# Classification I

## Learning Outcomes

---

- Upon successful completion of this lab, you will have demonstrated the abilities to:
  - Build decision tree model for different data sets
  - Evaluate model by holdout and cross-validation techniques
  - Investigate the overfitting issue for the decision tree model

## Instructions:

1. Read the tutorial **here** **(http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial6/tutorial6.html)**
2. Download the following data sets from the UCI Machine Learning Repository:
   - **https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)** **(https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data))**
   - **https://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+(Version+1)** **(https://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+(Version+1))**

**Part I:**

1. Build a decision tree model and evaluate the model using:
   1. **Holdout**
      A. Use 90% of data set for train and 10 % for the test, and perform it 5 times, the final results are the average of performance trials
      B. You should report the *Accuracy*, *Precision* and *F-measure* for each trial as well as their final average (use a table and then a bar chart)
   2. **Cross-validation**
      A. Perform 10-fold cross-validation for evaluating the model
      B. You should report the *Accuracy*, *Precision* and *F-measure* for each trial as well as their final average ((use a table and then a bar chart))

**Part II:**

1. Select the Entropy as the impurity measure and repeat **Part I**
2. Compare the final *Accuracy* of cross-validation of Part I and II using some figures

**Part III:**

1. Use the **holdout** method (train: 90 % data set, test: 10 % set)
   1. Investigate the effect of tree depth on the accuracy of the model (see the tutorial)
      A. Change the tree depth (e.g, 2, 5, 8, ..., 50) and draw training and test accuracy

B. Explain your observation

## Report:

1. Your report should have a cover letter including the group member names
2. Organize all your *diagrams* and *interpretations* in your lab report *(PDF format)*
3. Include your code and report in a folder (you can zip the folder) and submit it

## Resources:

1. **https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html (https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html)**
2. **https://scikit-learn.org/stable/modules/cross_validation.html (https://scikit-learn.org/stable/modules/cross_validation.html)** (alternatively, you can code the cross-validation by yourself)
3. **https://matplotlib.org/3.1.1/gallery/lines_bars_and_markers/barchart.html#sphx-glr-gallery-lines-bars-and-markers-barchart-py (https://matplotlib.org/3.1.1/gallery/lines_bars_and_markers/barchart.html#sphx-glr-gallery-lines-bars-and-markers-barchart-py)**