The “Adult” dataset hosted on (UCI’s Machine Learning Repository) contains approximately 32000 observations, with 15 variables. The dependent variable that in all cases we will be trying to predict is whether or not an “individual” has an income greater than $50,000 a year.

Here is the set of variables contained in the data.

- age – The age of the individual
- type_employer – The type of employer the individual has. Whether they are government, military, private, an d so on.
- fnlwgt – The \# of people the census takers believe that observation represents. We will be ignoring this variable
- education – The highest level of education achieved for that individual
- education_num – Highest level of education in numerical form
- marital – Marital status of the individual
- occupation – The occupation of the individual
- relationship – A bit more difficult to explain. Contains family relationship values like husband, father, and so on, but only contains one per observation. I’m not sure what this is supposed to represent
- race – descriptions of the individuals race. Black, White, Eskimo, and so on
- sex – Biological Sex
- capital_gain – Capital gains recorded
- capital_loss – Capital Losses recorded
- hr_per_week – Hours worked per week
- country – Country of origin for person
- income – Boolean Variable. Whether or not the person makes more than \$50,000 per annum income.