Supplementary Material for PIPES: A Meta-dataset of Machine Learning Pipelines

Anonymous Authors

Abstract. This document presents the supplementary material for the paper "PIPES: A Meta-dataset of Machine Learning Pipelines", submitted to IJCNN 2025. It contains the details that were not included in the paper. In addition to this document, the codes can be found on Github: https://anonymous.4open.science/status/PIPES-C4BE.

I. API PIPES

One of the main objectives of PIPES is to contribute to the community by providing reproducible results in the area and promoting the advancement of research. In Figure 1, we illustrate how users can retrieve and insert more metadata. The projected initial use of PIPES[1] will be through an API (Application Programming Interface).

The datasets component contains the entire collection of metadata stored in the backend, while data retrieval and insertion requests are handled through the API. An example demonstrates this interaction: A POST request sends new metadata, including the dataset ID, name, and location. Afterward, a GET request retrieves all metadata, such as dataset ID, fold information, training and testing times, and a specific metric with its corresponding value.

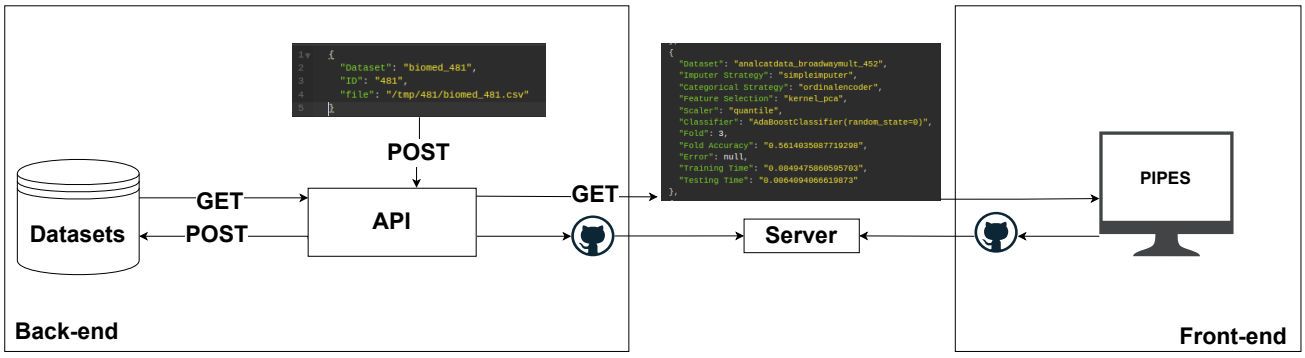

Fig. 1. API design of PIPES.

This API-driven approach makes it easy to interact with the metadata repository. The front-end interface will be designed for better user interaction with metadata, and everything will be connected through GitHub.

II. Description of the Datasets

A. Datasets used in meta-dataset construction

For the construction of the meta-dataset, 381 datasets were selected, 273 are binary-class problems, and 108 are multiclass problems. The datasets are presented in Table I.

[1]https://anonymous.4open.science/status/PIPES-C4BE

TABLE I: Datasets that were used in building meta-dataset.

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
727	2dplanes	2	11	40768
183	abalone	28	9	4177
1455	acute-inflammations	2	7	120
41156	ada	2	49	4147
1037	ada_prior	2	15	4562
179	adult	2	15	48842
42493	airlines	2	8	26969
45035	albert	2	32	58252
46279	algerian_forest_fires	2	14	243
40707	allbp	3	30	3772
4135	Amazon_employee_access	2	10	32769
458	analcatdata_authorship	4	71	841
452	analcatdata_broadwaymult	7	8	285
461	analcatdata_creditscore	2	7	100
1014	analcatdata_dmft	2	5	797
984	analcatdata_draft	2	5	366
1025	analcatdata_germangss	2	6	400
852	analcatdata_gsssexsurvey	2	10	159
966	analcatdata_halloffame	2	17	1340
449	analcatdata_homerun	5	27	163
450	analcatdata_lawsuit	2	5	264
986	analcatdata_marketing	2	33	364
1008	analcatdata_reviewer	2	8	379
728	analcatdata_supreme	2	8	4052
748	analcatdata_wildcat	2	6	163
2	anneal	5	39	898
45613	appendicitis_test_edsa	2	8	106
1059	ar1	2	30	121
1061	ar4	2	30	107
1064	ar6	2	30	101
949	arsenic-female-bladder	2	5	559
950	arsenic-female-lung	2	5	559
947	arsenic-male-bladder	2	5	559
951	arsenic-male-lung	2	5	559
1459	artificial-characters	10	8	10218
7	audiology	24	70	226
40981	Australian	2	15	690
840	autoHorse	2	26	205
831	autoMpg	2	8	398
756	autoPrice	2	16	159
9	autos	6	26	205
975	autos	2	26	205
46331	autos_clean	6	26	205
1547	autoUniv-au1-1000	2	21	1000
1555	autoUniv-au6-1000	8	41	1000
1551	autoUniv-au6-400	8	41	400
1549	autoUniv-au6-750	8	41	750
1552	autoUniv-au7-1100	5	13	1100
1554	autoUniv-au7-500	5	13	500
1553	autoUniv-au7-700	3	13	700
745	auto_price	2	16	159
463	backache	2	32	180
1121	badges2	2	11	294

