Progetto G

Dato il dataset che contiene l'evoluzione della simulazione della rete idrica in figura,

realizzare un progetto di data analysis che prevede la profilazione dei nodi della rete tramite le seguenti tecniche di clustering:

1. PCA (Principal Component Analysis)
2. Clustering K-Means

Il dataset contiene per ogni nodo (82 nodi in tutto) le misure orarie. Raggruppare le misure per nodo e mediare in modo da assegnare ad ogni nodo un'unica riga di misura da utilizzare per il raggruppamento.

Il progetto deve mostrare eventuali correlazioni tra lo spazio delle features, in termini di grafici e matrice di correlazione. Oltre al plot con la proiezione dei dati nelle due componenti principali, mostrare Scree Plot e Biplot.

Applicare inoltre un metodo di StandardScalar ai dati prima delle tecniche di clustering (nella standardizzazione di tipo scale ogni variabile di input si modifica sottraendo la media e dividendo per la deviazione standard per spostare la distribuzione in modo che abbia una media pari a zero e una deviazione standard pari a uno).

Utilizzare il metodo Elbow e Silhouette Score, per scegliere il numero corretto di cluster. Valutare la possibilità di selezionare un sott'insieme di features per applicare la cauterizzazione. Applicare la clusterizzazione (k-means) sulle proiezioni delle prime due componenti principali. Nello scatter plot che visualizza i cluster, mostrare anche il centro di massa per ognuno dei cluster.

Dataset:

1M_one_res_small_no_leaks_ordered_new_delimited_merged.csv

Output progetto:

- Relazione di progetto (circa 10 pagine) con la descrizione dell'analisi progettuale. Questa deve includere la descrizione delle features, l'intera pipeline di analysis, e risultati ottenuti. Il codice deve essere consegnato su file separati, la relazione può comunque riportare eventuali funzioni di importanza e le relative descrizioni.

- Codice sorgente del progetto.

Modalità di consegna dei progetti:

- Invio e-mail all'indirizzo domenico.garlisi@unipa.it, si suggerisce di specificare nell'oggetto NOME_GRUPPO-ID-PROGETTO, es: MARIO-ROSSI-1 allegando:
    - PDF della relazione
    - ZIP file contenente i codici sorgente in python.