BIG DATA ANALYSIS
10/01/2022

Nome:	Cognome:
Matricola:	

Parte 1	
Parte 2	
Totale	

Regole:

1. E' vietato comunicare con altri durante la prova.
2. [Per chi è online] Nel primo notebook occorre copiare e firmare la seguente dichiarazione: "Dichiaro che questo elaborato è frutto del mio personale lavoro, svolto in maniera individuale e autonoma".
3. [Per chi è online] Durante la prova la connessione con la piattaforma di comunicazione adottata. In caso vengano rilevati comportamenti anomali lo studente viene ammonito e eventualmente la prova annullata.
4. Al termine della prova, lo studente rinomina il notebook con il proprio nome e cognome e le manda via email al docente: francesco.guerra@unimore.it, oggetto: BDA: 10-1-2022.
5. I risultati sono pubblicati entro il giorno 16/1/2022.

Note:

Durata della prova: 2 ore. Il file csv che si trova al link
https://bit.ly/2022BDAfraud

Parte 0: Il Dataset

Il datataset (preso e modificato da kaggle -- https://www.kaggle.com/surekharamireddy/fraudulent-claim-on-cars-physical-damage) contiene dati relativi a frodi assicurative di auto. La variabile da predire è "fraud".

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. Analizzare la variabile che indica l'età del guidatore, e considerare solo i guidatori con età inferiore a 91 anni. Rappresentare con un istogramma la distribuzione dei valori. Raggruppare poi le età in gruppi, in questo modo: gruppo1 18-21; gruppo 2 22-25; gruppo3 26-30; gruppo 4 41-40; gruppo 5 41-50; gruppo 6 51-90, visualizzare la distribuzione delle età nei gruppi e indicare la percentuale di frodi nel gruppo. (punti 3)

3. Considerare il dataset originale e considerare la divisione in uomini e donne, e all'interno di ogni gruppo la divisione in under o over 40 (si includano anche le persone con quaranta anni in questo gruppo). Indicare a quale gruppo occorre fare maggiore attenzione perché è più facile avere una frode all'interno di esso (motivare la decisione) (punti 4)

4. Verificare con un opportuno diagramma se è vero che la distribuzione delle frodi aumenta all'aumentare del pagamento richiesto per l'indennizzo (attributo `claim_est_payout`) (punti 2)

Parte 2: Trasformazione e Predizione (20 punti)

1. Si vuole predire il valore di fraud sulla base degli attributi presenti nel dataset. Ricaricare il dataset originale, eliminare eventuali attributi inutili (giustificare la scelta), eliminare le istanze che contengono valori nulli, rendere tutti gli attributi numerici, e dividerlo in modo che 2/3 degli elementi siano contenuti in un nuovo dataset "train" e 1/3 nel dataset "test".

Allenare il train con il modello Decision Tree e valutare l'accuracy ottenuta calcolata sia sul dataset train sia sul dataset test. Confrontare i risultati ottenuti con quelli ottenuti con una predizione basata sul modello Logistic Regression (ignorare eventuali warning). Effettuare alcune considerazioni sui risultati ottenuti, tenendo in considerazione anche l'analisi della confusion matrix e la predizione effettuata da un dummy classifier. L'accuratezza è la metrica migliore per misurare la qualità del modello in questo scenario, o sarebbe opportuno utilizzare un'altra metrica? (punti 4)

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

3. Trovare i parametri migliori del classificatore Logistic Regression. Agire sui parametri penalty e C. Verificare se l'accuratezza che si ottiene con la nuova configurazione supera quella standard ottenuta al punto 1 (punti 4)

4. Introdurre una discretizzazione degli attributi `claim_est_payout` e `vehicle_price`, e utilizzare la funzione MaxAbsScaler per scalare i valori del dataset tra 0 e 1 e confrontare se l'accuratezza ottenuta con il Decision Tree Classifier e con la Logistic Regression migliora (punti 3).

5. Creare una pipeline in cui si aggiungano al dataset normalizzato due colonne che rappresentano i valori degli attributi `claim_est_payout` e `vehicle_price` discretizzati in 10 intervalli. Si valuti se l'accuratezza migliora utilizzando LogisticRegression come modello (punti 2).

6. Aggiungere alla pipeline la funzione SelectKBest (https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html?highlight=selectkbest#sklearn.feature_selection.SelectKBest). Utilizzare la funzione di gridSearchCV per selezionare il K migliore e anche gli intervalli migliori in cui discretizzare i valori di `claim_est_payout` e `vehicle_price` (punti 3).

7. Creare una nuova pipeline che applichi un simpleImputer (anziché la rimozione delle righe), al dataset iniziale. Si aggiunga questa pipeline a quella del punto 6 e si valuti la strategia migliore tra mean, median e most frequent (si decida una configurazione qualsiasi per gli altri parametri.(punti 3).