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
11	balance-scale	3	5	625
1461	bank-marketing	2	17	45211
833	bank32nh	2	33	8192
725	bank8FM	2	9	8192
1462	banknote-authentication	2	5	1372
185	baseball	3	17	1340
481	biomed	2	9	209
45037	BitcoinHeist_Ransomware	2	8	24780
46280	blastchar	2	20	7043
1463	blogger	2	6	100
1464	blood-transfusion-service-center	2	5	748
251	BNG(breast-w)	2	10	39366
255	BNG(cmc)	3	10	55296
137	BNG(tic-tac-toe)	2	10	39366
778	bodyfat	2	15	252
853	boston	2	14	506
825	boston_corrected	2	21	506
186	braziltourism	7	9	412
13	breast-cancer	2	10	286
1465	breast-tissue	6	10	106
15	breast-w	2	10	699
844	breastTumor	2	10	286
327	bridges	6	12	105
45717	bwin_amlb	2	14	530
40663	calendarDOW	5	33	399
45578	California-Housing-Classification	2	9	20640
43979	california	2	9	20634
40664	car-evaluation	4	22	1728
1466	cardiotocography	10	36	2126
45547	Cardiovascular-Disease-dataset	2	12	70000
455	cars	3	8	406
991	car	2	7	1728
1447	CastMetal1	2	38	327
820	chatfield_4	2	13	235
798	cholesterol	2	14	303
909	chscase_census2	2	8	400
908	chscase_census3	2	8	400
907	chscase_census4	2	8	400
906	chscase_census5	2	8	400
900	chscase_census6	2	7	400
939	chscase_whale	2	9	228
40701	churn	2	21	5000
1024	cjs	2	35	2796
786	cleveland	2	14	303
1220	Click_prediction_small	2	10	39948
1467	climate-model-simulation-crashes	2	21	540
890	cloud	2	8	108
23	cmc	3	10	1473
27	colic	2	23	368
897	colleges_aaup	2	16	1161
987	collins	2	23	500
42192	compas-two-years	2	14	5278
44162	compass	2	18	16644
41538	conference_attendance	2	7	246
40668	connect-4	3	43	67557

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
40669	corral	2	7	160
1446	CostaMadre1	2	38	296
44377	covertype_seed_0_nrows_2000_ nclasses_10_ncols_100_stratify_True	2	11	2000
41701	CPMP-2015-classification	4	27	527
796	cpu	2	8	209
735	cpu_small	2	13	8192
31	credit-g	2	21	1000
45938	credit-score-classification-Hzl	3	28	100000
4154	CreditCardSubset	2	31	14240
44089	credit	2	11	16714
6332	cylinder-bands	2	40	540
1075	datatrieve	2	9	130
42477	default-of-credit-card-clients	2	24	30000
803	delta_ailerons	2	6	7129
35	dermatology	6	35	366
37	diabetes	2	9	768
41430	DiabeticMellitus	2	98	281
818	diggle_table_a2	2	9	310
40713	dis	2	30	3772
45712	doa_bwin_balanced	2	14	708
23381	dresses-sales	2	13	500
45604	dummy	2	7	1000
42177	echocardiogram-uci	4	8	132
944	echoMonths	2	10	130
1011	ecoli	2	8	336
1471	eeg-eye-state	2	15	14980
46298	electrical_grid_stability_simulated _classification	2	13	10000
151	electricity	2	9	45312
846	elevators	2	19	16599
43551	Employee-Turnover-at-TECHCO	2	10	34452
1472	energy-efficiency	37	10	768
4340	Engine1	3	6	383
42169	epiparo_extract	6	20	224
188	eucalyptus	5	20	736
1044	eye_movements	3	28	10936
1473	fertility	2	10	100
45553	FICO-HELOC-cleaned	2	24	9871
1475	first-order-theorem-proving	6	52	6118
854	fishcatch	2	8	158
285	flags	8	29	194
46174	Flare	6	12	1066
41769	FOREX_audcad-day-Close	2	12	1834
901	fried	2	11	40768
904	fri_c0_1000_50	2	51	1000
888	fri_c0_500_50	2	51	500
837	fri_c1_1000_50	2	51	1000
766	fri_c1_500_50	2	51	500
866	fri_c2_1000_50	2	51	1000
920	fri_c2_500_50	2	51	500
806	fri_c3_1000_50	2	51	1000
937	fri_c3_500_50	2	51	500
797	fri_c4_1000_50	2	51	1000
805	fri_c4_500_50	2	51	500

