Homework Assignment 3

In this homework, you are expected to create a univariate feature selection schema using measures of impurity. We have learned two measures of impurity: entropy (***H(t, D)***), Gini Index (***Gini(t, D)***). These measures essentially show how random the distribution of a particular variable is or how well the values of a particular variable are separated. We also learned the Information Gain (IG) measure, which is a measure of the reduction in the overall entropy (or whichever measure of impurity is used).

In this homework, your task is to create a feature selection method called: ***IUFS*** (impurity-based univariate feature selection), which takes

- a dataset as (a pandas data frame, denoted as ***dataset***),
- the name of the target variable (denoted as ***target***),
- a measure of impurity (denoted as ***measure***) and
- the number of features to be selected (denoted as ***k***)

and returns the names (or indices) of the k columns to be selected based on the information gain.

Notes:

1. The input variable ***measure*** can only get 'entropy' or 'gini' as values. Otherwise, method terminates.
2. You can assume that your dataset consists of categorical variables.
3. You can find the definitions of measures of entropy and Gini index, as well as information gain in your textbook and slides. For entropy use base 2 for the logarithm.
4. You will be provided a starter code. You can use the car evaluation dataset provided in the starter code for your testing. Feel free to use other datasets with categorical variables (both target and descriptive features).

Bonus (+10 points):
Add a fifth parameter to IUFS method, called ***gain*** which can take 'IG' or 'GR' as values and will determine which selection measure to use between information gain (if gain=='IG') and gain ratio (if gain=='GR').