

**Advance Software Engineering (SE-406)**

**LAB A1-G3**

**Laboratory Manual**



**Department of Software Engineering**

**DELHI TECHNOLOGICAL UNIVERSITY(DTU)**

Shahbad Daulatpur, Bawana Road, Delhi-110042

**Submitted to: -**

Ms. Parul Sharma

**Submitted by:-**

Name: ASHISH KUMAR

Roll number: 2K18/SE/041

# **EXPERIMENT 5**

- ASHISH KUMAR

- 2K18/SE/041

**Aim:-** Data Acquisition for software defect prediction.

## **About dataset:-**

Link to dataset: <http://promise.site.uottawa.ca/SERepository/datasets/cm1.arff>

This is a PROMISE data set made publicly available in order to encourage repeatable, verifiable, refutable, and/or improvable predictive models of software engineering.

### 1. Sources:

- Creators: NASA, then the NASA Metrics Data Program,
- <http://mdp.ivv.nasa.gov>. Contacts: Mike Chapman, Galaxy Global Corporation
- Donor: Tim Menzies (tim@barmag.net)
- Date: December 2 2004

### 2. Number of instances: 10885

### 3. Number of attributes: 22 (5 different lines of code measure, 3 McCabe metrics, 4 base Halstead measures, 8 derived Halstead measures, a branch-count, and 1 goal field)

### 4. Attribute Information:

1. loc : numeric % McCabe's line count of code
2. v(g) : numeric % McCabe "cyclomatic complexity"
3. ev(g) : numeric % McCabe "essential complexity"
4. iv(g) : numeric % McCabe "design complexity"
5. n : numeric % Halstead total operators + operands
6. v : numeric % Halstead "volume"
7. l : numeric % Halstead "program length"
8. d : numeric % Halstead "difficulty"
9. i : numeric % Halstead "intelligence"
10. e : numeric % Halstead "effort"
11. b : numeric % Halstead
12. t : numeric % Halstead's time estimator
13. lOCODE : numeric % Halstead's line count
14. lOCOMMENT : numeric % Halstead's count of lines of comments
15. lOBLANK : numeric % Halstead's count of blank lines
16. lOCODEANDCOMMENT : numeric

- 17. uniq\_Op : numeric % unique operators
- 18. uniq\_Opnd : numeric % unique operands
- 19. total\_Op : numeric % total operators
- 20. total\_Opnd : numeric % total operands
- 21: branchCount : numeric % of the flow graph
- 22. defects : { false,true} % module has/has not one or more reported defects

- 5. Missing attributes: none
- 6. Class Distribution: the class value (defects) is discrete  
false: 2106 = 19.35%  
true: 8779 = 80.65%

Software Defect Prediction Data Analysis

Notebook Data Logs Comments (3)

▲ 13

Copy & Edit 166

⋮

```
total_op      10885 non-null object
total_opnd    10885 non-null object
branchCount   10885 non-null object
defects       10885 non-null bool
dtypes: bool(1), float64(12), int64(4), object(5)
memory usage: 1.8+ MB
```

In [4]:

```
data.head() #shows first 5 rows
```

Out[4]:

	loc	v(g)	ev(g)	iv(g)	n	v	l	d	i	e	b	t	IOCode	IOComment	IOBlank	locCodeAndComment	uniq_Op	un
0	1.1	1.4	1.4	1.4	1.3	1.30	1.30	1.30	1.30	1.30	1.30	1.30	2	2	2	2	1.2	1.
1	1.0	1.0	1.0	1.0	1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1	1	1	1	1
2	72.0	7.0	1.0	6.0	198.0	1134.13	0.05	20.31	55.85	23029.10	0.38	1279.39	51	10	8	1	17	36
3	190.0	3.0	1.0	3.0	600.0	4348.76	0.06	17.06	254.87	74202.67	1.45	4122.37	129	29	28	2	17	13
4	37.0	4.0	1.0	4.0	126.0	599.12	0.06	17.19	34.86	10297.30	0.20	572.07	28	1	6	0	11	16

In [4]: data.tail() #shows Last 5 rows

Out[4]:

	loc	v(g)	ev(g)	iv(g)	n	v	l	d	i	e ...	IOCode	IOComment	IOBlank	locCodeAndComment	uniq_Op	uniq_Opnd	total_
10880	18.0	4.0	1.0	4.0	52.0	241.48	0.14	7.33	32.93	1770.86 ...	13	0	2	0	10	15	
10881	9.0	2.0	1.0	2.0	30.0	129.66	0.12	8.25	15.72	1069.68 ...	5	0	2	0	12	8	
10882	42.0	4.0	1.0	2.0	103.0	519.57	0.04	26.40	19.68	13716.72 ...	29	1	10	0	18	15	
10883	10.0	1.0	1.0	1.0	36.0	147.15	0.12	8.44	17.44	1241.57 ...	6	0	2	0	9	8	
10884	19.0	3.0	1.0	1.0	58.0	272.63	0.09	11.57	23.56	3154.67 ...	13	0	2	1	12	14	

5 rows x 22 columns

In [5]: data.describe() #shows simple statistics (min, max, mean, etc.)

Out[5]:

	loc	v(g)	ev(g)	iv(g)	n	v	l	d	i	e
count	10885.000000	10885.000000	10885.000000	10885.000000	10885.000000	10885.000000	10885.000000	10885.000000	10885.000000	1.088500e+04
mean	42.016178	6.348590	3.401047	4.001599	114.389738	673.758017	0.135335	14.177237	29.439544	3.683637e+04
std	76.593332	13.019695	6.771869	9.116889	249.502091	1938.856196	0.160538	18.709900	34.418313	4.343678e+05
min	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000e+00
25%	11.000000	2.000000	1.000000	1.000000	14.000000	48.430000	0.030000	3.000000	11.860000	1.619400e+02
50%	23.000000	3.000000	1.000000	2.000000	49.000000	217.130000	0.080000	9.090000	21.930000	2.031020e+03
75%	46.000000	7.000000	3.000000	4.000000	119.000000	621.480000	0.160000	18.900000	36.780000	1.141643e+04
max	3442.000000	470.000000	165.000000	402.000000	8441.000000	80843.080000	1.300000	418.200000	569.780000	3.107978e+07

**Learning from experiment:-** We are successful in finding a defect dataset and we analyzed it properly.