



School of Innovation,  
Design and Engineering

## Module 2, Assignment 3 (1hp)

This document provides the instructions for the Module 2, Assignment 3 in DVA-448. The assignment, when completed, corresponds to 1 credit of the 2.5 credits in the course examination.

### Description:

We assessed the success of a predicted fault-proneness ranking of files by determining the percent of the actual bugs that were contained in the top 20% of files in the ranking.

An alternative to a ranking prediction is a *binary* classification, which predicts for each file whether or not it will be fault-prone. A prediction that a file will be fault-prone is a positive prediction; a prediction of not fault-prone is a negative prediction. The success of the predictions is evaluated by computing how close the predictions come to the true occurrences of faults in the files. The True Positives (TP) are the set of positive predictions for files that contain 1 or more bugs. The False Positives (FP) are the set of positive predictions for files that do not contain bugs. The True Negatives (TN) are the set of negative predictions for files that do not contain bugs. The False Negatives (FN) are the set of negative predictions for files that do contain bugs.

The ACCURACY of a prediction is defined to be the portion of all the predictions that turn out to be correct.

$$\text{ACCURACY} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN})$$

The PRECISION of a prediction is the portion of all the positive predictions that turn out to be correct.

$$\text{PRECISION} = \text{TP} / (\text{TP} + \text{FP})$$

If a prediction is perfect, it has no False predictions, so  $\text{FP} = \text{FN} = 0$ . In such a case,  $\text{Accuracy} = \text{Precision} = 1$

**Question 1:** Assume you have a system with 100 files, of which 80 are bug-free and 20 have bugs.

a) What is the impact on both Accuracy and Precision as the number of FP and FN increase?

b) If 10 files predicted to contain bugs are bug-free, and 10 files predicted to be bug-free actually contain bugs, what are the Accuracy and Precision of the prediction?

**Question 2:** How can the Accuracy & Precision concepts be applied for a file ranking prediction?

**Question 3:** False negatives and false positives are incorrect predictions. Do you think they represent equally serious mistakes? If not, why not? From a practical point of view, which is more of a problem? Recall, a false positive identifies a bug-free file as buggy, while false negatives identify buggy files as bug-free.

**Question 4:** If you know that there are likely to be many more bug-free files than buggy files, is there a strategy to assure a high Accuracy? A high Precision? What are the implications in terms of the usefulness of either of these measures of assessing a binary classification of this type?