BIG DATA ANALYSIS
2/02/2021

Nome:	Cognome:
Matricola:	

Parte 1	
Parte 2	
Totale	

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: 2-2-2021.
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 9/2/2021.

Note:

Durata della prova: 2 ore. Il file csv si trova al link
https://bit.ly/2021BDA2

Parte 0: Il Dataset

Il dataset (preso da kaggle -- https://www.kaggle.com/andrewmvd/heart-failure-clinical-data?select=heart_failure_clinical_records_dataset.csv) contiene dati relativi a pazienti deceduti per attacco cardiaco:
- age: age of the patient (years)
- anaemia: decrease of red blood cells or hemoglobin (boolean)
- high blood pressure: if the patient has hypertension (boolean)
- creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
- diabetes: if the patient has diabetes (boolean)
- ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
- platelets: platelets in the blood (kiloplatelets/mL)
- sex: woman or man (binary)
- serum creatinine: level of serum creatinine in the blood (mg/dL)
- serum sodium: level of serum sodium in the blood (mEq/L)
- smoking: if the patient smokes or not (boolean)
- time: follow-up period (days)
- [target] death event: if the patient deceased during the follow-up period (boolean)

La variabile da predire è death event.

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? _____ I casi raccolti nel dataset sono equamente distribuiti per età? _____ (punti 1).

2. Verificare se è vero che ci sono meno decessi tra le donne (sex = 0). Rappresentare graficamente se possibile quanto emerge dai dati.(punti 2)

3. Realizzare una pivot_table in cui rappresentare la percentuale di decessi considerando la variabile age (sulle righe e suddivisa in 5 gruppi), la variabile sex e la variabile smoking (entrambe sulle colonne) (punti 3)

4. Verificare se è vero che generalmente le persone anemiche (anaemia==true) sono anche diabetiche (diabetes == true).(punti 2)

5. La frequenza dei decessi è uniforme nelle età considerate nel dataset? Mostrare l’analisi attraverso un opportuno grafico (punti 2)

Parte 2: Trasformazione e Predizione (20 punti)

1. Si vuole predire il valore di death event sulla base degli attributi presenti nel dataset. Dividere il dataset in modo che 2/3 degli elementi siano contenuti in un nuovo dataset “train” e 1/3 nel dataset “test”. Eliminare gli eventuali attributi che non concorrono alla predizione (identificatori se presenti o altri attributi, giustificare la scelta).

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 una 5 Fold cross validation. (punti 1)

3. Considerare il dataset originale, eliminare l’attributo time, scalare il valore degli attributi a un intervallo (0,1) e allenare sui dati un modello di LinearRegression (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression). Analizzare poi i coefficienti del modello e individuare i 5 attributi che in valore assoluto hanno il valore più elevato. Costruire un nuovo dataset composto unicamente di quei 5 attributi, e usare la tecnica 5 Fold cross validation per valutare se l’accuratezza del modello Decision Tree migliora. (punti 5)

4. Considerare il dataset originale, eliminare l’attributo time, e creare una pipeline in cui il valore degli attributi age e platelets sia discretizzato in 6 intervalli e gli attributi non booleani vengano ricondotti a valori nell’intervallo (0,1) e normalizzati con la funzione Normalizer. Si applichi poi un modello DecisionTree e si valuti l’accuratezza. (punti 4)

5. Applicare una funzione per l’ottimizzazione dei parametri (sia del modello di classificazione sia della pipeline) e verificare se l’accuratezza migliora. (punti 3).

6. Creare una pipeline che aggiunga alle features della pipeline del punto 4, 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 3).