**Description:**

Our project is based on the “Adult Data Set” from the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/Adult>). The goal of the project is to “predict whether income exceeds $50K/yr based on census data”. We accomplished this goal by using a variety of different classification techniques in combination with this dataset.

This dataset was extracted from a 1994 census database, and the data is used to determine whether or not a person listed in the dataset falls into one of two categories (we later make these categories binary for simplicity sake). These two categories are income > $50k and income <= $50k. The dataset has a large number of attributes associated with each entry, some of which are not as useful as others. This dataset has 48842 instances, each instance as 14 (15 if you include the income) attributes, and there are missing values in some of the instances.

The dataset’s attributes are described in the table below. It includes the name of the attribute, a description of what the data represents, and the various values the data can take on.

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Data-type** |
| age | The age of the person | continuous |
| workclass | The workclass of the person | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked |
| fnlwgt | Weights on the current population survey | continuous |
| education | The person’s highest education level | 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 | An integer representing the person’s highest education level | continuous |
| martial-status | The person’s martial-status | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse |
| occupation | The category of the person’s 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 | The person’s current relationship status | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. |
| race | The race of the person | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| sex | The sex of the person. | Female, Male |
| capital-gain | The net capital-gain of the person | continuous |
| capital-loss | The net capital-loss of the person | continuous |
| hours-per-week | The number of hours the person works per week. | continuous |
| native-country | The native country of the person. | 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 |
| income | Whether or not the person makes more than 50k annually. | >50k, <=50k |

As you can see, each entry in this dataset as a very large number of potential attributes that it can contain and as a result some of the data is either not useful to us for the purposes of classification, or it is a bit redundant. In particular, the “fnlwgt “attribute is useless to us for the purposes of classification. It give no information in regard to classifying an instances annual income. Additionally, the “education” attribute is of little use to us as well – it contains a complicated number of string attributes that is better represented by the “education-num” attribute. The “education-num” attribute condenses all of that information down into an integer to represent the number of years of education. With this in mind, during our implementation, we remove these two attributes from our data during the pre-processing phase.