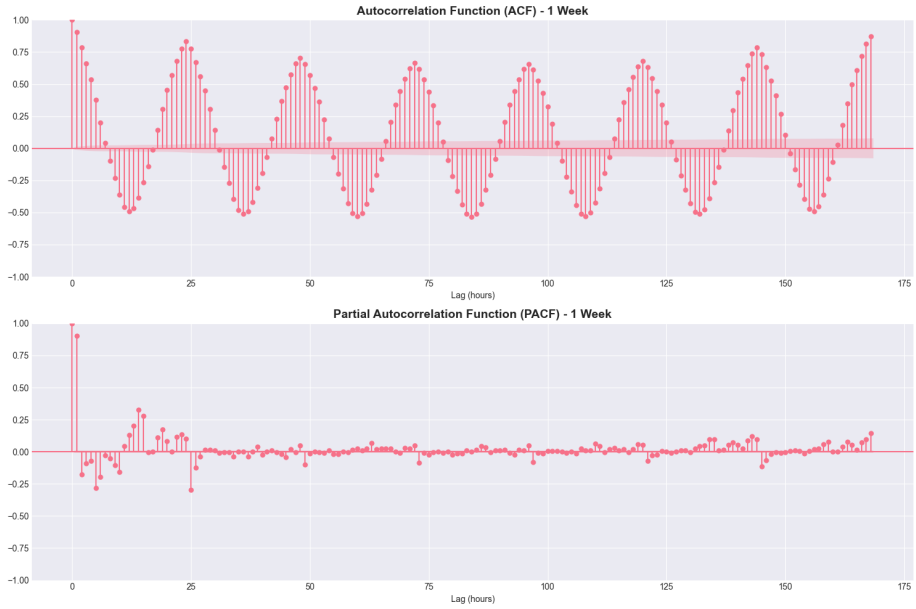

Figura 5: Funzioni di autocorrelazione (ACF) e autocorrelazione parziale (PACF) fino a 168 lag (1 settimana)

1.5 Stazionarietà

Per verificare la stazionarietà della serie temporale analizzata, sono stati condotti due tipi di test statistici: uno per la stazionarietà in media e uno per la stazionarietà in varianza.

Stazionarietà in media Per testare la stazionarietà in media è stato applicato l'*Augmented Dickey-Fuller (ADF) test*. L'output ottenuto è il seguente:

```
Augmented Dickey-Fuller Test
data:  ts_data
Dickey-Fuller = -18.53, Lag order = 34, p-value = 0.01
alternative hypothesis: stationary
```

Il valore del test Dickey-Fuller è significativamente negativo e il p-value (0.01) è inferiore al livello di significatività del 5%, suggerendo quindi il rifiuto dell'ipotesi nulla di non stazionarietà e confermando che la serie è stazionaria in media.

Stazionarietà in varianza Per verificare la stazionarietà in varianza è stato condotto il *test ARCH* (LM-test). L'output riportato è il seguente:

```
ARCH LM-test; Null hypothesis: no ARCH effects
data:  ts_data
Chi-squared = 30093, df = 12, p-value < 2.2e-16
```

Il risultato indica un p-value molto basso ($<$ 2.2e-16), suggeriscd che la varianza della serie storica non è costante nel tempo, indicando la necessità di modellare la volatilità in maniera appropriata.

2 Modellazione ARIMA

A seguito dell'analisi esplorativa e della valutazione di stazionarietà, è stata intrapresa la modellazione tramite approccio ARIMA (AutoRegressive Integrated Moving Average), con l'obiettivo di catturare sia la componente autoregressiva sia la struttura degli errori serialmente correlati presenti nella serie.

La selezione del modello è stata effettuata in ambiente R utilizzando la funzione `auto.arima()` del pacchetto `forecast`, che permette la ricerca automatica della combinazione ottimale dei parametri (p, d, q) e dei parametri stagionali $(P, D, Q)_m$ in funzione delle caratteristiche sia stagionali che non stagionali della serie.

Il modello finale selezionato è un **ARIMA(2,1,2)** con regressori esterni costituiti dalle componenti di Fourier giornaliere, settimanali e annuali, oltre a dummy per festività e weekend. La matrice dei regressori ha dimensione 41328×41, comprensiva di tutte le componenti stagionali e indicatori di giorni speciali.

2.1 Parametri del Modello e Metriche di Adattamento

I principali parametri stimati includono:

- AR1 = 0.5281, AR2 = -0.9616
- MA1 = -0.5922, MA2 = 0.9205
- Coefficienti di Fourier giornalieri, settimanali e annuali con valori compresi tra -79.70 e 32.81
- Dummy festività e weekend tra -2.20 e 1.13

Le metriche di bontà di adattamento sul training set sono:

- MAE = 25.84
- RMSE = 38.07
- AIC = 418198.1, BIC = 418595.1

2.2 Previsioni e Analisi

Le previsioni generate dal modello ARIMA esteso sono state confrontate con gli ultimi 30 giorni osservati, mostrando una buona capacità di catturare la stagionalità giornaliera e settimanale. La distribuzione dei valori previsti varia da 0 a circa 210 pedoni/ora, con mediana intorno a 68 e media circa 76. QUesti valori, reativamente contunti rispetto ai picchi, suggerisce un buon forecast sul breve periodo ma scarso sul lungo.

