Data Mining Techniques for Software Quality Prediction in Open Source Software

An Initial Assessment

Elisabetta Ronchieri, **Marco Canaparo**

July 12, 2018

INFN - CNAF, Bologna, Italy

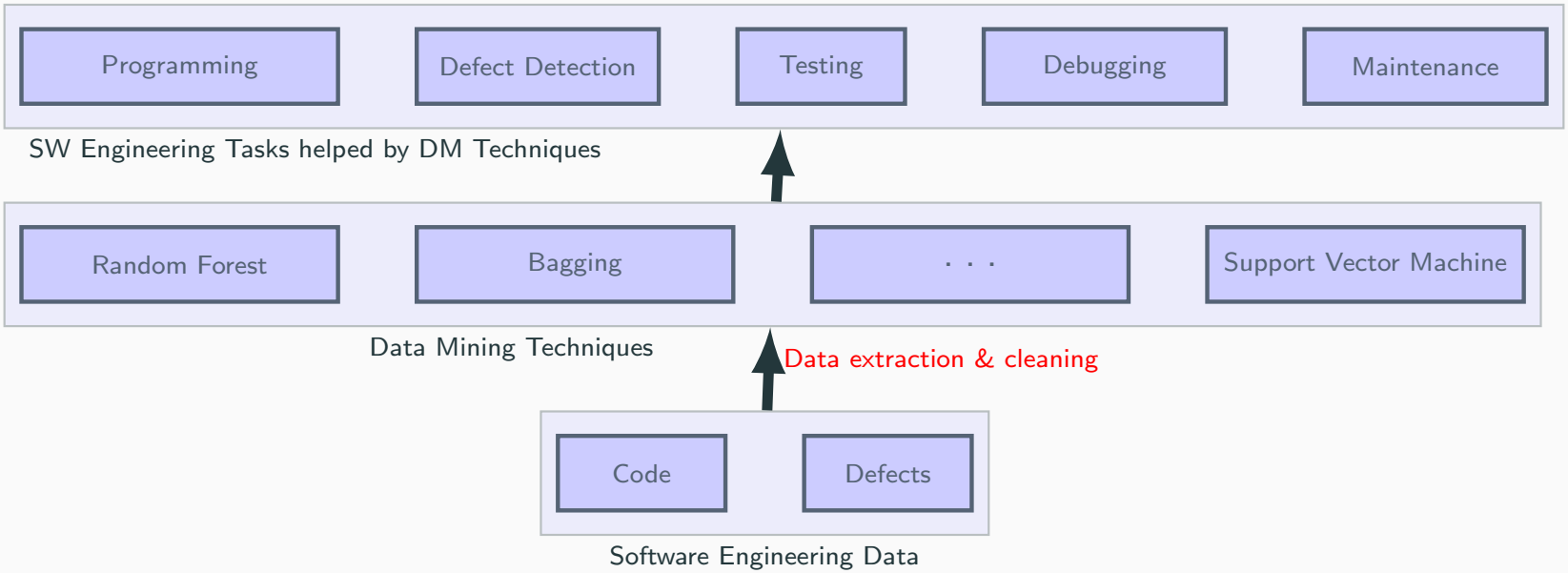
CHEP, 9-13 July, 2018, Sofia, Bulgaria

Table of contents

Motivation

Background

- **Open source** software widespread in scientific environment.
- Need to develop and use software with **high level of quality**.
- **Software quality prediction** to identify potential defect-prone software modules.



Why this Study?

In literature, there are **various studies** about software quality prediction and data mining techniques.

However, it does **not** provide **a comprehensive study** to evaluate and compare predictive techniques to be selected by quality managers.

We aim at providing an initial comparative performance analysis of different data mining techniques for software quality prediction **through a well-documented methodology**.

Research Methodology

Procedure: two Major Phases

Research in **data mining** for software engineering issues:

- Focusing on software quality prediction in order to identify defect-prone software modules

Selection of:

- Software Metrics to be used in our study
- Online datasets, with the necessary cleaning operation
- Data mining techniques
- Free data mining tools and packages
- Performance criteria

Study setup

Used Techniques and Metrics

We have collected **techniques and metrics** according to existing studies.

