IS706 – Software Mining and Analysis
Programming Assignment 1 –
Data Mining for Software Engineering

Objective.

In this programming assignment, you will learn to use classification algorithms to solve a prediction problem. You would need to read one existing work and perform experiments following it. To perform experiments, you would need to use existing data mining tools. Dataset and references to tools needed for the lab would be given to you. At the end of the lab you would need to submit a document describing the results of your experiments.

Instructions.

Part I:

1. Please check the resource folder. The resource folder contains a paper [1], a read me file, and a dataset.
2. Please read at least the abstract and introduction of the paper to understand the problem that you are trying to solve. Please read the read me file to get an understanding of the dataset.
3. Please download SVM-light from the internet [2].
4. Please learn how to use SVM-light.
5. Divide the dataset into two. Use one half of the dataset as training dataset and the other half as testing dataset.
6. Train a model using the training dataset with SVM-light.
7. Test the model using the testing dataset.

8. Note/compute the precision, recall, and F-measure of the model on the testing dataset.

Part II:

1. Learn about ten-fold cross validation. Rather than dividing the data into two halves, use ten-fold cross validation to estimate precision, recall, and F-measure.
2. Learn about stratified ten-fold cross validation. Use stratified ten-fold cross validation to estimate precision, recall, and F-measure.
3. There are many classification algorithms that are available. Experiment with at least 3 classification algorithms (c.f., [3,4,5,6]) using ten-fold cross validation and stratified ten-fold cross validation. Note the precision, recall, and F-measure of these algorithms.

Submission.

Create a report for Part I and II above:

- For Part I, simply report the results.
- For Part II, provide descriptions of standard ten-fold cross validation, and stratified ten-fold cross validation. Also, report results obtained using ten-fold cross validation and stratified ten-fold cross validation with SVM-light and at least 3 other classification algorithms. Submit the report using eLearn.

Submit your report + the code that you have written to solve this programming assignment + a readme file describing how your code can be run as **one zip file**.

References.

1. Le, Tien-Duy B., David Lo, and Ferdian Thung. "Should I follow this fault localization tool's output? Automated prediction of fault localization effectiveness." Empirical Software Engineering 20.5 (2015): 12371274.
2. SVM-light. Available from: http://svmlight.joachims.org/
3. WEKA. Available from: http://www.cs.waikato.ac.nz/ml/weka/
4. R. Available from: http://cran.r-project.org/
5. PyTorch. Available from: https://pytorch.org/
6. Other machine learning or data mining frameworks / tools