IT160IU Data Mining
Fall 2025

Assessment Task: Programming

Due date: Confirmed by the Lab instructor
Weighting: 20% of final mark

Objectives

This assessment task addresses the following objectives from the subject outline:

1. Understanding the process of data mining
2. Develop skills in implementing machine learning algorithms.
3. Develop skills in processing and transforming datasets.
4. Develop skills in evaluating the data mining methods.

Overview

In this assignment, you will build a data mining framework from scratch. This framework contains one *clustering/classification* method and one *sequence mining* method. The suggested dataset is one of the following resources:

- https://www.kaggle.com/datasets/podsyp/production-quality
- https://www.kaggle.com/datasets/abdullah0a/world-happiness-data-2024-explore-life
- https://www.kaggle.com/datasets/oktayrdeki/heart-disease

Given a certain training dataset following a specific data format, your framework should be built for a classification/prediction process. You need to identify the target label for classification/prediction. You will test your data mining framework with the real-world dataset to evaluate the quality of the processes.

In order to build this data mining framework, you need to finish four steps as follows.

- Step 1: Identify attributes for data mining, make training and testing datasets.
- Step 2: Implement a classification/prediction algorithm (Able to refer to the Weka library to find the best algorithm)
- Step 3: Improve the results in Step 2 by clustering, other algorithms, or analysing data.
- Step 4: Test the built models, compare and evaluate their performance. Write a report.

Step 1: Pre-processing (20 pts)

Input: a raw dataset
Output: clean data
Process:

- Cleaning and preparing data (15pts)
- Data analysis should be involved. (5pts)

Step 2: Implement a classification/prediction algorithm (20 pts)

In this step, you will implement a classification/prediction algorithm, using the Weka library. The input data needs to be converted to the ARFF format.

You have to use Weka in your code. Ref:

- https://waikato.github.io/weka-wiki/use_weka_in_your_java_code/
- https://www.youtube.com/playlist?list=PLea0WJq13cnBVfsPVNyRAus2NK-KhCuzJ

Step 3: Implement another algorithm (20 pts)

In this step, you will implement another algorithm for classification/prediction or another solution improving the above results.

Step 4: Model Evaluation and Report (20 pts)

Evaluate using 10-fold cross-validation.
You should evaluate the performance of the classification/prediction models and give some remarks on experimental results.

Hint: You should consider the run-time of building models and making predictions for performance evaluation as well.

Your report should include introduction, body (description of processing steps and evaluation), conclusions, and references.

Group collaboration and report presentation (20 pts)

Project Organization and Submission

Your project structure should look like this:

```
yourMemberIDs_Names/
        |--------- report.pdf
        |--------- code/
                |  -------- executable and program files + datasets
```

Your program should be able to generate binary files with corresponding classification method names, e.g., DECISIONTREE.

Double check before you submit:
Your program should be able to handle both absolute paths and relative paths to different training and test files. Do not assume that the files are located in the same folder of your programs.
Now you can zip your assignment folder and submit it on the IU Blackboard.
Congratulations, you just finished the programming assignment of IT160IU!