Data Mining Techniques:

- Support Vector Machine (e.g. SMO)
- Decision Tree (e.g. J48)
- Naive Bayes
- Ensemble Classifier (e.g. Random Forest)
- Deep Learning
- ...

Metrics:

- McCabe (e.g. Cyclomatic Complexity, Essential Complexity)
- Halstead (e.g. Base Measures, Derived Measures)
- Size (e.g. Lines of Code, Comments Density)
- Chidamber and Kemerer (e.g. Number of Children, Depth of Inheritance)
- ...

Selected Free Online Datasets

We have considered **not only** the NASA Defect Datasets.

Repository	#Projects	#Metrics	#Modules	%Defective Modules
NASA Defect Datasets[1,2]	11	[30,41]	[101, 5589]	[0.41, 48.80]%
Eclipse Datasets[3]	5	17 each	[324, 1863]	[9.26, 39.81]%
Android Datasets[4]	6	102 each	[74, 124]	[0, 27.02]%
Elastic Search Datasets[4]	12	102 each	[1860, 7263]	[0.16, 11.47]%

[1] https://github.com/klainfo/NASADefectDataset

[2] http://openscience.us/repo

[3] http://bug.inf.usi.ch/

[4] http://www.inf.u-szeged.hu/∼ferenc/papers/GitHubBugDataSet

Used Performance Criteria

Accuracy: percentage of modules correctly classified as either faulty or non-faulty $\frac{TP+TN}{N}$

Precision: percentage of modules classified as faulty that are actually faulty $\frac{TP}{FP+TP}$

Recall or Completeness: percentage of faulty modules that are predicted as faulty $\frac{TP}{TP+FN}$

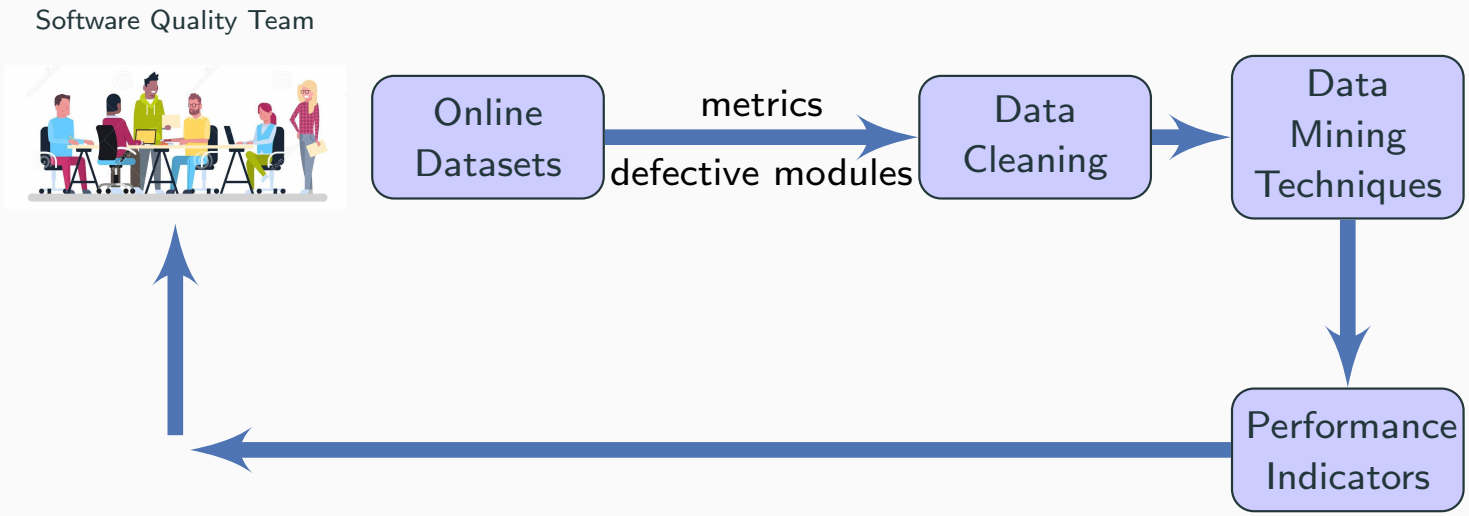
Mean Absolute Error: determines how close the values of predicted and actual fault rate differ

