Individual Data Analytics Assignment in KNIME

Total points: 25

Due Date: June 20, 2021 – 23:00

Late submission penalty: 25% per day

Problem Context

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Dataset Content

The datasets consists of several medical predictor variables and one target variable, `Outcome`. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Tasks:

1. Using KNIME platform Examine Summary Statistics
2. Build a Decision Tree Workflow in KNIME
3. Create validation set: Split your dataset into two parts of train and test
4. Train and build a Decision Tree Classification model on your dataset
5. Evaluate the Performance of your Decision Tree Model by Generate a Confusion Matrix and Determine Accuracy Rate

What to submit:

1. Report (in a word document)
    a. Summary Statistics of your dataset
    b. Explain the validation set strategy you have used
    c. Validation: Confusion Matrix results for your trained decision tree model and its interpretation
2. KNIME Workflows file of your project

Note: For the step by step, instruction refer to the material on the Moodle about Classification using Decision Tree in KNIME.