

# Linee guida progetto esame di Streaming Data Management and Time Series Analysis

Prof. Matteo Pelagatti e Prof. Antonio Candelieri

AA 2020-2021

Il progetto ha come obiettivo la definizione, lo sviluppo e la validazione di un sistema predittivo per dati di tipo time series. Lo studente dovrà implementare:

- almeno un modello ARIMA
- almeno un modello UCM
- almeno un modello predittivo non-lineare (Machine Learning)

Il dataset è contenuto nel file "TrainingSet.csv" presente sul Moodle del corso.

Il file è quindi costituito da 3 colonne:

- DATA [yyyy-mm-dd]
- Ora 1,...,24
- VALORE numeric

Il periodo di riferimento è da **2018-09-01** a **2020-08-31**.

## Obiettivo

Predire l'andamento della serie temporale VALORE dal **2020-09-01** al **2020-10-31** (2 mesi) usando il MAE come misura dell'errore

## Output: formato file previsione

Lo studente dovrà inviare un file csv con le seguenti colonne:

- Data [yyyy-mm-dd] (tutte le date dal **2020-09-01** al **2020-10-31**)
- Ora 1,...,24
- ARIMA (previsione fornita da quello che si ritiene essere il miglior modello ARIMA)
- UCM (previsione fornita da quello che si ritiene essere il miglior modello UCM)
- ML (previsione fornita da quello che si ritiene essere il miglior modello Machine Learning)

**NOTA: le previsioni verranno validate in modo automatico ed è quindi cruciale attenersi strettamente al formato previsto:**

- **Utilizzare il punto e virgola come separatore**
- **Usare il formato yyyy-mm-dd per le date**
- **Le ore sono interi da 1 a 24**
- **Nessun nome di riga**
- **Intestazione del file: Data; Ora; ARIMA; UCM; ML**

Il file dovrà essere denominato come "SDMTSA\_<matricola>\_<submission>.csv", dove <matricola> va sostituito con il numero di matricola dello studente e <submission> con un numero intero che rappresenta il numero di sottomissione (da usare nel caso in cui lo stesso studente voglia "aggiornare" la propria previsione, fino alla submission finale).

## Modalità di consegna

Inviare una mail ai Prof. Pelagatti e Candelieri con allegati:

1. Un file csv contenente le previsioni dei (migliori) modelli ARIMA, UCM e Machine Learning identificati
2. Un elaborato che indirizzi i seguenti punti:
  - definizione dello split training-validation scelto dallo studente (ad esempio 70%-30%, 80%-20% o altro)
  - descrizione sintetica delle metodologie utilizzate e configurazioni considerate (ad esempio, architettura delle reti neurali)
  - per ogni strategia, confronto valore di errore (MAE) su training set e validation set. Confronto errore tra le strategie.
  - considerazioni finali
3. Codice implementato

La consegna dovrà avvenire almeno una settimana prima della data dell'appello che si intende sostenere.