BIG DATA AND TEXT ANALYSIS
10/01/2025

Nome:	Cognome:
Matricola:	

Parte 1	
Parte 2	
Totale	

Regole:

1. La prova dura 2 ore.
2. Durante la prova è vietato comunicare con gli altri e utilizzare funzionalità per il suggerimento del codice e strumenti come ChatGPT e Gemini.
3. Al termine della prova, lo studente rinomina il notebook con il proprio nome e cognome, lo scarica nel formato **.ipynb** e lo manda via email a **federica.rollo@unimore.it** con oggetto: **BDTA: 10-01-2025**.
4. I risultati saranno pubblicati entro il giorno 22/01/2025.

Parte 0: Il Dataset

Preso da kaggle - https://www.kaggle.com/datasets/beridzeg45/diamonds-prices-prediction/, il dataset è scaricabile attraverso il link http://bit.ly/4jaJgD6 e contiene dati relativi ad alcuni tipi di diamanti. Sono presenti le seguenti feature:

`Shape`: forma del diamante.
`Cut`: qualità del taglio del diamante (in ordine crescente: Very Good, Excellent, Ideal, Astor).
`Color`: grado di colore del diamante da D (incolore) a Z.
`Clarity`: grado di chiarezza basato sulle imperfezioni.
`Carat Weight`: peso del diamante in carati.
`Length/Width Ratio`: proporzione tra lunghezza e larghezza.
`Depth %`: profondità del diamante come percentuale della sua larghezza.
`Table %`: larghezza della facciata superiore in percentuale.
`Polish`: qualità della finitura superficiale del diamante.
`Symmetry`: precisione della forma del diamante.
`Girdle`: spessore del bordo del diamante.
`Culet`: dimensione della sfaccettatura inferiore.
`Length`: lunghezza del diamante in millimetri.
`Width`: larghezza del diamante in millimetri.
`Height`: altezza del diamante in millimetri.
`Price`: prezzo del diamante in dollari ($).
`Type`: tipo di diamante (target).
`Fluorescence`: livello di fluorescenza UV del diamante.

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 ogni attributo sono sempre correttamente specificati - non esistono

“missing values”)? _____ Il dataset è bilanciato per quanto riguarda la classe da predire? _____ (punti 1)

2. Calcolare in una nuova colonna il volume approssimato del diamante come `Length` * `Width` * `Height` e verificare con un opportuno grafico se c'è una relazione tra il volume e il prezzo: i diamanti più grandi sono quelli più costosi? (punti 2)

3. Considerare soltanto i record con valore non nullo di `Cut` e discretizzare la variabile `Carat Weight` in 5 gruppi. Verificare attraverso una tabella pivot se è vero che il prezzo medio aumenta all'aumentare della qualità del taglio e del peso in carati. (punti 3)

4. Si vuole analizzare il prezzo a carato per ogni tipo di diamante: creare una nuova feature che rappresenta il prezzo per carato (`Price` / `Carat Weight`) e visualizzare attraverso dei boxplot come varia questo prezzo per ogni tipo (`Type`) di diamante (punti 4)

Parte 2: Trasformazione e Predizione (20 punti)

1. Si vuole predire la tipologia di diamante (`Type`). Ricaricare il dataset originale, eliminare eventuali attributi inutili (giustificare la scelta), eliminare gli attributi con più del 50% di valori nulli, eliminare le istanze che contengono valori nulli, trasformare opportunamente valori categorici e dividere il dataset in modo che 3/4 degli elementi siano contenuti in un nuovo dataset “train” e 1/4 nel dataset “test” preservando le proporzioni delle classi nella colonna target.
Confrontare la predizione ottenuta sia sul dataset train sia sul dataset test dai classificatori DecisionTree, KNeighborsClassifier e da un dummy classifier a scelta. Effettuare alcune considerazioni sui risultati ottenuti, tenendo in considerazione i valori di F1 (con *average= “weighted”*) e della confusion matrix. (punti 4)

2. Confrontare i valori di F1 ottenuti nel punto precedente con quelli che si ottengono con una 10 Fold cross validation. (punti 1)

3. Attraverso la tecnica Permutation Feature Importance (PFI) e considerando il classificatore KNeighborsClassifier, analizzare la feature importance del dataset utilizzato al punto 1. Applicare 5 permutazioni per ogni feature. Quali risultano essere le 2 feature più importanti? (punti 4)

4. A partire dal dataset utilizzato al punto 1, trovare i valori migliori dei parametri `weights` e `n_neighbors` del classificatore KNeighborsClassifier. Come varia il valore di F1? (punti 2)

5. Creare una pipeline in cui, a partire dal dataset utilizzato al punto precedente, i valori degli attributi `Length`, `Width`, `Height` sono discretizzati in 5 intervalli, la variabile `Price` è scalata nell'intervallo 0-1 e tutti gli altri attributi sono lasciati invariati. Applicare il KNeighborsClassifier con i valori migliori dei parametri analizzati nel punto precedente e confrontare i risultati. (punti 3)

6. Creare una pipeline che, a partire dal dataset iniziale a cui sono stati rimossi gli attributi con più del 50% di valori nulli, trasforma le colonne testuali in valori numerici, applica il SimpleImputer per sostituire i valori nulli, trasforma tutte le feature attraverso lo Standard Scaler e applica il KNeighborsClassifier. (punti 3)

7. Aggiungere alla pipeline del punto precedente (dopo lo Standard Scaler) la decomposizione TruncatedSVD. Valutare il valore migliore per il numero di componenti di TruncatedSVD tra 2, 4 e 6 e i valori migliori di n_neighbors e weights del KNeighborsClassifier. (punti 3)