Le festività e i weekend sono stati analizzati separatamente, risultando coerenti con i pattern storici osservati.

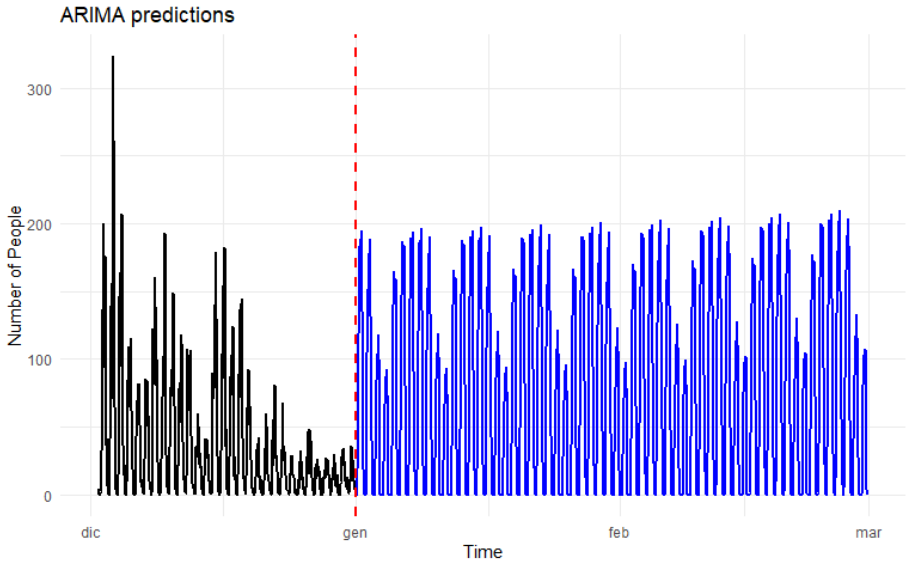

Figura 6: Confronto tra osservazioni storiche (nero) e previsioni ARIMA (blu). La linea tratteggiata rossa indica l'inizio del periodo predittivo.

3 Modellazione UCM (Unobserved Components Model)

Per confrontare i risultati ottenuti tramite ARIMA, è stato implementato un modello UCM con i seguenti componenti:

- **Local level trend**: stazionario (Q = 0)
- **Stagionalità oraria**: 24 ore (dummy)
- **Regressori di Fourier**: giornalieri (K=12), settimanali (K=6) e annuali (K=1)

- **Dummy per festività e weekend**
- **Trasformazione dei dati**: Log(y+1), per stabilizzare la varianza rendere additivi eventuali valori moltiplicativi e prevernire valori negativi in fase di forecasting.

Il modello è stato stimato tramite il filtro e lo smoothing di Kalman. La previsione è stata generata per le ultime 1.439 ore (circa 60 giorni).

3.1 Statistiche di Forecast

- **Periodo di training**: 15 aprile 2015 - 31 dicembre 2019
- **Periodo di forecast**: 1 gennaio 2020 - 29 febbraio 2020
- **Numero di regressori**: 42 (Fourier giornalieri, settimanali, annuali + dummy festività)
- **Metriche in-sample** (scala originale):
  - MAE = 35.37
  - RMSE = 59.26

3.2 Forecast UCM

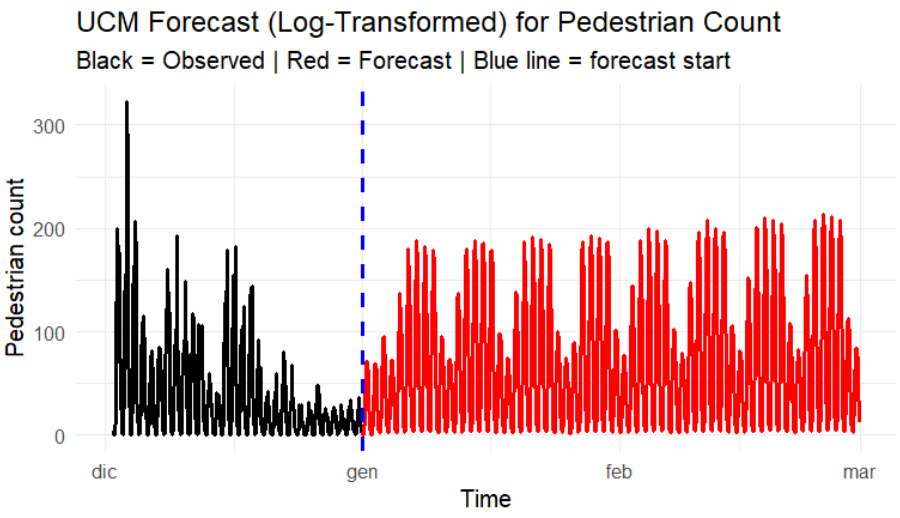

Figura 7: Previsioni generate dal modello UCM per il conteggio pedonale.

