BIG DATA ANALYSIS
12/01/2021

Regole:

1. E' vietato comunicare con altri durante la prova.
2. 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. 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: 12-1-2021.
5. L'orale deve essere svolto entro l'inizio delle lezioni del secondo semestre. Per la prenotazione rivolgersi al docente via email.
6. I risultati sono pubblicati entro il giorno 16/1/2021.

Note:

Durata della prova: 2 ore. Il file csv che si trova al link
http://bit.ly/2021BDAGEN

Parte 0: Il Dataset

Il datataset (preso da kaggle -- https://www.kaggle.com/manishkc06/patient-treatment-classification?select=training_set.csv) contiene dati relativi a pazienti in cura in un ospedale, utilizzando le seguenti feature:

Name / Data Type / Value Sample/ Description
HAEMATOCRIT /Continuous /35.1 / Patient laboratory test result of haematocrit
HAEMOGLOBINS/Continuous/11.8 / Patient laboratory test result of haemoglobins
ERYTHROCYTE/Continuous/4.65 / Patient laboratory test result of erythrocyte
LEUCOCYTE /Continuous /6.3 / Patient laboratory test result of leucocyte
THROMBOCYTE/Continuous/310/ Patient laboratory test result of thrombocyte
MCH/Continuous /25.4/ Patient laboratory test result of MCH
MCHC/Continuous/33.6/ Patient laboratory test result of MCHC
MCV/Continuous /75.5/ Patient laboratory test result of MCV
AGE/Continuous/12/ Patient age
SEX/Nominal – Binary/F/ Patient gender
SOURCE/Nominal/ {1,0}/The class target 1.= in care patient, 0 = out care patient

La variabile da predire è SOURCE.

Parte 1: Analisi (8 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? _____ Sono presenti tutte le età da 1 a 99? _____ Le età sono rappresentate con frequenza simili? _____ (punti 1).

2. Dividere i valori assunti dalla variabile AGE in 10 gruppi. Verificare se per ogni gruppo sono presenti un numero simile di pazienti rispetto la classe da predire. Verificare inoltre la distribuzione della classe da predire rispetto al genere (SEX). (punti 2)

3. Verificare se è vero che le donne si ammalano meno degli uomini. Rappresentare graficamente se possibile quanto emerge dai dati.(punti 2)

4. Realizzare una pivot_table in cui rappresentare come si comporta la classe da predire rispetto i 10 gruppi di AGE (sulle righe), e il SEX (sulle colonne) (punti 3)

Parte 2: Trasformazione e Predizione (22 punti)

1. Si vuole predire il valore di SOURCE sulla base degli attributi presenti nel dataset. Ricaricare il dataset originale, rendere 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. (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 decision tree. Agire sui parametri criterion, max_features e min_samples_split. 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 AGE e `THROMBOCYTE`, 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 il valore di AGE sia discretizzato in 4 intervalli, il valore di `THROMBOCYTE` sia discretizzato in 10 intervalli e poi il dataset venga ricondotto a valori nell'intervallo (0,1) e normalizzato con la funzione Normalizer. Si applichi poi un modello DecisionTree. (punti 4) [Alternativa (punti 2): non applicare la discretizzazione]

6. Verificare se con un modello di regressione lineare (applicando eventualmente una approssimazione all'intero) si ottengono risultati migliori (punti 2)

7. Applicare una funzione per l'ottimizzazione dei parametri (sia al DecisionTree sia alla regressione lineare, su parametri a piacere o dell'algoritmo o della normalizzazione) e verificare se l'accuratezza migliora. (punti 2).

8. Creare una pipeline che aggiunga alle features della pipeline del punto 5, le feature che derivano dalla applicazione di una PCA (https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html mantenendo due dimensioni) e le feature che derivano dalla applicazione della funzione SelectKBest (https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html?highlight=selectkbest#sklearn.feature_selection.SelectKBest scegliendo K=2). (punti 2).