**Data selection**

The selected dataset for data mining process should in sufficient of instances and attributes to investigate the potential insight patterns. In addition to emphasis the comprehensive of pattern, the attributes should in high proportion in categorical type.

The dataset used for this project was obtained from UCI Machine Learning Repository website named “Adult Data Set”. The dataset and relative information can be find from the link: <https://archive.ics.uci.edu/ml/datasets/Adult>

US census dataset was initial extracted by Barry Becker from the 1994 census bureau database. It is important to notice that the extraction of dataset was based on the condition of following: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0)).

This dataset has been widely used in the Machine Learning studies especially in the experiment required large scale of data. “Adult Data Set” contains categorical attributes can effectively be used in the analysis of logistic regression (S. Sathiya Keerthi, 2002). In classifier aspect, “Adult Data Set” has been used to demonstrate the new algorithm work as expected to obtained better results than the algorithm without any wrapper function. (Jie Cheng, 1999) Moreover, “Adult Data Set” work as a large database was used to illustrated a hybrid classifier of decision-Tree and Naïve-Bayes are advanced than the function of individuals. (Ron, 2011) Therefore, “Adult Data Set” is a wise choice for experimenting in classification.

**Data description**

The data set is initially aimed to using in the prediction of the whether annual income exceeds $50000 depends on the attributes included within the census data. The dataset using in the analysis includes 32561 instances and 14 attributes. 5 attributes are numerical, and 9 attributes are in categorical. There is 2399 missing data existing in this dataset.

The full list of attributes shows in table below:

|  |  |
| --- | --- |
| attribute | Information |
| age | continuous |
| workclass | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked |
| fnlwgt | continuous |
| 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 |
| marital\_status | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. |
| 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 |
| capital\_loss | continuous |
| hours\_per\_week | continuous |
| 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. |

“fnlwgt” means final weight. Participators include within the dataset with similar demographic characteristics at similar “fnlwgt”.

**Preliminary analysis**

The first observation is in “native\_country”, majority data are concentrated in US, this is because the census was conducted in US, therefore, people who take the census was ancestral home in US. This result in this attribute does not take important role to influence the analysis outcome.

The other observation is the value of “captital\_gain” and “capital\_loss” are almost distributed in 0. In “workclass” most of them are in private. “relationship” and “sex” are highly correlatd with each other, for instance, husband in “relationship” is male in “sex”. The correlated between these attributes should be aware while in the process of analysis.

The pre-processing will deal with missing values and details will be discussed in the next section.