4 Modellazione con Algoritmi di Machine Learning

Oltre ai modelli statistici ARIMA e UCM, è stata condotta un'estesa sperimentazione con algoritmi di Machine Learning per valutare la capacità predittiva sulla serie temporale pedonale. Gli algoritmi confrontati includono: LSTM, KNN, Random Forest e XGBoost.

I risultati principali sono riportati di seguito.

4.1 Metriche dei Modelli

- **LSTM**: MAE = 14.702
- **KNN**: MAE = 18.489
- **Random Forest**: MAE = 13.548
- **XGBoost**: MAE = 13.323

Sebbene Random Forest e XGBoost abbiano il MAE più basso, la sola metrica non cattura la fedeltà alla dinamica reale della serie, poiché i modelli mostrano previsioni compresse con variabilità molto inferiore rispetto ai dati osservati.

4.2 Analisi delle Statistiche di Forecast

- **LSTM**: Mean = 176.71, Std = 150.59, Min = 0.00, Max = 555.74
- **KNN**: Mean = 50.80, Std = 50.68, Min = 0.00, Max = 303.40
- **Random Forest**: Mean = 15.79, Std = 10.62, Min = 0.98, Max = 40.96
- **XGBoost**: Mean = 13.45, Std = 10.91, Min = 1.41, Max = 42.56
- **Dati reali**: Mean = 101.51, Std = 103.17, Min = 0.00, Max = 906.00

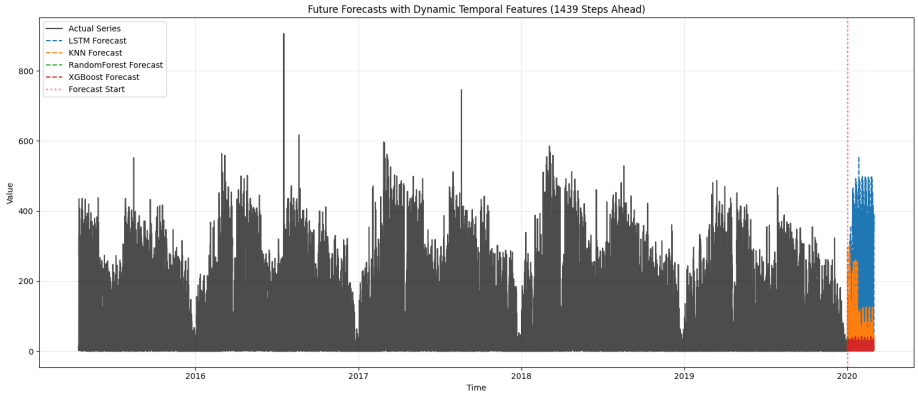

Figura 8: Forecast dei modelli con la serie originale.

RF e XGBoost risultano eccessivamente smorzati: generano forecast troppo piatti, con un range molto inferiore alla variabilità reale. Il KNN ha una variabilità maggiore ma ancora insufficiente.

LSTM, al contrario, mantiene una distribuzione più coerente con la dinamica reale, catturando picchi, stagionalità e ampiezza delle oscillazioni.

4.3 Scelta del Modello LSTM

Il modello selezionato per il forecasting finale è l'LSTM. La scelta è motivata non solo dalla buona accuratezza (MAE = 14.702), ma soprattutto dalla capacità del modello di adattarsi alla forma globale della serie, mantenendo:

- una variabilità predittiva compatibile con quella reale;
- una struttura stagionale più credibile rispetto ad altri modelli ML;
- una capacità di catturare sia picchi sia pattern giornalieri e settimanali;
- un comportamento dinamico non appiattito come nei modelli tree-based.

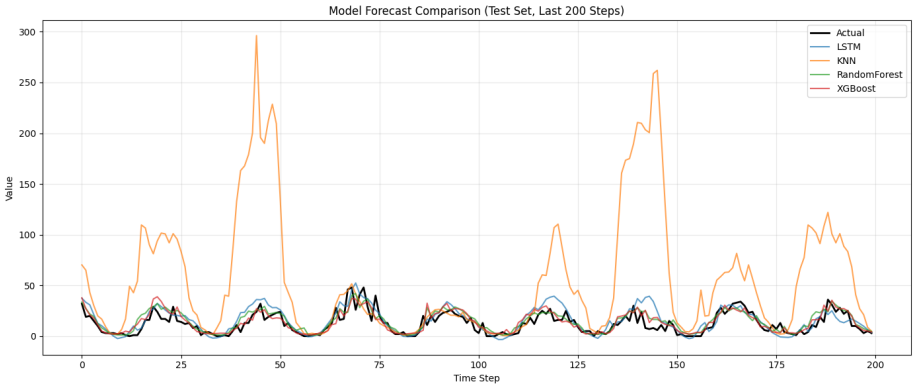

Figura 9: Adattamento dei modelli nelle ultime 200 osservazioni.