Data Set Information:

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0))

Prediction task is to determine whether a person makes over 50K a year.

There are two data sets for this project.

1) train_test.csv : It is used to construct for classifying the person who makes over 50K a year.. There are 13 input variables and 1 output variable. The number of instances is 10,000. You can divide the data to training sample and test sample data.

2) submit: It is used to submit the predicted output variable(whether a person makes over 50K a year (binary: 'yes','no')) and the probability of 'yes' (or score). There are 13 input variables. The output variable is not provided. The number of instances is 2,000.

Attribute Information:

Output variable (desired target), y : yes: >50K, no: <=50K. (Binary variable)

1. **age:** continuous.
2. **workclass**: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. **fnlwgt**: continuous.
4. **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.
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**: continuous.
11. **capital-loss**: continuous.
12. **hours-per-week**: continuous.

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.

Relevant Papers:

Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996