Drug Review Dataset

Il **dataset** si trova su Kaggle - https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018 e contiene dati relativi ad alcuni farmaci, tra cui la **condizione** del paziente che ha usato il farmaco, una **recensione** testuale del farmaco, un **punteggio** assegnato dal paziente al farmaco e un valore di **utilità** della recensione assegnata da altri pazienti.

Il dataset è già diviso in train e test.

Le recensioni sono di lunghezza varia e possono essere analizzate per predire il punteggio e/o predire la condizione del paziente.

1. Analisi dei dataset di train e test

1. Quante righe e quante colonne nei dataset? Ci sono valori nulli? La lunghezza del testo nelle recensioni varia molto da recensione a recensione?
2. Considera la colonna condition, quanti sono i possibili valori di questa feature? Visualizza in un istogramma la distribuzione dei 10 valori più frequenti per entrambi i dataset.
3. Considerando le 100 recensioni con il valore di utilità più alto, calcolare il punteggio medio per condizione del paziente. Cosa si può osservare da questi dati? Si nota una differenza tra i due dataset?

2. Trasformazione del dataset e predizione del punteggio

1. A partire dal dataset originale, eliminare eventuali attributi inutili (giustificare la scelta), eliminare eventuali istanze che contengono valori nulli.
2. Creare una pipeline in cui:
    a. i valori di *drugName* e *condition* sono trasformati con un label encoder,
    b. il testo delle recensioni è trasformato utilizzando TfidfVectorizer,
    c. i valori di *usefulCount* sono scalati con MinMaxScaler(0,1),
    d. viene applicato il modello **LinearRegression**.
3. [OPTIONAL] Valutare con GridSearchCV la configurazione migliore per TfidfVectorizer (es. analyzer e ngram_range).

3. Trasformazione del dataset e predizione della condizione del paziente

1. A partire dal dataset originale, eliminare eventuali attributi inutili (giustificare la scelta), eliminare eventuali istanze che contengono valori nulli.
2. Creare una pipeline in cui:
   a. i valori di *drugName* sono trasformati con un label encoder,
   b. il testo delle recensioni è trasformato utilizzando TfidfVectorizer,
   c. i valori di *usefulCount* sono scalati con MinMaxScaler(0,1),
   d. viene applicato il modello **LinearSVC**.
3. [OPTIONAL] Valutare con GridSearchCV la configurazione migliore per TfidfVectorizer (es. analyzer e ngram_range) e per il modello LinearSVC.