Final Activity Outline

The goal of this activity is for students to integrate what they learned in the machine learning course, their intuition about the problem, as well as prior data analysis background, to predict whether a person’s income is greater than of less than 50K a year, based on characteristics about that person.

## About the Dataset

This dataset was taken from [UC Irvine’s Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/adult) and subsequently simplified. It contains data about ~30,000 adults in the US in 1994, and whether their income is above or below $50K/year. The data it contains is: **age (integer), workclass (string), education (string), marital status (string), occupation (string), race (string), sex (string), capital income (integer), hours per week worked (integer), native country (string)**. Below is extra information about the dataset:

* **workclass** contains the following values: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
* **education** contains the following values: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
* **marital status** contains the following values: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
* **occupation** contains the following values: 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
* **race** contains the following values: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
* **sex** contains the following values: Female, Male
* **capital income** refers to money you make/lose from your possessions (i.e. real estate or stocks you own). It can be positive or negative.
* **native country** contains the following values: 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

### Interacting with the Dataset

We have written Python code to allow you to easily interact with the database. Initialize it with the line “database = Database()”. You can then interact with the “database” variable as if it were a list (you can index into it, iterate over it in a loop, and get its len). Each entry in the database is a person object, and has the following properties:

* age
* workclass
* maritalStatus
* occupation
* race
* sex
* capitalIncome
* hoursPerWeek
* nativeCountry

For example, to access the 11th person’s age, do: “database[10].age”

Ask the mentors if you have any questions!

## Guidelines for Writing the Function

You can write your function in any way you want. You can use decision stumps/trees, weighted majority algorithm, and/or k nearest neighbors. You can also use your intuition (i.e. if someone works less than a certain number of hours per week, automatically predict that they will make less than 50K). You can use global variables if you want (for example, so you can store and modify the weights in weighted majority algorithm). However, make sure you end every global variable name and/or function name with your name, so that they don’t clash with other students’ variable names when we copy them into one document. For example, if my name is “Jon”, my function name should be “predictIncomeJon” and my global variable names should be “variableNameJon”.

Let us know if you have any questions!

## Evaluation Technique

We will run all your functions in an overall weighted majority algorithm. This will have two outcomes: (1) as a class we will create a stronger learner comprised of all of your learners; and (2) we will be able to determine how accurate each learner is.

Good luck!