DATA325 Project 2: Census Income

Submission: All materials need to be submitted into Moodle by March 10th at 11:59PM.

You may have up to one (1) partner.

The U.S. Constitution requires that a census be conducted every ten years in order to allocate congressional representatives. While the Constitution only requires an "actual enumeration" of citizens, the census has expanded to include a number of demographic questions. The U.S. Census Bureau is still compiling the data from the 2020 census. As described by the Census Bureau, the results of the 2020 census will,

"…determine congressional representation, inform hundreds of billions in federal funding, and provide data that will impact communities for the next decade."

In this project, you will use census data to predict whether or not someone has an annual income of more than $50,000. The data for making your predictions are contained in two files. The file "census_train.csv" contains 35,000 rows representing unique individuals, and 15 columns, representing demographic information about those individuals (including whether their income is above or below $50,000). The file "census_test.csv" contains 13,840 rows, but only 14 columns since the "income" column has been removed. A complete description of the variables in the data set is contained on the next page.

There are two deliverables for this project. The first is a short technical paper (not to exceed 1000 words) describing your modeling process. This should be a formal submission paper; there should be no typos, each graph or figure should have titles and axes, etc. The second is a ".csv" file containing a vector of your predictions for whether the individuals in the test set make more than $50,000. That is, you will create a length 13,480 vector of 0's and 1's and write them to a file with:

```
write.csv(prediction_vector, "my_predictions.csv", row.names = FALSE)
```

Your project will be evaluated for predictive quality (accuracy), writing quality, and mathematical clarity. Not all columns in this data set contain numerical values, so some will need to be translated into appropriate forms before beginning data analysis. There are also instances of missing or incomplete data, and some issues with how the data have been entered that you will need to address. You may wish to start by consulting the labs we have done in class and the textbook. You are also welcome to use any other techniques or packages you would like, but make sure that you can explain your analysis well.

Notes: This data set is great for practicing data science techniques. Because of this, if you search the internet you will find the original data set as well as articles that describe exactly how to apply statistical learning methods to it. There are several reasons ***NOT TO***

DO THIS. For one, it is cheating and a serious violation of the Wooster Ethic. But also, it ruins the fun. This project is a great chance for you to go more in depth about something we have discussed in class and to challenge yourself to make the best predictions possible! You are welcome to search the internet for general advice regarding predictive modeling, but if you encounter census data, look away! **If you have more specific questions about the data set, please ask me.**

Description of Variables:

`age`: continuous.

`workclass`: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

`fnlwgt`: continuous. A weight that represents how common people with these exact age and racial demographics are in the United States.

`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.

`education-num`: continuous. Numerical representation of education level.

`marital-status`: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. ("civ" and "AF" represent "civilian" (not in military) or "Armed Forces" (in military)).

`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.

`race`: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

`sex`: Female, Male.

`capital-gain`: continuous. (Income from the sale of a capital asset, e.g., stocks or property)

`capital-loss`: continuous. (A loss occurred when a capital asset, e.g., stocks or property, decreases in value.)

`hours-per-week`: continuous. Number of hours worked per week.

`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, Holland-Netherlands.

`income`: whether or not annual income from all sources is above or below $50,000