F-measure: $\frac{(2xRecallxPrecision)}{Precision+Recall}$

Each measure can be defined on the basis of the confusion matrix below

		Predicted by model +	**Predicted by model** -
Actual value	+	True Positive (TP)	False Negative (FN)
	-	False Positive (FP)	True Negative (TN)

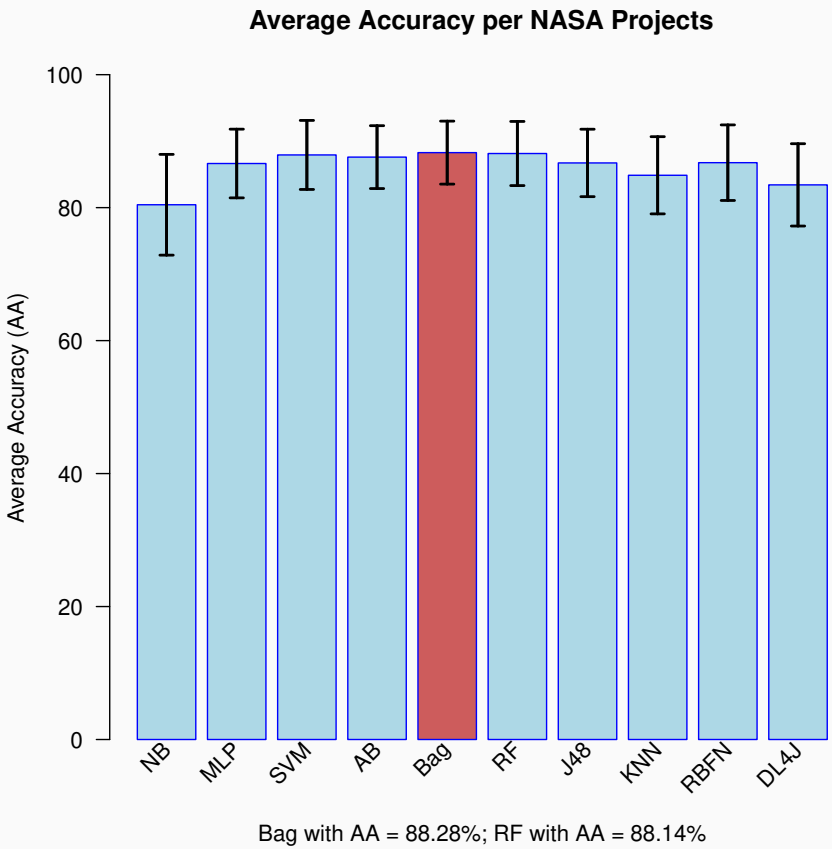
Solution Workflow



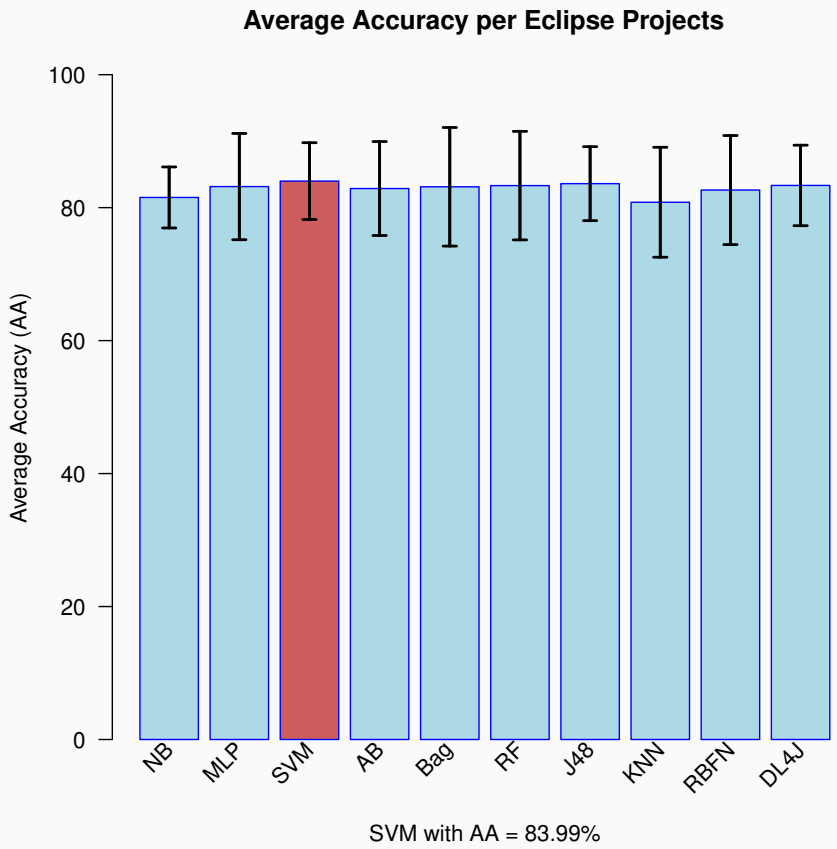
Replaced metrics' undefined values with their mean value.

