961701 Everything starts with Data

Workshop 4 Data pre-processing with mix data type.

Due date 24th July 2021 before midnight 23:59pm

Description:

In this workshop, we are going to perform a data pre-processing task for the "census income" dataset, also known as the "Adult" dataset. The aim of this workshop is to get your familiar with the basic pre-processing (i.e., try to avoid using method from library to do the tasks). The output of this workshop is a KNN predictive mode. The model assists the stakeholders regarding the predict whether income exceeds $50K/yr based on census data.

Example of your preprocessing task are listed as:

1. Convert training and test data in to arff format. (you may use any tool to convert)
2. Identify the missing value and missing value imputation
3. Deal with outlier
4. Convert class independent and dependent variables to nominal or ordinal feature (e.g., 0, 1).
5. Applied one-hot and thermo encoding.

Etc.

Data source:

https://archive.ics.uci.edu/ml/machine-learning-databases/adult/

Training data = adult.data

Testing data = adult.test

Tool:

You can use any of your favorite tools (e.g., excel, python, java, etc.)

By the end of this workshop you should write one page discussing the following questions?

1. Which feature is the class target / dependent variables?
2. How many missing values in this data set? (list for each features)
3. How did you remedy the missing value and outliers (show steps)
4. What data pre-processing (i.e. normalization, discretization) techniques you have use in for this data set.
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)?

Workshop submission:

Please submit the following items to the google classroom.

1. One page describes of what you have done for the pre-processing stage in pdf format and how, such that you should address the pre-processing methods and the given questions. Please include your name, last name, and student id properly

2. The pre-processing dataset in csv format, please rename it to adult_pre.csv

3. Your IBK model from Weka software.