Homework 2

1. Description

The data set was extracted from census data of the United States in 1994, describing some social information about the citizens registered.

2. Data Format

(1) Age real [17.0, 90.0]
(2) Workclass {Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked}
(3) Education {Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool}
(4) Education-num real [1.0, 16.0]
(5) Marital-status {Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse}
(6) Occupation {Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces}
(7) Relationship {Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried}
(8) Race {White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black}
(9) Sex {Female, Male}
(10) Capital-gain real [0.0, 99999.0]
(11) Capital-loss real [0.0, 4356.0]
(12) Hours-per-week real [1.0, 99.0]
(13) Native-country {United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands}
(14) Class {>50K, <=50K}

3. Task

The task is to predict whether the citizen income exceeds fifty thousand dollars a year.

4. Grading

Each group is required to hand in a zip file that contains a report and a rule file before ***4/30***. The zip file should be entitled by the concatenation of the

homework number and the names of the team members. For example, “H2 李 XX 王 XX…” where “H1” and ”李 XX 王 XX…” denote homework 2 and the names respectively.

(1) Report: 35 %

The report should at least include four sections, namely introduction, preprocessing tactics, method & settings (maybe postprocessing) and results. The report length is **not more than 4 pages**.

(2) Implementation:

(1.1) Pre(post)processing Tactics: 35%

Since the original dataset is dirty, you may employ some proper data pre(post)processing tactics to facilitate the analysis, such as the tactics for discretization.

(1.2) Effectiveness Evaluation: 30%

Each group should additionally attach a file containing all the rules which are considered the most helpful in predicting whether the citizen income exceeds fifty thousand dollars a year. Basically, each rule is listed line by line and for two attributes in a rule should be separated by “;”. Every attribute is consisted of the attribute name and the associated values where each value is separated by “:”.

Ex1. age:>=:50;workclass:private;sex:male;class:>50K
Ex2. age:<:25:>:30;education:bachelors;hours-per-week:<=:40;class:<=50K

Ex1 classifies the income of a male who worked in a private company for over 50 years into “=>50K” whereas Ex2 assigns “<=50K” label to a guy between 25 and 30 years old who owned a bachelor degree and worked no more than 40 hours in a week. Note that for those testing instances which are not being able to be classified into any category are deemed incorrect.