## Data Mining: Assignment 2

## Census-Income

**Tasks:** Data Preparation, Classification

**Assigned:** September 11th,2018

**Due:** October 23rd,2018

**Points:** 75 points

For this project we will look again at census-income data. The cleaned dataset obtained from assignment 1 will be used as an input data set for this assignment.

**The training set:** The data set obtained from assignment 1.

**The test set:** The test data set (census-income-test.xlsx) (99,762 objects).

The followings are **classification tasks** you have to predict

1. Predict if a person's income is greater than 50,000.
2. Predict if a person's marital status is never married.
3. Predict if a white person's income is less than 50,000.

Write a report covering in detail all steps of the project. The results (rule set, decision tree, probabilities, confusion matrix, etc.) have to be reproducible using your report. Carefully describe every assumption and every step in your report.

**1. Data Preprocessing [30 points]**

1.1 Define and prepare your class attributes for each classification tasks if needed. [10 points]

1.2 Remove attributes that are not needed/useful for the analysis. Explain why you remove those attributes. [10 points]

1.3 Describe the final data set that is used for classification (include the scale/range for the new combined variables) [10 points]

**2. Modeling [45 points]**

2.1 Create at least 3 different classification models (different techniques) for each of the classification tasks. [15 points]

2.2 Implement models obtained from 2.1 on the test data set for each classification task. [15 points]

2.3 Compare classification results obtained from 2.2 for each classification task. [15 points]