Workshop 4

1. Which feature is the class target/dependent variable?

   workclass, education, education-num, occupation

2. How many missing values in this data set (List for each feature)?

Feature	adult	adult.test
workclass	1836	963
occupation	1843	966
native-country	583	274

3. How did you remedy the missing value and outliers (show steps)?
   i. Data Exploration by using python to see each feature have a null value the result shows there is no null value in both datasets.
   ii. Exploration in each feature containing a string value shows a missing value ("?") in feature workclass, occupation, and native-country in both datasets.
   iii. Remove a missing value from the dataset.
   iv. Normalization min-max scaling by using R program
   v. Use the Weka to find an outlier and extreme value by using “InterquartileRange” filter.
   vi. Remove an outlier and extreme value by using “RemoveWithValues” filter.
   vii. Check an outlier again do until the dataset doesn’t have any outlier and extreme value.
   viii. Applied one-hot encoding by using “NominalToBinary” filter.
   ix. Classifier IBK and get the result.
4. What data pre-processing (i.e., normalization, discretization) techniques you have use in for this data set.

   Normalization min-max scaling with age, fnlwgt, edu_num, hours_per_week

5. Use Weka software, convert you pre-processed data to .arff and use KNN algorithm or IBK for datamining tool.
6. What is the test accuracy did you get after you applied the adult.test (before pre-process and after pre-processing)?

Before pre-processing		After pre-processing	
adult	79.4202%	adult-pre	79.0462%
adult.test	79.5651%	adult.test-pre	79.0348%

Sonram Sirirat 640631037