Relazione progetto

Data Mining and Organization

Studente: Amatori Matteo – 7024736

Anno accademico: 2019-2020

Data understanding and pre-processing:

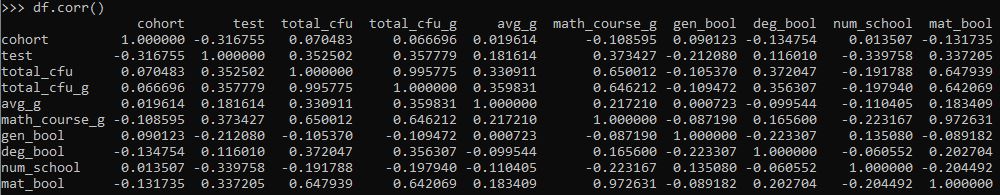
Durante questa fase del progetto ho analizzato il dataset, in primis aprendolo tramite un editor di testo per capirne il formato, la forma e, con l’ausilio della descrizione del dataset fornita, cosa vi fosse contenuto al suo interno.  
In seguito, tramite Python, ho provveduto ad effettuare le seguenti operazioni:

* Importazione del dataset all’interno di una dataframe della libreria Pandas di Python, tramite i comandi:

df = pd.read\_csv(dataset\_loc, sep = ";")

dove la variabile dataset\_loc contiene un riferimento relativo al file del dataset fornito;

* Creazione di una variabile classe, basata sulla variaible math\_course\_g che contiene “yes” nel caso in cui il valore della variabile di riferimento sia maggiore o uguale a 0, “no” altrimenti. Quest’operazione è stata effettuata nell’eventualità di utilizzare un metodo di classificazione anche su Weka, percorso inizialmente pensato, che poi ho deciso di non intraprendere perché ho continuato con il progetto in Python;
* Randomizzazione delle righe del dataset, in quanto nel caso in cui si decida di effettuare un’analisi con una divisione del dataset in training e test set è bene che la distribuzione delle righe sia quanto più casuale possibile;
* Aggiunta di alcune variabili, che consentirà di effettuare una visualizzazione della correlaizone tra le variabili del dataset:
  + gen\_bool: variabile booleana basata sul valore della variabile gender;
  + deg\_bool: variabile booleana basata sul valore della variabile degree
  + num\_school: variabile che trasforma ogni valore della variabile school in un valore numerico;
  + mat\_bool: una variabile booleana simile a quella citata nel punto 2, che contiene 1 nel caso in cui la variabile math\_course\_g sia maggiore di 0, 0 nel caso in cui essa sia 0.
* Analisi della correlazione tra le variabili del dataset: quest’operazione è effettuata tramite il comando di Python df.corr(). Il risultato ottenuto è il seguente:



Osservazioni sulla correlazione tra le variabili:

Da un’analisi sui valori contenuti nella matrice di correlazione tra le variabili presenti nel dataset e quelle aggiunte durante la fase di preprocessing, risulta che la variabile mat\_bool è correlata in maniera positiva con le variabili total\_cfu, test, num\_school e avg\_g.

Sono quindi queste le variabili da qui ho deciso di partire per effettuare una classificazione basata su Decision Tree in Python.