Initial Assessment

Preliminary Results

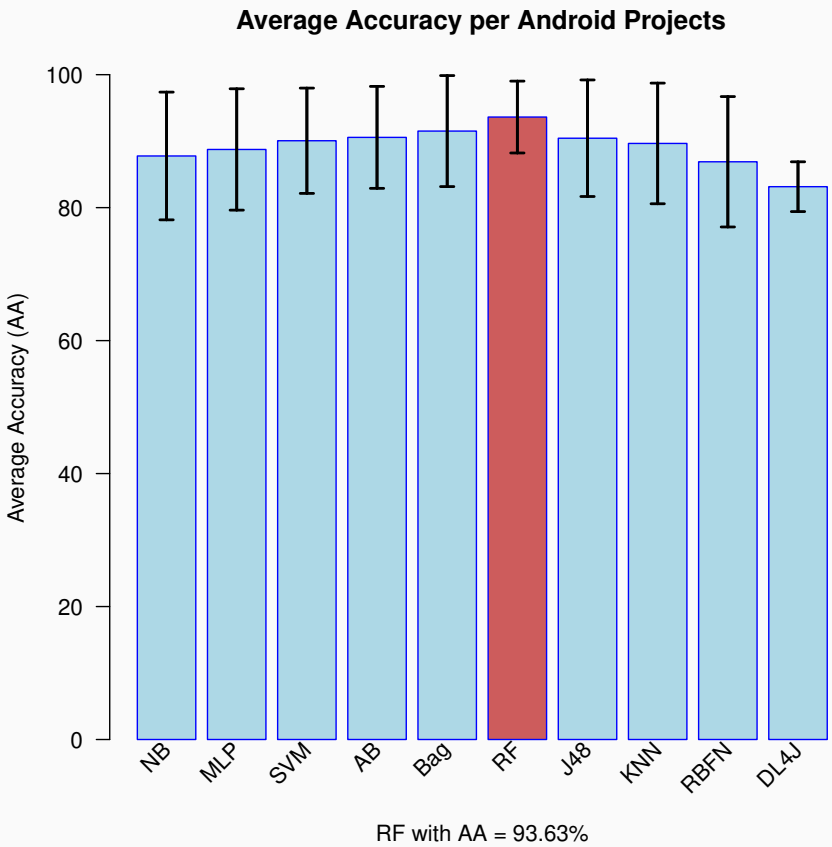


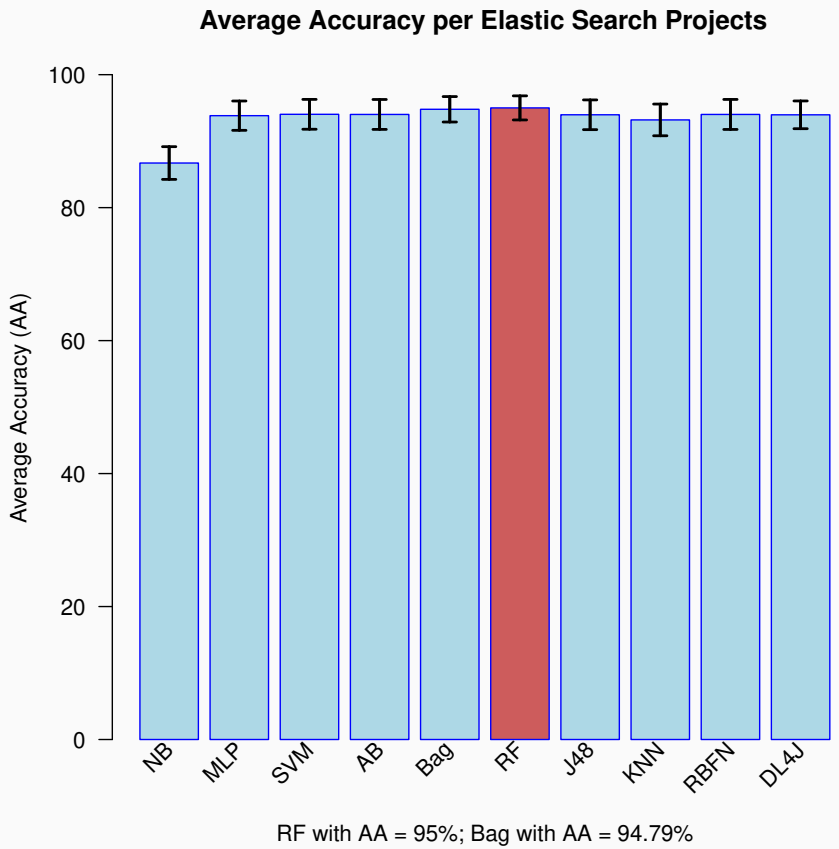
Bag with AA = 88.28%; RF with AA = 88.14%



SVM with AA = 83.99%

Preliminary Results





Conclusions

- We have shown an initial comparison of data mining techniques in the context of software defect prediction. We have leveraged existing literature to collect online dataset, used techniques, performance criteria.
- Best average accuracy amongst all the datasets: **Ensemble Learning**
- Experimenting the same techniques on software used in HEP, results will be shown in the IEEE 2018 Nuclear Science Symposium.
- Study is on going to identify which metrics have more weight in determining defect proneness.
- Data Mining can be used together with traditional statistical analysis techniques.

Questions?

Glossary

- Software quality, according to IEEE: Quality is the degree to which a system meets specified requirements or customer or users needs or expectations ISO: Quality is the degree to which a set of inherent characteristics fulfils requirements. Quality is the conformance to requirements and fitness for use. One way to monitor software quality is to find software faults or defects and then correct those faults.
- Data Mining is the process of discovering interesting patterns and knowledge from large amount of data contained in datasets.
- Defect: An imperfection or deficiency in a work product where that work product does not meet its requirements or specifications and needs to be repaired or replaced.

