**Problem**

Submitters often misclassify an improvement request as a bug and vice versa. Hence automated classification of the submitted reports would be of great practical utility.

bugs (meant for corrective maintenance of the software), it is common to change requests that ask for adaptation of the software to new platforms (adaptive maintenance) or to incorporate new features (perfective maintenance). Add to that requests to update the documentation or refactor the code, or simply discussions and requests for help.

**Algorithms used**

naive Bayes (NB), linear discriminant analysis (LDA), k-nearest neighbors (kNN), support vector machine (SVM) with various kernels, decision tree (DT) and random forest (RF) separately to classify the reports from three open-source projects.

What is F score (In paper performance is evaluated on F score)

In statistical analysis of binary classification, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It is calculated from the precision and recall of the test, where the precision is the number of correctly identified positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive.

Highest F-measure varies within 69 􀀀 76% while both highest average classication accuracy and highest weighted average F-measure values vary within 75􀀀83% depending on the project.

1. True Positive (TP): Number of reports correctly labeled to a class.

2. True Negative (TN): Number of reports correctly rejected from a class.

3. False Positive (FP): Number of reports incorrectly labeled to a class.

4. False Negative (FN): Number of reports incorrectly rejected from a class.

Precision: It is the ratio of the number of true positives to the total number of reports labeled by the classier as belonging to the positive class.

Recall: It is the ratio of the number of true positives to the total number of reports that actually belong to the positive class.

F-Measure or F1-score: It is the harmonic mean of precision and recall.

Accuracy: It measures how correctly the classier labeled the records.

**Data analyzed**

In an elaborate study involving more than 7000 issues spanning 5 projects, researchers found that 33:8% of all reports are misclassified.

**Result in paper**

Our experiments show that random forests perform best while SVM with certain kernels also achieve high performance.

**Datasets Sources**

like Bugzilla1, Gnats2 and Jira3 are commonly used for storing bug reports and tracking them till closure. The issue reports selected belong to three open-source projects HttpClient4, Lucene5 and Jackrabbit6.

**Labels used**

Bug/Not Bug. and tracking, in this paper we will group issues into only two types, BUG and NUG

where NUG includes all categories except BUG, i.e., problems that do not require corrective maintenance.

**Dataset Meaning**

The corpus is organized as a set of 5 csv files, each corresponding to a single project. Every row in a csv is a 3-tuple: (ID, CLASSIFIED, TYPE) referring respectively to the report identifer (that links to its full information in the repository), the report type that the researchers considered most appropriate and the type as mentioned in the repository. In many cases, CLASSIFIED and TYPE do not agree, pointing to a misclassication in the repository. For our study, we chooseonly the reports extracted from Jira. They belong to three prominent open-source projects from Apache, namely, HttpClient (that aims to extend and provide robust support to the base Http protocol; it is no longer being developed and has been replaced by the Apache HttpComponents9 project), Lucene (that is a high-performance, full-featured text search engine library coded in Java) and Jackrabbit (which is an open source content repository for the Java platform).

So we consider the summary part and ignore the longer description for automated analysis. A cursory look at the summaries of the

**Test/Train set**

Generation of training and test subsets: We use k-fold cross-validation to measure the predictive power of a classier. In order to generate k folds, k iterations are run. In each iteration, stratied sampling is used to generate a test subset. If there are N reports in a project, each of the rst k 􀀀 1 test subsets contain bN=kc in-stances while the k-th test subset holds N 􀀀 (k 􀀀 1) bN=kc instances. The test subsets do not overlap. Thus each instance is tested exactly once. In an iteration, all instances not in the test subset form the training subset.

Project Overflow

Diagram

Description automatically generated