BIG DATA AND TEXT ANALYSIS
18/02/2025

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: 18-02-2025**.
4. I risultati saranno pubblicati entro il giorno 25/02/2025.

Parte 0: Il Dataset

Preso e modificato da Kaggle (https://www.kaggle.com/datasets/yasserh/loan-default-dataset), il dataset è scaricabile attraverso il link https://bit.ly/4gFUCwt e contiene dati relativi ad alcune richieste di prestito. Lo scopo è predire se il prestito è in stato di inadempimento. Sono presenti le seguenti feature:

`ID`: identificatore del richiedente
`year`: anno della richiesta di prestito
`Gender`: sesso del richiedente
`age`: età del richiedente
`approv_in_adv`: indica se il prestito è stato approvato in anticipo (pre) oppure no (nopre)
`loan_type`: tipo di prestito
`loan_purpose`: scopo del prestito
`open_credit`: indica se il richiedente ha conti a credito aperti (opc) oppure no (nopc)
`business_or_commercial`: indica se il prestito ha scopi commerciali (b/c) oppure no (nob/c)
`loan_amount`: importo del prestito
`rate_of_interest`: tasso d'interesse applicato al prestito
`term`: durata del prestito in mesi
`lump_sum_payment`: indica se alla fine del prestito è richiesto un pagamento forfettario
`property_value`: valore della proprietà per cui si chiede il prestito
`construction_type`: tipo di costruzione (sb - sito costruito, mh - fabbricato)
`occupancy_type`: tipo di occupazione (pr -residenza primaria, sr -res. second., ir -inv. immob.)
`total_units`: numero di unità del bene oggetto del finanziamento (1U, 2U, 3U, 4U)
`income`: reddito annuo del richiedente
`credit_type`: tipo di credito del richiedente
`Credit_Score`: credit score del richiedente
`co-applicant_credit_type`: tipo di credito del co-richiedente
`submission_of_application`: modalità di presentazione della richiesta
`Region`: regione geografica in cui si trova la proprietà (nord, sud, centro, nord-est)
`Status`: indica se il prestito è in stato di inadempimento (1) oppure no (0)

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 rispetto alla classe da predire? _____ (punti 1)

2. Calcolare nella nuova colonna "ratio" il rapporto tra l'importo del prestito (`loan_amount`) e il valore della proprietà (`property_value`). Calcolare il valore medio di questo rapporto in base al numero di unità della costruzione (`total_units`). Si può affermare che maggiore è il numero di unità, maggiore è il rapporto? (punti 2)

3. Verificare, soltanto per i richiedenti con più di 55 anni, attraverso una tabella pivot se è vero che chi ha un reddito annuo maggiore e un credit score maggiore ha più probabilità di essere in regola con i pagamenti (`status`). Per fare questa analisi discretizzare le feature `income` e `Credit_score` in 5 intervalli. (punti 3)

4. Si immagina che chi chiede un prestito per una proprietà che è la sua residenza primaria (`occupancy_type`) sia più attento ad essere in regola con i pagamenti. I dati confermano questa considerazione? Visualizzare in un opportuno grafico e soltanto per i prestiti relativi a residenze primarie, la correlazione tra l'importo del prestito e il valore della proprietà. Quali conclusioni si possono trarre? (punti 4)

Parte 2: Trasformazione e Predizione (20 punti)

1. Si vuole predire se il prestito è in stato di inadempimento oppure no. Ricaricare il dataset originale, eliminare eventuali attributi inutili (giustificare la scelta), eliminare le istanze che contengono valori nulli, trasformare opportunamente i valori categorici e dividere il dataset in train (3/4 del dataset) e test (1/4), preservando le proporzioni delle classi nella colonna target. Confrontare la predizione ottenuta sia sul dataset train sia sul dataset test dai classificatori ExtraTreeClassifier, KNeighborsClassifier e da un dummy classifier a scelta. Effettuare alcune considerazioni sui risultati ottenuti, tenendo in considerazione i valori di F1 e della confusion matrix. (punti 4)

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

3. Calcolare, nelle predizioni ottenute dai modelli dei classificatori ExtraTreeClassifier e KNeighborsClassifier, la probabilità di avere un prestito in stato di inadempimento se si è donna. La probabilità è la stessa per gli uomini? Riportare alcune considerazioni sul confronto dei due modelli. (punti 3)

4. Analizzare la correlazione tra le feature del dataset, creare un dataframe che contiene, oltre alla colonna target, le 2 feature più correlate positivamente al target e le 2 feature più correlate negativamente al target. La predizione dell' ExtraTreeClassifier migliora? (punti 4)

5. A partire dal dataset utilizzato al punto 1, trovare i valori migliori dei parametri `criterion` e `max_depth` del classificatore ExtraTreeClassifier. Come varia il valore di F1? (punti 2)

6. Creare una pipeline in cui, a partire dal dataset utilizzato al punto 1, i valori degli attributi `income` e `loan_amount` sono discretizzati in 7 intervalli, `Credit_Score` è scalato con StandardScaler e tutti gli altri attributi sono lasciati invariati. Applicare il modello ExtraTreeClassifier e confrontare i risultati. (punti 3)

7. Aggiungere alla pipeline precedente la funzione TruncatedSVD per ridurre la dimensionalità del dataset. Valutare i valori migliori di `n_components` di TruncatedSVD e dei parametri `criterion` e `max_depth` di ExtraTreeClassifier, confronta i risultati con i precedenti. (punti 3)