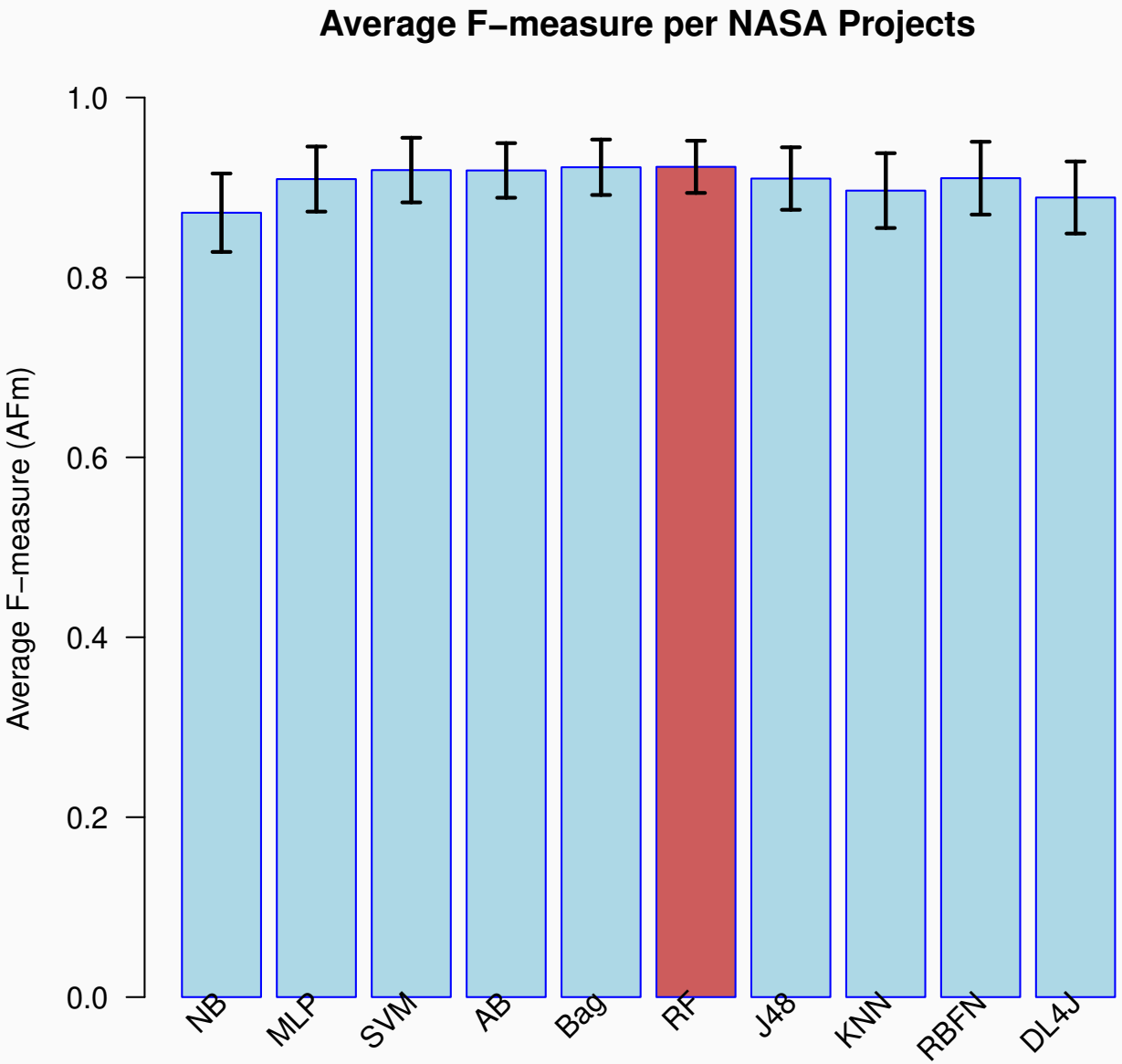
Selected Free Online Datasets: NASA Defect Raw data

Projects	Metrics	Modules	%Defective Modules
AR1[2]	30	121	7.44%
AR6[2]	30	101	14.85%
CM1[1]	41	505	9.50%
KC3[1]	41	458	9.39%
KC4[1]	41	125	48.80%
MC2[1]	41	161	32.30%
MW1[1]	41	403	7.68%
PC1[1]	41	1107	6.86%
PC2[1]	41	5589	0.41%
PC3[1]	41	1563	10.24%
PC4[1]	41	1458	12.21%