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
714	fruitfly	2	5	125
40646	GAMETES_Epistasis_2-Way _20atts_0.1H_EDM-1_1	2	21	1600
46116	German-Credit-Risk-with-Target	2	10	1000
4538	GesturePhaseSegmentationProcessed	5	33	9873
1005	glass	2	10	214
1026	grub-damage	2	9	155
329	hayes-roth	3	5	160
49	heart-c	2	14	303
51	heart-h	2	14	294
1512	heart-long-beach	5	14	200
53	heart-statlog	2	14	270
1513	heart-switzerland	5	13	123
41169	helena	100	28	65196
45023	heloc	2	23	10000
55	hepatitis	2	20	155
23512	higgs	2	29	98050
823	houses	2	9	20640
821	house_16H	2	17	22784
843	house_8L	2	9	22784
858	hungarian	2	14	294
57	hypothyroid	4	30	3772
43893	ibm-employee-attrition	2	35	1470
41945	ilpd-numeric	2	11	583
1480	ilpd	2	11	583
46281	insurance_company	2	86	9822
46282	internet_usage	2	69	10108
59	ionosphere	2	35	351
382	ipums_la_97-small	8	61	7019
61	iris	3	5	150
41168	jannis	4	55	83733
375	JapaneseVowels	9	15	9961
1073	jEdit_4.0_4.2	2	9	274
1053	jm1	2	22	10885
41001	jungle_chess_2pcs_endgame_complete	3	47	44819
1066	kc1-binary	2	95	145
1045	kc1-top5	2	95	145
1067	kc1	2	22	2109
1063	kc2	2	22	522
1065	kc3	2	40	458
43947	KDDCup09_upselling	2	46	5032
981	kdd_internet_usage	2	69	10108
41162	kick	2	33	72983
807	kin8nm	2	9	8192
46173	King-rook-vs-King	18	7	28056
1448	KnuggetChase3	2	40	194
1481	kr-vs-k	18	7	28056
184	kropt	18	7	28056
1441	KungChi3	2	40	123
43890	law-school-admission-bianry	2	12	20800
1482	leaf	30	16	340
40496	LED-display-domain-7digit	10	8	500
40678	led7	10	8	3200
1222	letter-challenge-unlabeled.arff	3	17	20000
6	letter	26	17	20000

