BIG DATA AND TEXT ANALYSIS
10/09/2024

Nome:	Cognome:
Matricola:	

Parte 1	
Parte 2	
Totale	

Regole:

1. E' vietato comunicare con altri durante la prova.
2. 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 alla docente **federica.rollo@unimore.it** con oggetto: **BDTA: 10-09-2024**.
3. I risultati saranno pubblicati entro il giorno 20/09/2024.

Note:
Durata della prova: 2 ore. Il file csv che si trova al link https://bit.ly/3ATiuxi.

Parte 0: Il Dataset

Il dataset (preso da kaggle - https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers) contiene dati relativi ai clienti di una banca.
La variabile da predire è `Exited` e indica se il cliente ha abbandonato la banca.

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. Dopo aver discretizzato l'attributo `Age` in 5 gruppi, verificare se è vero che i clienti più anziani hanno meno probabilità di abbandonare la banca rispetto ai più giovani. (2 punti) Considerando i clienti con più di 60 anni, verificare se è vero che maggiore è il saldo (`Balance`) e minore è la probabilità che il cliente abbandoni la banca. (2 punti)

3. Riportare in una pivot table la media di `CreditScore` raggruppando per uomini e donne (sulle righe) e i valori di salario stimato discretizzati in 5 gruppi (sulle colonne). Si può dire che i clienti con `CreditScore` più elevato sono quelli con il salario più alto? Si notano differenze tra uomini e donne? (punti 3)

4. Considerando soltanto i clienti che hanno una carta di credito e più di 100000 euro di credito, confrontare in un istogramma la distribuzione del `CreditScore` dei clienti francesi e di quelli spagnoli. Chi ha `CreditScore` maggiore? (punti 2)

Parte 2: Trasformazione e Predizione (20 punti)

1. Si vuole predire l'abbandono dei clienti della banca. Ricaricare il dataset originale, eliminare eventuali attributi inutili (giustificare la scelta), eliminare le eventuali 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.

Allenare il train con il modello DecisionTree e valutare l'accuratezza ottenuta sia sul dataset train sia sul dataset test. Confrontare i risultati ottenuti con quelli ottenuti con una predizione basata sul modello KNeighborsClassifier. Effettuare alcune considerazioni sui risultati ottenuti, tenendo in considerazione i valori di accuracy, F1 score, l'analisi della confusion matrix e la predizione effettuata da un dummy classifier a scelta. (punti 4)

2. Confrontare l'accuratezza ottenuta nel punto precedente con l'accuratezza che si ottiene con una 10 Fold cross validation. (punti 1)

3. Considerando i dati del test set e utilizzando il modello DecisionTree, la probabilità di predire l'abbandono del cliente della banca è la stessa per uomini e donne? (punti 2) Valutare se l'accuratezza della predizione negli uomini è la stessa ottenuta nelle donne. (punti 2) Come varia l'accuratezza se elimino l'attributo `Gender`? (punti 1)

4. A partire dal dataset iniziale (in cui sono stati eliminati eventuali attributi inutili ed eventuali istanze con valori nulli) aggiungere una nuova feature nel dataset con il valore di (`EstimatedSalary`*`Tenure`+`Balance`)/2. L'accuratezza del modello DecisionTree migliora? Come cambia l'accuratezza se i valori della nuova feature vengono discretizzati in 10 gruppi? (punti 2)

5. A partire dal dataset iniziale (in cui sono stati eliminati eventuali attributi inutili ed eventuali istanze con valori nulli) trovare i valori migliori dei parametri `criterion` e `max_depth` del classificatore DecisionTree. Come varia l'accuracy? (punti 2)

6. Creare una pipeline in cui gli attributi `Balance` e `EstimatedSalary` sono discretizzati in 6 intervalli, l'attributo `Tenure` è scalato nell'intervallo 0-1 e tutti gli altri attributi sono lasciati invariati. La pipeline deve applicare il modello DecisionTree con i parametri migliori trovati al punto 5. Valutare l'accuratezza della classificazione. (punti 3)

7. Creare una nuova pipeline che seleziona N componenti tra quelle ottenute dalla pipeline del punto 6 utilizzando la funzione TruncatedSVD e applica il modello DecisionTree. Attraverso la funzione di gridSearchCV, valutare il valore migliore per il numero di componenti (`n_components`) tra 2, 4 e 6 e il numero migliore di gruppi in cui discretizzare gli attributi `Balance` e `EstimatedSalary`. (punti 3)