Big Data Analytics

Attività – Data Analytics

Data processing e exploratory data analytics su Data set provenienti da più sorgenti

Obiettivo in breve: L'attività consiste nello sviluppare un progetto di Data Analytics in un ambito a proprio piacere e di proprio interesse finalizzato allo storytelling ovvero a trovare risposta a più quesiti di ricerca.

Quanti e quali dataset? Il progetto deve partire da due o più dataset provenienti da *almeno due sorgenti distinte*. Diversi siti Web consentono di scaricare dataset pubblici come ad esempio open data, è possibile effettuare lo scraping di dati pubblicati in pagine Web, infine sul sito di Kaggle all'indirizzo https://www.kaggle.com/datasets sono disponibili diversi data set. Le pagine da dove è possibile scaricare i dataset ne forniscono una descrizione e, in alcuni casi, anche suggerimenti di quesiti che potrebbero essere indagati usando il dataset stesso.

Descrizione:

***Attività* 1**

- Scegliere uno (o più) dataset orientato all'analisi di dati in formato tabulare.
- Usando PANDAS implementare le operazioni di data processing necessarie (principalmente join e selezioni) per mettere in collegamento i dataset e per preparare i dati al passo successivo
- Usando pacchetti Python quali Pandas, scipy, matplotlib, seaborn, sklearn, statsmodel, implementare attività di data cleaning, exploratory data analysis estraendo dati statistici e di visualizzazione dei risultati attraverso il quale sia possibile "raccontare qualcosa sui dati" (storytelling), eventualmente partendo da dei quesiti di ricerca. L'uso dei pacchetti non deve necessariamente essere limitato alle istruzioni viste a lezione. Le documentazioni dei pacchetti stessi ed i testi consigliati forniscono spunti d'uso interessanti!!
- Produrre un notebook Jupyter (https://jupyter.org/) che contenga:
  - Un'introduzione all'argomento scelto, alle sorgenti dati e agli obiettivi del progetto specificando i quesiti di ricerca o i motivi alla base della scelta del dataset.
  - Una sezione per ogni fase del progetto di data analytics. Ricordo che a lezione abbiamo affrontato argomenti quali: Data wrangling e cleaning con Pandas; Misure di centralità e dispersione; Visualizzazione dei dati; Correlazione tra variabili e misure di similarità; Gestione ed imputazione di valori mancanti; Outlier detection; Feature rescaling e feature engineering.

***Attività* 2**

- Scegliere un dataset relativo ad una serie temporale
- Usando PANDAS implementare le eventuali operazioni di data preprocessing necessarie a rendere uniforme la frequenza di campionamento della serie e a gestire eventuali dati mancanti o outliers.
- Comprendere il significato della serie temporale ed individuare e il comportamento di eventuali componenti sistematiche e non-sistematiche della serie.
- Rendere la serie stazionaria ed, eventualmente, applicare un metodo autoregressivo per fare previsione (se applicabile).
- Produrre un notebook Jupyter (https://jupyter.org/) che contenga:
  - Introduzione e descrizione del dataset.
  - Una sezione per ogni fase del progetto di data analytics

Consegna: Upload del notebook al corrispondente link nella pagina Moodle del corso.

Valutazione: Le attività verranno valutate sulla base dei seguenti criteri:
1. Motivazione. L'elaborato fa credere al lettore che l'argomento sia rilevante o importante (i) in generale e (ii) rispetto alla scienza dei dati?
2. Comprensione. Dopo aver letto l'elaborato, un lettore non informato si sente informato il tema? Un lettore che già conosce l'argomento ha l'impressione di aver imparato di più?
3. Storytelling. La parte in prosa dell'elaborato è convincente?
4. Codice. Il codice è ben scritto, ben documentato, riproducibile e aiuta il lettore a capire? Fornisce buoni esempi di tecniche specifiche?

Scadenza per premio partecipazione: **06/13/2024**