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).

BIG DATA ANALYSIS
08/01/2020

Parte 0: Il Dataset

Il datataset trainMobile.csv (preso da kaggle -- https://www.kaggle.com/iabhishekofficial/mobile-price-classification) contiene dati relativi a telefoni cellulare, utilizzando le seguenti feature:

battery_power: Total energy a battery can store in one time measured in mAh
blue: Has bluetooth or not
clock_speed: speed at which microprocessor executes instructions
dual_sim: Has dual sim support or not
fc: Front Camera mega pixels
four_g: Has 4G or not
int_memory: Internal Memory in Gigabytes
m_dep: Mobile Depth in cm
mobile_wt: Weight of mobile phone
n_cores: Number of cores of processor
pc: Primary Camera mega pixels
px_height: Pixel Resolution Height
px_widthPixel Resolution Width
ram: Random Access Memory in Mega Bytes
sc_h: Screen Height of mobile in cm
sc_w: Screen Width of mobile in cm
talk_time: longest time that a single battery charge will last
three_g: Has 3G or not
touch_screen: Has touch screen or not
wifi: Has wifi or not
price_range: This is the target variable with value of 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).

Il dataset è costituito da attributi con valori numerici. La variabile da predire è price_range.

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 attributo sono sempre correttamente specificati – non esistono "missing values")? _____ Il dataset è bilanciato per quanto riguarda la classe da predire? _____ (punti 1).

2. La variabile sc_w assume valori discreti o continui? Analizzare la distribuzione dei valori e verificare se i telefoni costosi hanno mediamente una dimensione superiore di schermo. Verificare se eliminando gli elementi con sc_w uguale a 0 il risultato cambia. (punti 2)

3. E' vero che mediamente i telefoni meno costosi hanno anche una batteria meno potente? Realizzare 4 istogrammi (uno per ogni valore di price_range) che rappresentino la distribuzione dei valori di battery power per ogni categoria. (punti 3)

4. Verificare se tutti i telefoni che hanno il 4G hanno anche il 3G (punti 2)

5. Quanti sono i telefoni 4G che non hanno wifi e bluetooth? (punti 2)

Parte 2: Trasformazione e Predizione (20 punti)

1. Si vuole predire il valore di price_range sulla base degli attributi presenti nel dataset. Dividere il dataset in modo che 3/4 degli elementi siano contenuti in un nuovo dataset "train" e 1/4 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. Effettuare alcune considerazioni sui risultati ottenuti, tenendo in considerazione anche l'analisi della confusion matrix. (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. Utilizzare la funzione di gridSearchCV per trovare i parametri migliori del classificatore decision tree. Agire sui parametri criterion, max_features e min_samples_split. Vericare se l'accuratezza che si ottiene con la nuova configurazione supera quella standard ottenuta al punto 1 (punti 4)

4. 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 migliora (punti 3).

5. Discretizzare il valore di ram in 4 intervalli e verificare se l'accuratezza ottenuta con il Decision Tree Classifier migliora (punti 2).

6. Creare una pipeline in cui il valore di ram sia discretizzato in 4 intervalli, il valore di battery_power 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]

7. Si verifichi l'accuratezza ottenuta con il file test.csv. Controllare le colonne del file. I risultati corretti sono nel file class.csv. (punti 2).

Note:

Durata della prova: 2 ore. Creare una cartella esame e scaricare in essa il file csv che si trova al link
http://bit.ly/MB2020BDA

Creare nella cartella un jupyter notebook e rispondere nel file notebook alle domande. Indicare CHIARAMENTE nel notebook a quale domanda si sta dando una risposta.
Salvare frequentemente il file notebook creato attribuendogli il proprio nome-cognome.
Al termine della prova spedire a francesco.guerra@unimore.it il file della prova o il notebook direttamente o la versione html (file / download as / HTML).

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).

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).

BIG DATA ANALYSIS
26/01/2018

NOME	
COGNOME	
MATRICOLA	

Parte 0: Il Dataset

Il file bdastudents.csv (separatore ;) contiene una libera variazione del dataset "Student Perfomance" disponibile nell'UCI Machine Learning Repository
http://archive.ics.uci.edu/ml/datasets/Student+Performance
Il file rappresenta alcuni dati su studenti che frequentano 2 insegnamenti in 2 scuole diverse e il campo G3 la valutazione finale (da predire).
Lo schema del dataset è

School - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

sex - student's sex (binary: 'F' - female or 'M' - male)

age - student's age (numeric: from 15 to 22)

address - student's home address type (binary: 'U' - urban or 'R' - rural)

famsize - family size (binary: 0' - less or equal to 3 or 1 - greater than 3)

Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

guardian - student's guardian (nominal: 'mother', 'father' or 'other')

traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

failures - number of past class failures (numeric: n if 1<=n<3, else 4)

schoolsup - extra educational support (binary: yes or no)

famsup - family educational support (binary: yes or no)

paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

activities - extra-curricular activities (binary: yes or no)

higher - wants to take higher education (binary: yes or no)

internet - Internet access at home (binary: yes or no)

romantic - with a romantic relationship (binary: yes or no)

famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

freetime - free time after school (numeric: from 1 - very low to 5 - very high)

goout - going out with friends (numeric: from 1 - very low to 5 - very high)

Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

health - current health status (numeric: from 1 - very bad to 5 - very good)

absences - number of school absences (numeric: from 0 to 93)

G1 - first period grade (numeric: from 0 to 2)

G2 - second period grade (numeric: from 0 to 2)

G3 - final grade (numeric: from 0 to 2)

Note:

Durata della prova:2 ore. Dove possibile rispondere nel file notebook.
Creare una cartella esame e scaricare in essa il file csv che si trova al link
http://bit.ly/BDAgen2

Salvare frequentemente il file notebook creato attribuendogli il proprio nome-cognome.
Al termine della prova spedire a francesco.guerra@unimore.it il file della prova o il notebook direttamente o la versione html (file / download as / HTML).

Parte 1: Analisi (10 punti)

1. Caricare il dataset e denominarlo con una variabile chiamata “dataset”

2. Quante sono le istanze contenute nel dataset? _____ Il dataset è bilanciato rispetto alle scuole e ai generi degli studenti analizzati? (punti 1).

3. Creare un nuovo attributo “GRate” che misuri per ogni studente la differenza tra la valutazione ricevuta nel primo e nel secondo periodo (punti 2)

Realizzare un grafico che rappresenti per ogni età questa differenza.

_____

_____

4. Sono mediamente più bravi (attributo G3) i ragazzi o le ragazze? Esistono delle variazioni rilevanti nelle due scuole considerate? (punti 2)

_____

_____

_____

5. Tra i genitori degli studenti considerati, il livello di “educazione” maschile e femminile varia? Sono generalmente più scolarizzati i padri o le madri? Visualizzare poi un grafico che rappresenti il concetto (punti 2).

6. Indicare cosa visualizza l’istruzione

```
ds["G3"].groupby([ds["G3"],ds["address"]] ).count().plot()
```

Si tratta di una operazione significativa? (punti 3).

_____

_____

Parte 2: Trasformazione e Predizione (20 punti)

1. Scikit-learn utilizza un array numpy per effettuare le proprie predizioni. Gli elementi dell’array numpy devono essere di tipo numerico. Creare un dataset chiamato “numeric” che contiene solo le features numeriche.

Creare poi un nuovo dataset “reduced” dall’originale con le colonne G1 e G2 e un dataset “lessReduced” togliendo da numeric unicamente le colonne G1 e G2(punti 1).

2. Si vuole predire G3 sulla base degli altri attributi presenti nel dataset. Dividere i dataset numeric, lessReduced e reduced in modo che 2/3 degli elementi siano contenuti in un nuovo dataset “train” e 1/3 nel dataset “test” (punti 2).

Valutare l’accuracy ottenuta con il modello LogisticRegression su tutti i dataset
(from sklearn.linear_model import LogisticRegression)

Il valore di accuratezza ottenuto è pari a _____ .La confusion matrix presenta qualche valore significativo (punti 1)?

3. Che valore di accuratezza si ottiene con un 10 Fold cross validation e il modello basato su Decision Tree _____

E' più affidabile la valutazione fatta con la cross validation o quella fatta con una suddivisione arbitraria del dataset in due parti, training set e test set? Per quale motivo? (punti 2).

_____

_____

4. Considerare il dataset numeric. Considerare l'intervallo di valori assunto dall'attributo age e dividerlo in tre parti. Associare a ogni istanza il valore 0,1,2 a seconda del fatto che l'età sia nel primo, nel secondo o nel terzo intervallo. Eliminare l'attributo age originale, non discretizzato e calcolare l'accuratezza con il metodo 10 cross fold validation.

_____

Trasformare la feature discretizzata in 3 feature booleane, una per ogni valore discretizzato. Il valore assegnato sarà 1 nella colonna che rappresenta il valore in esame. 0 nelle altre colonne. Calcolare l'accuratezza con il metodo 10 cross fold validation (punti 4).

_____

5. Aggiungere al dataset "numeric" gli attributi Mjob e Fjob il cui valore categorico deve essere mappato utilizzando una formula di conversione a scelta. Confrontare il risultato ottenuto con quelli ottenuti in precedenza (punti 4).

