Weather Dataset

Francesco Guerra
francesco.guerra@unimore.it

Big Data Analysis

Weather

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

Weather

1. Inserire nuove features nel dataset che rappresentino:
   - escursione termica gionarliera (MaxTemp-MinTemp)
   - differenza di unimidità (Humidity3pm- Humidity 9am)
   - vento medio (WindSpeed3pm,WindSpeed9am)

   Eliminare l'attributo Location e trasformare gli attributi booleani in numerici. Dividere il dataset in train e test (20%). Calcolare che accuratezza si ottiene con un modello randomForest (100 alberi) e confrontarla con un dummyClassifier (strategia stratified).

2. Verificare se i risultati migliorano dopo avere scalato i valori tra 0 e 1 e normalizzato con Normalizer

3. Nel dataset originale, il valore -1 per l'attributo Cloud3pm rappresenta un errore. Occorre individuare le istanze che assumono quel valore, usare il modello randomforest per stimare il valore di Cloud3pm. Sostituire il valore predetto a -1 e rieseguire tutte le normalizzazioni del punto precedente e valutare se si introduce un miglioramento. Valutare cosa succede se invece di effettuare la sostituzione dei valori si eliminano le istanze non corrette

Weather

4. Realizzare una pipeline che effettui il punto 2 dell'esercizio
5. Realizzare una pipeline che effettui i punti 1 e 2 dell'esercizio
6. Verificare se il modello ha risultati migliori utilizzando un randomforest con 10-100-250 alberi
7. Provare altre trasformazioni per migliorare il risultato.