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
43595	Loan-Predication	2	12	614
941	lowbwt	2	10	189
1412	lungcancer_GSE31210	2	24	226
10	lymph	4	19	148
46264	mabbob_ela_as_2d_classify	5	46	1120
733	machine_cpu	2	7	209
1120	MagicTelescope	2	12	19020
45557	Mammographic-Mass-Data-Set	2	5	961
310	mammography	2	7	11183
1056	mc1	2	39	9466
1054	mc2	2	40	161
1449	MeanWhile1	2	38	253
1442	MegaWatt1	2	38	253
44224	MembershipWoes	2	15	10362
757	meta	2	22	528
279	meta_stream_intervals.arff	11	75	45164
14	mfeat-fourier	10	77	2000
1020	mfeat-karhunen	2	65	2000
18	mfeat-morphological	10	7	2000
22	mfeat-zernike	10	48	2000
40966	MiceProtein	8	82	1080
41671	microaggregation2	5	21	20000
42530	Midwest_Survey	10	28	2778
1450	MindCave2	2	40	125
43974	MiniBooNE	2	51	72998
40680	mofn-3-7-10	2	11	1324
164	molecular-biology_promoters	2	58	106
334	monks-problems-2	2	7	601
335	monks-problems-3	2	7	554
1046	mozilla4	2	6	15545
880	mu284	2	11	284
24	mushroom	2	23	8124
40681	mux6	2	7	128
881	mv	2	11	40768
1071	mw1	2	38	403
43892	national-longitudinal-survey-binary	2	17	4908
44226	NewspaperChurn	2	19	15855
886	no2	2	8	500
23517	numerai28.6	2	22	96320
26	nursery	5	9	12960
311	oil_spill	2	50	937
45067	okcupid_stem	3	14	26677
1491	one-hundred-plants-margin	100	65	1600
45060	online_shoppers	2	18	12330
42738	open_payments	2	6	73558
28	optdigits	10	65	5620
45548	Otto-Group-Product-Classification-Challenge	9	94	61878
1487	ozone-level-8hr	2	73	2534
1021	page-blocks	2	11	5473
40706	parity5_plus_5	2	11	1124
1488	parkinsons	2	23	195
810	pbc	2	19	418
1068	pc1	2	22	1109
1069	pc2	2	37	5589

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
1050	pc3	2	38	1563
1049	pc4	2	38	1458
32	pendigits	10	17	10992
42585	penguins	3	7	344
738	pharynx	2	11	195
4534	PhishingWebsites	2	31	11055
1489	phoneme	2	6	5404
1451	PieChart1	2	38	705
1452	PieChart2	2	37	745
1453	PieChart3	2	38	1077
1443	PizzaCutter1	2	38	661
1444	PizzaCutter3	2	38	1043
1490	planning-relax	2	13	182
915	plasma_retinol	2	14	315
750	pm10	2	8	500
871	pollen	2	6	3848
722	pol	2	49	15000
1100	PopularKids	3	11	478
45714	PriceRunner	10	6	35300
1003	primary-tumor	2	18	339
446	prnn_crabs	2	8	200
952	prnn_fglass	6	10	214
470	profb	2	10	672
45558	Pulsar-Dataset-HTRU2	2	9	17898
752	puma32H	2	33	8192
816	puma8NH	2	9	8192
721	pwLinear	2	11	200
1494	qsar-biodeg	2	42	1055
45077	qsar	2	41	1055
1495	qualitative-bankruptcy	2	7	250
42172	regime_alimentaire	2	20	202
42665	ricci_vs_destefano	2	6	118
1496	ringnorm	2	21	7400
43949	rl	2	13	4970
717	rmftsa_ladata	2	11	508
1519	robot-failures-lp4	3	91	117
1520	robot-failures-lp5	5	91	164
40922	Run_or_walk_information	2	7	88588
1498	sa-heart	2	10	462
40900	Satellite	2	37	5100
182	satimage	6	37	6430
466	schizo	2	15	340
1499	seeds	3	8	210
36	segment	7	20	2310
1500	seismic-bumps	3	8	210
826	sensory	2	12	576
747	servo	2	5	167
44773	sf-police-incidents_seed_0_nrows _2000_nclasses_10_ncols_ 100_stratify_True	2	9	2000
40685	shuttle	7	10	58000
38	sick	2	30	3772
41946	Sick_numeric	2	30	3772
902	sleuth_case2002	2	7	147

