Project

The **dataset** is avalable at: `https://archive.ics.uci.edu/ml/datasets/Adult`

- Training data: adult.data
- Testing data: adult.test

Due: November 21, 2022 at 11:59 pm

Submit the following items on Canvas:

1) Project report in pdf format (maximum of 10 pages, excluding cover page, references, and appendix, if any).
2) R code:
    - Comment your code.
    - By clicking "run", your code should process data, run your method, and produce final prediction.

Project Description

We want to understand the relationship between a person's income and other various features such as age, job type, location, and education. To start, you can visit the above link and explore the dataset. You will need to process data, analyze data, and build predictive models to predict income for data points in the testing set. More details are given on the next page.

Write a report covering details of all steps of the project. The results have to be reproducible using your report and code. Carefully describe every assumption you make (if any), and every step in your report. If you are using any method, R package, or model we did not cover in class, make sure to describe what you are using and provide appropriate references.

1. Project Summary [5]

Provide a summary of your work in the beginning of the report. What have you done for this project? What is the aim of the project? Any interesting findings?

2. Data Understanding [20]

- Verify data quality. Are there any missing values, duplicate data, outliers, or invalid values? If there is any, provide justification for your detection. How do you handle them, and why do you handle them in such way?
- Describe the meaning and type of data (scale of measurement, values, etc.). What are the most important features and why?
- Give simple and appropriate statistics (e.g., range, mode, mean, median, variance, and frequency counts) for each of the important features. Can you find anything interesting about the data?
- Visualize some of the features. Present at least three meaningful visualizations, and provide interpretation for each of them.
- Explore the relationships between the features. You can use scatter plots, correlation, boxplots, cross-tabulation, or group-wise averages. Provide at least three relationships and interpret them.

3. Classification [15]

- The goal is to train models using the training set, and predict income for the testing set.
- Try at least three different methods.
- Compare the performance of different methods. If you observe one method performs better than the other, analyze the varied performance.

4. Prediction Outcome [5]

- Report your final model's classification error on the testing set. Your code should reproduce the result.

Additional/Creative Work [5]