Big Data Analytics

Attività – Data Analytics

Svolgere la seguente attività e inviare via mail la consegna al docente (federica.mandreoli@unimo.it):

Data processing e exploratory data anlytics su Data set di Gaggle
Sul sito di Kaggle all'indizzo 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.
Alcuni di questi dataset sono in formato CSV e si prestano ad attività di data processing e exploratory data analysis.

Un elenco non esaustivo è il seguente:

- UFC data: https://www.kaggle.com/rajeevw/ufcdata
- Apple app store strategy game: https://www.kaggle.com/tristan581/17k-apple-app-store-strategy-games
- Africa Economic, Banking and Systemic Crisis Data:
- https://www.kaggle.com/chirin/africa-economic-banking-and-systemic-crisis-data
- Sofia air quality dataset: https://www.kaggle.com/hmavrodiev/sofia-air-quality-dataset
- Military Spending of Countries (1960-2019) https://www.kaggle.com/nitinsss/military-expenditure-of-countries-19602019 (Da integrare eventualmente con dati sugli stati (come popolazione, reddito medio, ecc.)
- European Social Survey (ESS) 8 ed2.1 (2016/17): https://www.kaggle.com/pascalbliem/european-social-survey-ess-8-ed21-201617#variables.csv
- CO2 and GHG emission data of different countries from 1750 – 2019: https://www.kaggle.com/srikantsahu/co2-and-ghg-emission-data
- Daily statistics for trending YouTube videos: https://www.kaggle.com/datasnaek/youtube-new

L'attività da svolgere consiste nel:

- Scegliere un dataset
- Usando PANDAS implementare le operazioni di data processing necessarie per mettere in correlazione i dataset e per preparare i dati al passo successivo (join e selezioni)
- Usando pacchetti Python quali Pandas, scipy, matplotlib e sciborn implementare attività di 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 fornisco spunti d'uso interessanti!!

Esempi completi di progetti di data analytics sono i seguenti:
https://jiglesia3.github.io/
https://krixly.github.io/
https://andresgogo.github.io/

Consegna: PDF commentato con discussione e codice Python in un unico zip. In alternativa è possibile (e gradito) produrre e consegnare un notebook jupyter .ipynb (https://jupyter.org/).

Scadenza per premio partecipazione: **22/12/2019**