Project #2

This assignment can be completed by team.
Total score: 100
Due date: 4/26

Problem description

This data set consists of information about adult people collected by the census bureau. We want to determine whether a person makes over $50,000 a year or not. More specifically, we want to know the attributes or a profile of those people who make over $50,000.

The attributes of the data set are as follows:

- Age
- Work class: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- Education level: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- Years of education
- Marital status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- 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.
- Relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
- Gender: Female, Male.
- Capital gain ($ amount) per year
- Capital loss ($ amount) per year
- Hours-per-week
- Income: >50K, <=50K.

Required activities

Conduct the classifications using various methods including Naïve Bayes, Decision tree induction, and Multilayer perceptron to answer the questions. What is the profile of person who likely to make over $50k? Write a brief report in Word format that includes the following:

(a) Your team name, member name(s), email addresses, and the percentage contribution to this assignment if the assignment was completed by a team. (If a team cannot reach a consensus on the individual contribution, include the individual's claimed percent contribution with a brief description on the specific tasks performed.)
(b) Choose either Python machine learning package or MATLAB (or both if you want). Give a brief description about the software or tool used for your data analysis.
(c) Perform the following tasks: **(i)** Conduct the classification using each of those methods specified above and come up with the best classifier from each method. **(ii)** Analyze the results of the classification from each method, comparing the performance. **(iii)** Based on your analysis, answer the question specified above in plain English (to people who don't have much

background on machine learning). **(iv)** Write a brief description about the process of your data analysis activities and detailed analysis results.
(d) The source code, scripts written, or major GUI snapshots taken from the tool used.

Warning: Although code reuse from source codes available on the Internet is allowed, copying code from another student or team in this class is strictly prohibited. Any student or team violating this policy will receive a **ZERO** score for this assignment.

What and How to submit this assignment

Turn in **your report**. If you have more than one file, include all the files, and zip/compress them into one file by **your (or team's) name**. Then submit the **zipped** or **compressed file** to **Titanium**. For example, if your team name is "ABC", then the zip file name should be **ABC.zip**. If the assignment was completed by team, only **ONE** of **your team members** needs to submit your team's work.

Grading policy

Your work will be graded based on the quality of your (or the team's) work as well as the completion of the requirements, the level of understanding on the problem and results, and the written report.