

Data set	WHERE		CART		Random Forest	
	default	Tuned	default	Tuned	default	Tuned
antV0	30	<b>89</b>	27	<b>89</b>	28	<b>89</b>
antV1	32	<b>74</b>	36	<b>74</b>	41	<b>74</b>
antV2	78	78	52	79	64	<b>80</b>
camelV0	<b>83</b>	<b>83</b>	26	<b>83</b>	50	47
camelV1	22	<b>81</b>	24	26	25	28
ivyV0	16	<b>89</b>	17	25	16	21
jeditV0	35	<b>75</b>	49	48	44	45
jeditV1	24	<b>87</b>	28	78	28	39
jeditV2	2	<b>98</b>	3	18	5	5
log4jV0	94	8	97	<b>100</b>	99	<b>100</b>
luceneV0	61	73	67	<b>79</b>	67	76
poiV0	74	71	<b>79</b>	77	77	76
poiV1	53	<b>90</b>	73	89	82	36
synapseV0	66	66	71	<b>100</b>	65	<b>100</b>
velocityV0	34	<b>43</b>	34	38	34	36
xercesV0	13	<b>85</b>	14	16	15	14
xercesV1	<b>56</b>	26	49	26	49	26

Figure 1: Exp A: Precision results (best results shown in bold).

Data set	WHERE		CART		Random Forest	
	default	Tuned	default	Tuned	default	Tuned
antV0	39	32	32	30	<b>41</b>	23
antV1	11	8	38	<b>49</b>	37	8
antV2	0	<b>52</b>	44	46	47	51
camelV0	0	1	9	28	4	<b>31</b>
camelV1	35	36	31	35	<b>37</b>	32
ivyV0	27	28	28	<b>39</b>	27	33
jeditV0	50	<b>56</b>	<b>56</b>	54	<b>56</b>	<b>56</b>
jeditV1	37	33	36	<b>49</b>	43	<b>49</b>
jeditV2	4	7	5	<b>15</b>	9	9
log4jV0	62	<b>72</b>	47	55	57	52
luceneV0	70	<b>75</b>	56	<b>75</b>	70	74
poiV0	82	71	79	70	<b>83</b>	76
poiV1	5	<b>78</b>	21	<b>78</b>	19	<b>78</b>
synapseV0	0	2	40	<b>56</b>	41	<b>56</b>
velocityV0	<b>51</b>	<b>51</b>	49	<b>51</b>	<b>51</b>	<b>51</b>
xercesV0	22	21	21	22	<b>24</b>	21
xercesV1	25	52	18	53	18	<b>71</b>

Figure 2: Exp A: F-value results (best results shown in bold).

Features	Precision		F		SUM	
	default	tuned	default	tuned	default	tuned
noc				1		1
moa				1		1
ca				2		2
max_cc				2		3
cbo		1		3		4
ce		1		5		6
avg_cc		1		5		5
npm	1		1	5	2	7
lcom	1	2	1	5	2	7
amc	4	3	4	4	8	7
cbm	4	2	4	5	8	7
rfe	3	5	3	10	6	15
wmc	5	6	5	8	10	14
ic	9	3	9	3	18	6
cam	9	3	9	8	18	11
dit	8	6	8	8	16	14
lcom3	9	4	9	10	18	14
loc	9	5	9	11	18	16
dam	13	5	13	12	26	17
mfa	16	6	16	11	32	17

Figure 3: Exp A: Counts of features selected by different goals. Given that we are processing 17 data sets, the maximum counts for any one cell in the “precision” or “F” column is 17.

Datasets	Tuned_Where	Naive_Where	Tuned_CART	Naive_CART	Tuned_RanFst	Naive_RanFst
ant	50/ 94.15	1.50	60/ 3.85	0.08	50/ 7.85	0.17
antV1	60/ 195.89	2.72	70/ 6.77	0.08	60/ 11.53	0.24
antV2	50/ 429.43	7.54	50/ 7.64	0.20	70/ 16.84	0.37
camel	50/ 562.87	7.66	50/ 8.97	0.19	60/ 15.14	0.31
camelV1	50/ 1431.12	24.09	60/ 16.42	0.23	70/ 28.56	0.74
ivy	60/ 68.85	0.99	60/ 3.53	0.07	50/ 6.21	0.18
jedit	50/ 302.27	5.26	50/ 5.44	0.09	60/ 13.71	0.30
jeditV1	50/ 394.04	6.70	70/ 7.10	0.09	90/ 16.71	0.30
jeditV2	50/ 455.78	8.63	80/ 11.42	0.15	60/ 15.89	0.41
log4j	70/ 131.84	1.65	60/ 3.08	0.05	60/ 7.47	0.20
lucene	70/ 268.27	3.19	70/ 6.33	0.09	70/ 12.19	0.26
poi	60/ 341.75	4.87	50/ 5.94	0.10	50/ 12.20	0.34
poiV1	60/ 537.60	7.06	50/ 7.81	0.11	50/ 11.86	0.27
synapse	50/ 115.54	1.88	70/ 4.41	0.07	70/ 8.32	0.16
velocity	100/ 283.81	2.70	50/ 3.54	0.06	50/ 7.61	0.19
xerces	50/ 116.81	2.01	60/ 5.95	0.08	70/ 11.73	0.21
xercesV1	50/ 759.72	13.18	60/ 9.16	0.14	60/ 14.42	0.37

