Objective of the project: To Create a KNIME workflow that utilizes the data preparation, cleaning and manipulation operations on the Adult training data set. This data set can be found at the UCI machine learning repository.

http://archive.ics.uci.edu/ml/datasets/Adult

This data set was developed by Barry Becker by extracting from the 1994 Census database.

Our prediction task is to determine whether a person makes over 50K a year.

Dataset Description:

Listing of attributes:

>50K, <=50K.

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.

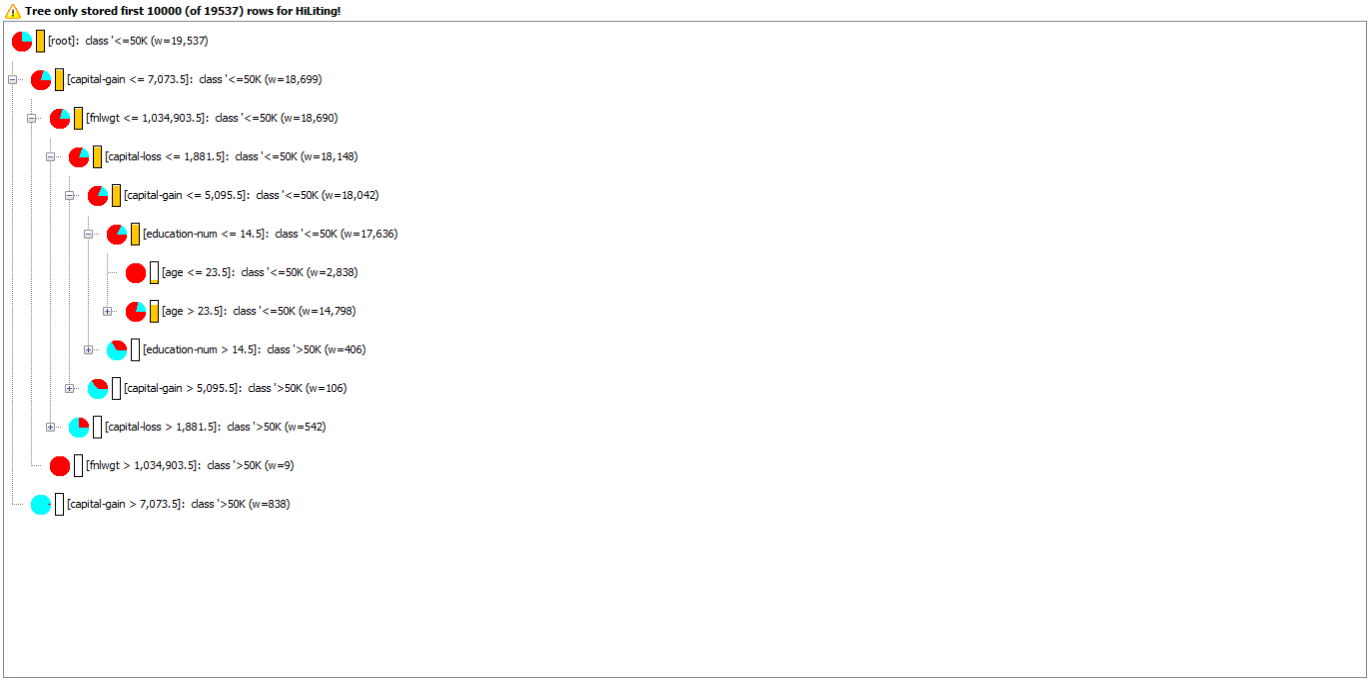
hours-per-week: continuous.

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, Trinidad & Tobago, Peru, Hong, Holland-Netherlands.

Data Mining Method used: Decision Tree Classification Algorithm.

Data Partitioning: 60 % training and 40 % test data.

Decision Tree splitting rules:



The splitting of the tree from the root node takes place based upon the capital-gains attribute of the adults and if we consider the above rule than we can make that if the capital-gains of an adult is greater than or equal to $7073.5 then we can predict that the income of the adult can be more than >$ 50 K and <= $ 50 K if less than $ 7073.5 of capital-gains.

From this, we can derive that we have even other attributes predicting the income scale of adults in US apart from capital-gains like education-num and age but maximum weightage can be given to capital-gains attribute.

Insight: Capital-gains is the major contributing factor for the level of the US adult to be greater or less $ 50,000.