ESERCITAZIONE D1 - DATA ANALYST

# DATASET STRUTTURATI

Un esempio di dataset strutturato scelto è “Air Quality Data Set”, estratto dalla seguente risorsa <https://archive.ics.uci.edu/ml/datasets/Air+quality>. Contiene le risposte di un dispositivo multi-sensore per gas installato sul campo in una città italiana. Vengono registrate le medie orarie delle risposte e i riferimenti delle concentrazioni di gas di un analizzatore certificato. Con un totale di 9358 indicativi. Gli attributi del dataset sono i seguenti 15:

* Date (DD/MM/YYYY)
* Time (HH.MM.SS)
* True hourly averaged concentration CO in mg/m^3 (reference analyzer)
* PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
* True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)
* True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
* PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
* True hourly averaged NOx concentration in ppb (reference analyzer)
* PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
* True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
* PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
* PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
* Temperature in Â°C
* Relative Humidity (%)
* AH Absolute HumidityGli indicativi (oggetti) sono

Un altro dataset strutturato è “Leaf Data Set”, estratto da <https://archive.ics.uci.edu/ml/datasets/Leaf>. Questo set di dati consiste in una raccolta di caratteristiche di forma e texture estratte da immagini digitali di esemplari di foglie provenienti da un totale di 40 specie vegetali diverse (come, indicativi). Gli attributi considerati sono 16:

* Class (Species)
* Specimen Number
* Eccentricity
* Aspect Ratio
* Elongation
* Solidity
* Stochastic Convexity
* Isoperimetric Factor
* Maximal Indentation Depth
* Lobedness
* Average Intensity
* Average Contrast
* Smoothness
* Third moment
* Uniformity
* Entropy

Un ultimo esempio di dataset di tipo strutturato è “Communities and Crime Data Set”, <https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime>. Descrive la comunità all'interno degli Stati Uniti. I dati combinano i dati socio-economici del censimento degli Stati Uniti del 1990, i dati sulle forze dell'ordine dell'indagine LEMAS del 1990 e i dati sulla criminalità dell'FBI UCR del 1995. È caratterizzato da 128 attributi e 1994 indicativi.

Per motivi di spazio riporto solo i primi 5 attributi, come esempio:

* state numeric
* county numeric
* community numeric
* communityname string
* fold numeric

# DATASET NON STRUTTURATI

Un esempio di dataset non strutturato potrebbe essere una playlist con 100 canzoni (indicativi/oggetti) che devono essere ancora catalogate. Possibili attributi potrebbero essere genere, artista, nome, lunghezza.

Un secondo esempio di dataset non strutturato potrebbe essere l’insieme di foto ancora presenti nella fotocamera (indicativi/oggetti), e quindi non ancora catalogate (elaborate). Una possibile elaborazione potrebbe essere fatta, una volta scaricate su PC, con data, luogo, grandezza file, tipo di file, tipologia di foto (paesaggio, ritratto, macro).