BIG DATA ANALYSIS
04/09/2018

NOME	
COGNOME	
MATRICOLA	

Parte 0: Il Dataset

Il file weather.csv () contiene dati meteo rilevati in alcune città australiane. Si vuole predire se il giorno successivo pioverà.

Lo schema del dataset è il seguente

- Month: mese in cui avviene la rilevazione del dato
- Location: città in cui avviene la rilevazione
- MinTemp, MaxTemp: temperature minima e massima
- Rainfall: quantitativo di pioggia caduta
- WindGustSpeed, WindSpeed9am, WindSpeed3pm: misurazioni relative al vento
- Humidity9am, Humidity3pm: misurazioni relative all'umidità
- Pressure9am, Pressure3pm: misurazioni relative alla pressione
- Cloud3pm: nuvolosità in ottavi: https://it.wikipedia.org/wiki/Okta
- Temp9am, Temp3pm: misurazioni relative alla temperatura
- RainToday, Yes/No
- RainTomorrow, la classe da predire

Note:

Durata della prova:2 ore. Rispondere nel file notebook.
Creare una cartella esame e scaricare in essa il file csv che si trova al link
http://bit.ly/BDAweather2018
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).
.

Parte 1: Analisi (10 punti)

1. Caricare il dataset introducendo un opportuno nome per le colonne e denominarlo con una variabile chiamata “dataset”

2. Quante sono le istanze contenute nel dataset? _____ Il dataset è completo (cioè non esistono valori nulli)? _____ Il dataset è bilanciato per quanto riguarda la classe da predire? _____ Il numero di rilevazioni per città è bilanciato? _____ (punti 2)

3. Rappresentare in un grafico la frequenza delle rilevazioni mensili per città. (punti 3)

4. Calcolare per ogni città e per ogni mese l’umidità minima e la massima. (punti 2)

5. Creare un nuovo attributo “TemperatureRange” che mostri l’escursione termica giornaliera. Rappresentare in un grafico l’escursione massima mensile. (punti 3)

Parte 2: Trasformazione e Predizione (20 punti)

1. Si vuole predire la possibilità di avere pioggia il giorno successivo (RainTomorrow è la classe da predire).
Trasformare i valori degli attributi RainToday e RainTomorrow da No a 0, e da Yes a 1.
Creare un nuovo dataset chiamato reduce con le istanze del dataset per le quali c’è un valore di Cloud3pm maggiore o uguale a 0.

Dividere il dataset in modo che 4/5 degli elementi siano contenuti in un nuovo dataset “train” e 1/5 in un dataset “test”.
Valutare l’accuracy ottenuta con il classificatore Logistic Regression e il Decision Tree

Il valore di accuratezza ottenuto è pari a _____ La confusion matrix evidenzia delle peculiarità? (punti 4)

2. Che valore di accuratezza si ottiene con un 5 Fold cross validation e il classificatore basato su Decision Tree _____ e quello basato su Logistic Regression _____
Il valore di accuratezza maggiormente rappresentativo è quello che si ottiene con questa tecnica o con quella attuata in precedenza? (punti 2)

3. Si introduca un attributo che sostituisca per ogni rilevazione i due valori di temperatura “MinTemp, MaxTemp” con il valore medio delle registrazioni. Si faccia lo stesso con vento (WindSpeed9am, WindSpeed3pm). Che valore di accuratezza si ottiene? _____ (punti 4)

4. Si riparta dal dataset originario e si considerino due nuovi dataset ottenuti rimuovendo dal dataset di partenza gli elementi con Cloud3pm minore di 0. Il dataset con i valori negativi si chiamerà cloudP, l’altro cloudT.
Si alleni un regressore su cloudT per predire i valori di Cloud3pm. Si usi il modello per sostituire in cloudP il valore predetto per Cloud3pm. (punti 4)

5. Si consideri il dataset ottenuto concatenando cloudP e cloudT in un unico dataset e si confronti l’accuratezza che si ottiene con un 10 Fold cross validation e il classificatore basato su Decision Tree e quello basato su Logistic Regression con quella ottenuta al punto 3. (punti 3)

6. Utilizzare un algoritmo di regressione da applicarsi al dataset del punto 1 per effettuare la predizione. Arrotondare i valori predetti all’intero. Confrontare i risultati ottenuti con quelli ottenuti nei punti precedenti (punti 3).