In [1]:
```
# importing dependencies
import numpy as np
import pandas as pd
from IPython.display import display
```

In [2]:
```
# loading the data
columns_names = ["age", "workclass", "fnlwgt", "education", "education-num", \
                 "marital-status", "occupation", "relationship", "race", "sex", \
                 "capital-gain", "capital-loss", "hours-per-week", "native-country", "label"]

train_data = pd.read_csv('adult.data', header=None)
test_data = pd.read_csv('adult.test', header=None)
train_data.columns = columns_names
test_data.columns = columns_names
# display(train_data.head())
# display(test_data.head())
```

In [3]:
```
# removing unecessary columns
col_list = ['age', 'workclass', 'education-num', 'occupation', 'sex', 'label']
train_data = train_data[col_list]
test_data = test_data[col_list]
```

In [4]:
```
# removing missing data
train_data = train_data[train_data.workclass != ' ?']
train_data = train_data[train_data.occupation != ' ?']
test_data = test_data[test_data.workclass != ' ?']
test_data = test_data[test_data.occupation != ' ?']
```

In [5]:
```
# prepare datasets
def prepare_dataset(ds):
    dataset = ds

    # creating our custom train data DataFrame
    col_list = ['age', 'workclass', 'education-num', 'occupation', 'sex', 'label']
    dataset = dataset[col_list]


    # setting values
    workclass_state_1_values = [' Federal-gov', ' State-gov', ' Local-gov', ' Self-emp-inc']
    workclass_state_0_values = [' Never-worked', ' Private', ' Self-emp-not-inc', ' Without-pay']
    occupation_state_1_values = [' Exec-managerial', ' Prof-specialty', ' Protective-serv', ' Tech-support']
    occupation_state_0_values = [' ?', ' Adm-clerical', ' Armed-Forces', ' Craft-repair', ' Farming-fishing', \
                                 ' Handlers-cleaners', ' Machine-op-inspct', ' Other-service', ' Priv-house-serv',
 \
                                 ' Sales', ' Transport-moving']

    # discretizing age
    dataset.loc[dataset['age'] < 26, 'age'] = 0
    dataset.loc[dataset['age'] > 65, 'age'] = 0
    dataset.loc[dataset['age'] > 0, 'age'] = 1

    # discretizing sex
    dataset.loc[dataset['sex'] == ' Male', 'sex'] = 1
    dataset.loc[dataset['sex'] == ' Female', 'sex'] = 0

    # discretizing workclass
    for val in workclass_state_1_values:
        dataset.loc[dataset['workclass'] == val, 'workclass'] = 1
    for val in workclass_state_0_values:
        dataset.loc[dataset['workclass'] == val, 'workclass'] = 0

    # discretizing education-num
    dataset.loc[dataset['education-num'] < 10, 'education-num'] = 0
    dataset.loc[dataset['education-num'] >= 10, 'education-num'] = 1

    # discretizing occupation
    for val in occupation_state_1_values:
        dataset.loc[dataset['occupation'] == val, 'occupation'] = 1
    for val in occupation_state_0_values:
        dataset.loc[dataset['occupation'] == val, 'occupation'] = 0

    # discretizing labels
    dataset.loc[dataset['label'] == ' <=50K', 'label'] = 0
    dataset.loc[dataset['label'] == ' >50K', 'label'] = 1

    return dataset


train_data = prepare_dataset(train_data)
test_data = prepare_dataset(test_data)
# display(train_data)
# display(test_data)
```

In [6]:
```
# writing files
train_data.to_csv('/home/arthurcgusmao/my_train_data.dat', header=None, index=None, sep=',', mode='a')
test_data.to_csv('/home/arthurcgusmao/my_test_data.dat', header=None, index=None, sep=',', mode='a')
```