_____

_____

6. Partendo dal dataset originale, costruire due dataset contenenti solo le feature numeriche. Uno che rappresenti la scuola GP e l'altro la scuola MS. Costruire due modelli di predizione utilizzando il decision tree. Uno per gli studenti GP e l'altro per gli studenti MS. Allenare entrambi i modelli utilizzando 2/3 delle rispettive istanze come training.
Fondere i test in un unico file.
Verificare l'accuratezza ottenuta dal test in entrambi i modelli. Quale funziona meglio?
(punti 3)

_____

_____

7. Utilizzare un algoritmo di regressione da applicarsi al dataset del punto 1 per predire "G3". Arrotondare i valori ottenuti all'intero. Confrontare i risultati ottenuti con quelli ottenuti in precedenza (punti 3).

_____

_____

BIG DATA ANALYSIS
28/01/2020

Nome:	Cognome:
Matricola:	

Parte 1	
Parte 2	
Totale	

Note:
Durata della prova: 2 ore. Il file csv che si trova al link http://bit.ly/wea_2020
Rispondere nel file notebook alle domande.
Salvare frequentemente il file notebook creato attribuendogli il proprio nome-cognome.
Al termine della prova spedire a francesco.guerra@unimore.it il file della prova o il notebook direttamente o la versione html (file / download as / HTML) (oggetto della mail BDA_GEN_2)

Parte 0: Il Dataset

Il datataset weather_train.csv (preso da kaggle -- https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather) contiene dati relativi a rilevazioni meteo registrate in città spagnole una volta al giorno secondo il seguente schema:
'dt_iso', 'city_name', 'temp', 'temp_min', 'temp_max', 'pressure', 'humidity', 'wind_speed', 'wind_deg', 'rain_1h', 'rain_3h', 'snow_3h', 'clouds_all', 'weather_id', 'weather_main', 'weather_description', 'weather_icon'
Il dataset è costituito da attributi con valori numerici e categorici.
L'obiettivo è quello di prevedere il tempo complessivo di una giornata (valore della feature 'weather_main') sulla base degli altri parametri.

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 attributo sono sempre correttamente specificati – non esistono "missing values")? _____ Il dataset è bilanciato per quanto riguarda la classe da predire? _____ (punti 1).

2. Le rilevazioni con pressione e umidità uguale a 0 sono irreali. Quante sono queste rilevazioni? Eliminarle dal dataset (punti 1)

3. Analizzare la temperatura massima rilevata. Valutare se la distribuzione dei valori assume un andamento simile a una gaussiana. Considerare poi le rilevazioni che si collocano all'interno del 5% delle temperature più alte. Le città sono equamente presenti in quella fascia di rilevazioni? Come è il tempo complessivo nei giorni in cui la temperatura massima è in quella fascia per ogni città? (punti 4)

4. Verificare se quando nevica la temperatura sia prossima alla temperatura di congelamento (NOTA: il dataset riporta i valori in Kelvin) (punti 2)

5. Confrontare l'escursione termica media (temp_max-temp_min) registrata nei giorni in cui nevica, con quella delle giornate che sono all'interno del 5% delle temperature più alte (punti 2)

Parte 2: Trasformazione e Predizione (20 punti)

1. Si vuole predire il valore di 'weather_main' 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 attributi ["dt_iso","city_name","weather_description", "weather_icon","weather_id", "clouds_all"]
Convertire l’attributo 'weather_main' in numerico in maniera opportuna.

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. Effettuare alcune considerazioni sui risultati ottenuti, tenendo in considerazione anche l’analisi della confusion matrix. (punti 4)

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

3. Utilizzare la funzione Normalizer per normalizzare i valori del dataset e confrontare se l’accuratezza ottenuta con il Decision Tree Classifier migliora (punti 3).

4. Creare una pipeline con trasformatori PCA (si scelgano 5 attributi) e poi Normalizer. Si usi come modello il Decision Tree Classifier (punti 2) [2 punti ulteriori se gli attributi della PCA sono aggiunti agli attributi del dataset]

5. Utilizzare la funzione di gridSearchCV sulla pipeline per modificare il numero di attributi selezionati dalla PCA e alcuni parametri a piacere del classificatore. Verificare se l’accuratezza che si ottiene con la nuova configurazione supera quella standard ottenuta al punto 1 (punti 4)

6. Si verifichi l’accuratezza ottenuta dalla pipeline del punto 4 con il file weather_test. I risultati corretti sono nel file class.csv. Controllare le features presenti nei dataset. (punti 2).

7. Si sperimenti una pipeline come quella del punto 4 dove al posto del classificatore si utilizzi un regressore lineare. Il risultato dovrà essere approssimato all’intero per il calcolo dell’accuratezza (punti 2).