

			CALL_PAIRS	CYCLOMATIC_COMPLEXITY	DECISION_DENSITY	GLOBAL_DATA_COMPLEXITY	GLOBAL_DATA_DENSITY	HALSTEAD_LENGTH	NUM_OPERANDS	NUM_OPERATORS	NUMBER_OF_LINES	PATHOLOGICAL_COMPLEXITY	Defective
0	1		0	0	5	2	3	10	1	N			
0	1		0	0	14	5	9	10	1	N			
0	3		0	0	47	19	28	24	1	N			
2	3	2	0	0	71	33	38	23	1	N			
0	1		0	0	30	10	20	8	1	N			
1	3	2	0	0	43	16	27	24	1	N			

Remove columns that have the same value for every row because they do not provide any information for modelling.

CALL_PAIRS			CYCLOMATIC_COMPLEXITY				DECISION_DENSITY				HALSTEAD_LENGTH				NUM_OPERANDS		NUMBER_OF_LINES		Defective
0	1	0	5	2	3	10												N	
0	1	0	14	5	9	10												N	
0	3	0	47	19	28	24												N	
2	3	2	71	33	38	23												N	
0	1	0	30	10	20	8												N	
1	3	2	43	16	27	24												N	

Replace missing *DECISION_DENSITY* values with zero. Based on other MDP datasets without missing *DECISION_DENSITY* values, one can deduce that they likely occurred due to a division by zero error and can be replaced with zeros.

CALL_PAIRS			CYCLOMATIC_COMPLEXITY			DECISION_DENSITY			HALSTEAD_LENGTH			NUM_OPERANDS			NUMBER_OF_LINES			Defective					
0	1	0	5	2	3	10	14	5	9	10	47	19	28	24	71	33	38		23	30	10	20	8
0	1	0	5	2	3	10	14	5	9	10	47	19	28	24	71	33	38	23	30	10	20	8	N
0	3	0	5	2	3	10	14	5	9	10	47	19	28	24	71	33	38	23	30	10	20	8	N
2	3	2	5	2	3	10	14	5	9	10	47	19	28	24	71	33	38	23	30	10	20	8	N
0	1	0	5	2	3	10	14	5	9	10	47	19	28	24	71	33	38	23	30	10	20	8	N
1	3	2	5	2	3	10	14	5	9	10	47	19	28	24	71	33	38	23	30	10	20	8	N

Remove rows that are duplicates of other rows to assure models are tested on unseen data only. Also remove rows that are inconsistent, meaning all column values are the same except for the class label (one is classified as defective and the other is not).

CALL_PAIRS			CYCLOMATIC_COMPLEXITY				DECISION_DENSITY				HALSTEAD_LENGTH				NUM_OPERANDS				NUMBER_OF_LINES				Defective
0	1	0	14	5	9	10															N		
0	3	0	47	19	28	24															N		
2	3	2	71	33	38	23															N		
1	3	2	43	16	27	24															N		

The CM1 dataset is now ready for use in modelling.

Deleted Edited Unchanged*

*A lighter value indicates a missing data value