Figure 4: Time (in seconds) spent on different models over the objective of prec

Datasets	Tuned_Where	Naive_Where	Tuned_CART	Naive_CART	Tuned_RanFst	Naive_RanFst
ant	60/ 106.50	1.58	60/ 4.54	0.07	60/ 9.06	0.20
antV1	60/ 209.41	2.89	50/ 5.46	0.09	60/ 13.00	0.24
antV2	60/ 510.36	7.69	70/ 10.23	0.20	60/ 16.69	0.37
camel	60/ 655.36	8.30	70/ 11.79	0.19	60/ 14.82	0.32
camelV1	60/ 1576.70	24.24	50/ 13.33	0.23	50/ 23.39	0.71
ivy	60/ 70.31	1.06	60/ 3.72	0.06	50/ 6.32	0.17
jedit	70/ 380.39	5.40	60/ 5.98	0.09	60/ 12.74	0.29
jeditV1	70/ 515.97	7.01	50/ 5.47	0.09	50/ 11.91	0.30
jeditV2	100/ 769.22	7.18	60/ 7.25	0.12	50/ 13.00	0.35
log4j	60/ 98.41	1.37	50/ 2.16	0.05	50/ 5.19	0.15
lucene	50/ 176.28	2.82	60/ 4.58	0.07	60/ 8.88	0.24
poi	70/ 332.64	4.19	50/ 5.04	0.10	50/ 9.40	0.27
poiV1	70/ 583.11	7.48	80/ 9.44	0.11	60/ 12.53	0.28
synapse	60/ 133.98	1.99	50/ 3.37	0.06	50/ 7.30	0.15
velocity	50/ 170.95	2.80	60/ 4.23	0.07	50/ 7.81	0.20
xerces	60/ 145.12	2.10	50/ 5.34	0.09	70/ 12.09	0.20
xercesV1	70/ 1026.37	13.86	70/ 13.04	0.15	60/ 15.10	0.37

Figure 5: Time (in seconds) spent on different models over the objective of F

Learner Name	Parameters	Default	antV0	antV1	antV2	camelV0	camelV1	ivyV0	jeditV0	jeditV1	jeditV2	log4jV0	luceneV0	poiV0	poiV1	synapseV0	velocityV0	xercesV0	xercesV1
Where based Learner	threshold	0.5	0.53	0.95	0.37	0.3	0.13	0.57	0.74	0.55	0.85	0.39	0.99	0.57	0.8	0.83	0.86	0.92	0.94
	infoPrune	0.33	0.68	0.6	0.86	0.45	0.78	0.82	0.3	0.35	0.57	0.4	0.48	0.69	0.98	0.74	0.77	1	0.97
	min_sample_size	4	3	7	9	2	2	7	9	6	8	2	2	5	1	5	7	4	8
	min_Size	0.5	0.07	0.81	0.56	0.87	0.09	0.49	0.05	0.11	0.72	0.31	1	0.06	0.76	0.71	0.89	0.29	0.84
	wriddle	0.2	0.91	0.45	0.22	0.13	0.01	0.04	0.32	0.46	0.43	0.5	0.34	0.01	0.12	0.53	0.36	0.28	0.81
	depthMin	2	4	4	4	4	3	1	3	4	1	4	5	1	3	2	1	1	2
	depthMax	10	10	9	5	15	12	13	14	13	18	2	6	19	1	3	16	6	19
	wherePrune	False	True	True	True	True	True	True	True	True	True	False	True	False	True	True	True	True	False
	treePrune	True	True	False	True	True	False	True	False	False	False	True	False	False	True	True	False	True	False
	threshold	0.5	0.83	0.78	0.81	0.33	0.97	0.92	0.01	0.78	1	1	1	0.84	0.58	0.41	0.77	0.63	1
CART	max_feature	None	0.01	0.63	0.98	0.01	0.89	0.28	0.01	0.87	0.01	0.01	0.29	1.0	0.56	0.15	0.54	0.73	0.01
	min_samples_split	2	18	13	8	13	9	18	2	4	12	6	10	18	8	2	17	19	14
	min_samples_leaf	3	2	8	6	17	17	1	15	4	7	19	19	13	4	3	20	4	7
	max_depth	None	1	25	25	38	14	21	32	13	1	22	2	23	44	1	16	2	4
Random Forests	threshold	0.5	0.86	0.99	0.87	0.82	0.69	0.78	1	1	1	1	0.72	0.83	0.74	0.71	1	0.99	0.92
	max_feature	None	0.05	0.18	0.31	0.37	0.93	0.95	0.23	1	0.97	0.01	0.61	0.86	0.73	0.01	0.18	0.85	0.01
	max_leaf_nodes	None	16	24	18	23	20	18	40	47	10	50	31	38	17	34	22	40	15
	min_samples_split	2	3	1	14	13	10	13	1	7	1	4	11	1	17	7	14	16	19
	min_samples_leaf	1	7	3	2	5	5	12	2	3	5	15	16	2	5	3	6	2	3
	n_estimators	100	85	55	56	83	110	57	70	96	54	84	136	75	135	80	99	68	116
	n_estimators	100	85	55	56	83	110	57	70	96	54	84	136	75	135	80	99	68	116
	n_estimators	100	85	55	56	83	110	57	70	96	54	84	136	75	135	80	99	68	116

