**BIG DATA ANALYSIS**
28/01/2020

| Nome: | Cognome: |
|---|---|
| Matricola: | |

| | |
|---|---|
| Parte 1 | |
| Parte 2 | |
| Totale | |

**Note**:
Durata della prova: 2 ore. Il file csv che si trova al link http://bit.ly/wea_2020
Rispondere nel file notebook alle domande.
Salvare frequentemente il file notebook creato attribuendogli il proprio nome-cognome.
Al termine della prova spedire a francesco.guerra@unimore.it il file della prova o il notebook direttamente o la versione html (file / download as / HTML) (oggetto della mail BDA_GEN_2)

## Parte 0: Il Dataset

Il datataset weather_train.csv (preso da kaggle -- https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather) contiene dati relativi a rilevazioni meteo registrate in città spagnole una volta al giorno secondo il seguente schema:
'dt_iso', 'city_name', 'temp', 'temp_min', 'temp_max', 'pressure', 'humidity', 'wind_speed', 'wind_deg', 'rain_1h', 'rain_3h', 'snow_3h', 'clouds_all', 'weather_id', 'weather_main', 'weather_description', 'weather_icon'
Il dataset è costituito da attributi con valori numerici e categorici.
L'obiettivo è quello di prevedere il tempo complessivo di una giornata (valore della feature 'weather_main') sulla base degli altri parametri.

## Parte 1: Analisi (10 punti)

1. Quante sono le istanze contenute nel dataset? _____ Il dataset è completo (cioè per ogni istanza tutti i valori di attributo sono sempre correttamente specificati – non esistono "missing values")? _____ Il dataset è bilanciato per quanto riguarda la classe da predire? _____ (punti 1).

2. Le rilevazioni con pressione e umidità uguale a 0 sono irreali. Quante sono queste rilevazioni? Eliminarle dal dataset (punti 1)

3. Analizzare la temperatura massima rilevata. Valutare se la distribuzione dei valori assume un andamento simile a una gaussiana. Considerare poi le rilevazioni che si collocano all'interno del 5% delle temperature più alte. Le città sono equamente presenti in quella fascia di rilevazioni? Come è il tempo complessivo nei giorni in cui la temperatura massima è in quella fascia per ogni città? (punti 4)

4. Verificare se quando nevica la temperatura sia prossima alla temperatura di congelamento (NOTA: il dataset riporta i valori in Kelvin) (punti 2)

5. Confrontare l'escursione termica media (temp_max-temp_min) registrata nei giorni in cui nevica, con quella delle giornate che sono all'interno del 5% delle temperature più alte (punti 2)

## Parte 2: Trasformazione e Predizione (20 punti)

1. Si vuole predire il valore di 'weather_main' sulla base degli attributi presenti nel dataset. Dividere il dataset in modo che 2/3 degli elementi siano contenuti in un nuovo dataset “train” e 1/3 nel dataset “test”.

Eliminare gli attributi ["dt_iso","city_name","weather_description", "weather_icon","weather_id", "clouds_all"]
Convertire l’attributo 'weather_main' in numerico in maniera opportuna.

Allenare il train con il modello Decision Tree e valutare l’accuracy ottenuta calcolata sia sul dataset train sia sul dataset test. Confrontare i risultati ottenuti con quelli ottenuti con una predizione basata sul modello Logistic Regression. Effettuare alcune considerazioni sui risultati ottenuti, tenendo in considerazione anche l’analisi della confusion matrix. (punti 4)

2. Confrontare l’accuratezza ottenuta nel punto precedente con l’accuratezza che si ottiene con un una 10 Fold cross validation. (punti 1)

3. Utilizzare la funzione Normalizer per normalizzare i valori del dataset e confrontare se l’accuratezza ottenuta con il Decision Tree Classifier migliora (punti 3).

4. Creare una pipeline con trasformatori PCA (si scelgano 5 attributi) e poi Normalizer. Si usi come modello il Decision Tree Classifier (punti 2) [2 punti ulteriori se gli attributi della PCA sono aggiunti agli attributi del dataset]

5. Utilizzare la funzione di gridSearchCV sulla pipeline per modificare il numero di attributi selezionati dalla PCA e alcuni parametri a piacere del classificatore. Verificare se l’accuratezza che si ottiene con la nuova configurazione supera quella standard ottenuta al punto 1 (punti 4)

6. Si verifichi l’accuratezza ottenuta dalla pipeline del punto 4 con il file weather_test. I risultati corretti sono nel file class.csv. Controllare le features presenti nei dataset. (punti 2).

7. Si sperimenti una pipeline come quella del punto 4 dove al posto del classificatore si utilizzi un regressore lineare. Il risultato dovrà essere approssimato all’intero per il calcolo dell’accuratezza (punti 2).