[1] https://github.com/klainfo/NASADefectDataset

[2] http://openscience.us/repo

Results



RF with AFm = 0.93; Bag with AFm = 0.92

Selected Free Online Datasets: Eclipse and Android

Projects	Metrics	Modules	%Defective Modules
EclipseJDTCore[1]	17	997	20.66%
EclipsePDE[1]	17	1497	13.96%
Equinox[1]	17	324	39.81%
Lucene[1]	17	691	9.26%
Mylyn[1]	17	1863	13.20%
Android_Uni_Class_2013_1[2]	102	74	27.02%
Android_Uni_Class_2013_2[2]	102	99	17.17%
Android_Uni_Class_2013_3[2]	102	110	1.81%
Android_Uni_Class_2014_1[2]	102	116	6.03%
Android_Uni_Class_2014_2[2]	102	119	1.68%
Android_Uni_Class_2015_1[2]	102	124	0%

[1] http://bug.inf.usi.ch/

[2] http://www.inf.u-szeged.hu/∼ferenc/papers/GitHubBugDataSet

Selected Free Online Datasets: Elastic Search

Projects	Metrics	Modules	%Defective Modules
Elastic_Search_Class_2010_1	102	1860	0.16%
Elastic_Search_Class_2010_2	102	2379	4.20%
Elastic_Search_Class_2011_1	102	3083	9.21%
Elastic_Search_Class_2011_2	102	3734	4.15%
Elastic_Search_Class_2012_1	102	3712	2.45%
Elastic_Search_Class_2012_2	102	3979	3.90%
Elastic_Search_Class_2013_1	102	4189	9.95%
Elastic_Search_Class_2013_2	102	4741	9.98%
Elastic_Search_Class_2014_1	102	5908	11.47%
Elastic_Search_Class_2014_2	102	6798	9.53%
Elastic_Search_Class_2015_1	102	6917	5.86%
Elastic_Search_Class_2015_2	102	7263	0.90%

http://www.inf.u-szeged.hu/∼ferenc/papers/GitHubBugDataSet

Used Free Data Mining Tools

- Open-source tool
- Various implementation of machine and deep learning algorithms
- Mainly used to manipulate datasets and graphical representation

- Open-source data mining platform
- Large number of implemented data mining techniques
- Only Used to run techniques

- Free Python package for data mining and data analysis
- Built on top of numpy, scipy and matplotlib
- Mainly used to manipulate datasets and run techniques

.....

```
 [76] "hda"                  "hdda"                 "hdrda"
 [79] "HYFIS"                "icr"                  "J48"
 [82] "JRip"                 "kernelpls"            "kknn"
 [85] "knn"                  "krlsPoly"             "krlsRadial"
 [88] "lars"                 "lars2"                "lasso"
 [91] "lda"                  "lda2"                 "leapBackward"
 [94] "leapForward"          "leapSeq"              "Linda"
 [97] "lm"                   "lmStepAIC"            "LMT"
[100] "loclda"               "logicBag"             "LogitBoost"
[103] "logreg"               "lssvmLinear"          "lssvmPoly"
[106] "lssvmRadial"          "lvq"                  "M5"
[109] "M5Rules"              "manb"                 "mda"
[112] "Mlda"                 "mlp"                  "mlpKerasDecay"
[115] "mlpKerasDecayCost"    "mlpKerasDropout"      "mlpKerasDropoutCost"
[118] "mlpML"                "mlpSGD"               "mlpWeightDecay"
[121] "mlpWeightDecayML"     "monmlp"               "msaenet"
[124] "multinom"             "mxnet"                "mxnetAdam"
[127] "naive_bayes"          "nb"                   "nbDiscrete"
[130] "nbSearch"             "neuralnet"            "nnet"
[133] "nnls"                 "nodeHarvest"          "null"
[136] "OneR"                 "ordinalNet"           "ORFlog"
[139] "ORFpls"               "ORFridge"             "ORFsvm"
[142] "ownn"                 "pam"                  "parRF"
[145] "PART"                 "partDSA"              "pcaNNet"
[148] "pcr"                  "pda"                  "pda2"
[151] "penalized"            "PenalizedLDA"         "plr"
```