

## DataXpert@UiTM Siri ke-6 (Final Project)

### Predicting Yearly income

#### Columns Information:

**age** continuous

**workclass**

Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

**fnlwgt** final weight / count. Estimated no of people fall under this category.

**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** numeric representation of education column. Higher education level will have higher value.

**marital-status**

Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AFspouse.

**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**

Capital property includes depreciable property and any property which, if sold, would result in a capital gain (for example, cottages, buildings and securities such as mutual funds).

**capital-loss**

Capital property includes depreciable property and any property which, if sold, would result in a capital loss (for example, cottages, buildings and securities such as mutual funds).

**hours-per-week** working  
hours per week

**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, Holand-Netherlands

**income**

50K, <=50K

-----

-----

Answer the following questions:

- 1) Does working more hours lead to higher income?
- 2) Many people believe that the “A” students will end up working for “C” students. Do you think the statement is correct (based on this dataset)?
- 3) People with high-income level has a lot of investments?

+Develop and test your own hypothesis/questions to understand more about the characteristics of a person that has income >50K and <=50K per year.

**Develop a machine learning model to predict the yearly income.** Use 80:20 split for train and test data set.