Figure 6: Parameters tuned on different models over the objective of prec

Learner Name	Parameters	Default	antV0	antV1	antV2	camelV0	camelV1	ivyV0	jeditV0	jeditV1	jeditV2	log4jV0	luceneV0	poiV0	poiV1	synapseV0	velocityV0	xercesV0	xercesV1
Where based Learner	threshold	0.5	0.16	0.89	0.59	0.52	0.62	0.85	0.99	0.8	0.88	0.83	0.34	0.92	0.12	0.05	0.79	0.65	0.43
	infoPrune	0.33	0.32	0.4	0.51	0.76	0.68	0.66	0.32	0.35	0.73	0.14	0.13	0.19	0.97	0.02	0.34	0.89	0.51
	min_sample_size	4	8	1	4	1	9	9	4	4	1	8	8	5	6	3	7	4	7
	min_Size	0.5	0.76	0.89	0.57	0.7	0.63	0.62	0.9	0.48	1	0.39	0.95	0.21	0.05	0.59	0.44	0.47	0.53
	wriddle	0.2	0.57	0.85	0.81	0.4	0.51	0.88	0.78	0.7	0.96	0.69	0.81	0.79	0.89	0.5	0.41	1	0.27
	depthMin	2	4	3	2	4	3	5	3	3	2	5	4	1	1	3	3	5	4
	depthMax	10	8	10	18	19	6	12	13	5	10	1	7	4	9	17	12	7	8
	wherePrune	False	False	False	False	True	False	True	False	False	False	False	True	True	False	False	False	True	True
	treePrune	True	True	False	True	True	False	False	False	True	False	False	False	False	False	False	False	True	True
	threshold	0.5	0.01	0.4	0.01	0.01	1	0.99	0.61	0.77	0.58	0.74	0.07	1	0.01	0.01	0.01	0.86	0.01
CART	max_feature	None	0.32	0.1	0.49	0.01	0.96	0.45	0.77	0.15	0.44	0.73	0.27	0.32	0.96	0.04	0.07	0.64	0.29
	min_samples_split	2	3	11	16	9	19	14	16	4	13	4	14	12	20	12	3	15	18
	min_samples_leaf	1	8	10	14	15	1	18	13	4	8	20	13	5	12	11	9	5	7
	max_depth	None	26	48	12	1	27	35	11	44	11	27	32	9	1	46	15	11	7
Random Forests	threshold	0.5	0.01	0.49	0.14	0.49	0.91	0.96	0.65	1.0	1	0.52	0.7	0.95	0.01	0.29	0.52	0.99	0.94
	max_feature	None	0.68	0.49	1	0.44	0.9	0.83	0.24	0.96	0.01	0.47	0.01	0.48	0.01	0.29	0.08	0.81	0.96
	max_leaf_nodes	None	36	30	43	22	29	40	24	41	41	31	14	19	47	32	29	50	16
	min_samples_split	2	16	13	17	1	17	3	10	19	1	16	17	3	6	9	13	7	12
	min_samples_leaf	1	7	15	7	4	9	13	3	15	2	3	20	2	15	14	5	4	11
	n_estimators	100	64	54	84	106	110	76	111	75	149	132	126	87	124	150	142	97	57
	n_estimators	100	64	54	84	106	110	76	111	75	149	132	126	87	124	150	142	97	57
	n_estimators	100	64	54	84	106	110	76	111	75	149	132	126	87	124	150	142	97	57

Figure 7: Parameters tuned on different models over the objective of F