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
4153	Smartphone-Based_Recognition_of _Human_Activities	6	68	180
934	socmob	2	6	1156
40687	solar-flare	6	13	1066
40	sonar	2	61	208
44227	South_Asian_Churn_dataset	2	14	2000
42	soybean	19	36	683
737	space_ga	2	7	3107
44	spambase	2	58	4601
336	SPECT	2	23	267
46	splice	3	61	3190
40982	steel-plates-fault	7	28	1941
841	stock	2	10	950
42167	stress	3	13	202
770	strikes	2	7	625
43097	students_scores	2	8	1000
46026	Stylized_Meta_Album _APL_STY_Mini	20	7	16000
41146	sylvine	2	21	5124
1004	synthetic_control	2	61	600
48	tae	3	6	151
1115	teachingAssistant	3	7	151
42178	telco-customer-churn	2	20	7043
41526	test_dataset	2	61	15547
40499	texture	11	41	5500
4329	thoracic_surgery	2	17	470
40690	threeOf9	2	10	512
40474	thyroid-allbp	5	27	2800
50	tic-tac-toe	2	10	958
40945	Titanic	2	14	1309
40705	tokyo1	2	45	959
42544	Touch2	8	11	265
45545	Tour-and-Travels-Customer -Churn-Prediction	2	7	954
42345	Traffic_violations	3	21	70340
788	triazines	2	61	186
41976	TuningSVMs	2	81	156
1507	twonorm	2	21	7400
44232	UCI_churn	2	21	3333
1508	user-knowledge	5	6	403
1047	usp05	11	17	203
54	vehicle	4	19	846
44153	vehicle_reproduced	4	19	846
1523	vertebra-column	3	7	310
719	veteran	2	8	137
925	visualizing_galaxy	2	5	323
923	visualizing_soil	2	5	8641
56	vote	2	17	435
1016	vowel	2	14	990
1497	wall-robot-navigation	4	25	5456
940	water-treatment	2	37	527
60	waveform-5000	3	41	5000
1510	wdbc	2	31	569
1511	wholesale-customers	2	9	440
40983	wilt	2	6	4839

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
847	wind	2	15	6574
187	wine	3	14	178
753	wisconsin	2	33	194
43607	WMO-Hurricane-Survival -Dataset	2	23	5021
40693	xd6	2	10	973
181	yeast	10	9	1484
43786	Zombies-Apocalypse	2	13	200
62	zoo	7	17	101

III. Meta-Features

Extracted 145 meta-features, which are defined in the Table II.

TABLE II: Meta-Features

Meta-feature	Description	Group
attr_to_inst	Ratio between the number of attributes.	simple
cat_to_num	Ratio between the number of categoric and numeric features.	simple
freq_class.mean	Relative frequency of each distinct class.	simple
freq_class.sd	Relative frequency of each distinct class.	simple
inst_to_attr	Ratio between the number of instances and attributes.	simple
nr_attr	Total number of attributes.	simple
nr_bin	Number of binary attributes.	simple
nr_cat	Number of categorical attributes.	simple
nr_class	Number of distinct classes.	simple
nr_inst	Number of instances (rows) in the dataset.	simple
nr_num	Number of numeric features.	simple
num_to_cat	Number of numerical and categorical features.	simple
attr_conc.mean	Concentration coef. of each pair of distinct attributes.	information theory
attr_conc.sd	Concentration coef. of each pair of distinct attributes.	information theory
attr_ent.mean	Shannon's entropy for each predictive attribute.	information theory
attr_ent.sd	Shannon's entropy for each predictive attribute.	information theory
class_conc.mean	Concentration coefficient between each attribute and class.	information theory
class_conc.sd	Concentration coefficient between each attribute and class.	information theory
class_ent	Target attribute Shannon's entropy.	information theory
eq_num_attr	Number of attributes equivalent for a predictive task.	information theory
joint_ent.mean	Joint entropy between each attribute and class.	information theory
joint_ent.sd	Joint entropy between each attribute and class.	information theory
mut_inf.mean	Mutual information between each attribute and target.	information theory
mut_inf.sd	Mutual information between each attribute and target.	information theory
ns_ratio	Noisiness of attributes.	information theory
can_cor.mean	Canonical correlations of data.	statistical
can_cor.sd	Canonical correlations of data.	statistical
cor.mean	Absolute value of the correlation of distinct dataset column pairs.	statistical

Meta-feature	Description	Group
cor.sd	Absolute value of the correlation of distinct dataset column pairs.	statistical
cov.mean	Absolute value of the covariance of distinct dataset attribute pairs.	statistical
cov.sd	Absolute value of the covariance of distinct dataset attribute pairs.	statistical
eigenvalues.mean	Eigenvalues of covariance matrix from dataset.	statistical
eigenvalues.sd	Eigenvalues of covariance matrix from dataset.	statistical
g_mean.mean	Geometric mean of each attribute.	statistical
g_mean.sd	Geometric mean of each attribute.	statistical
gravity	Distance between minority and majority classes center of mass.	statistical
h_mean.mean	Harmonic mean of each attribute.	statistical
h_mean.sd	Harmonic mean of each attribute.	statistical
iq_range.mean	Interquartile range (IQR) of each attribute.	statistical
iq_range.sd	Interquartile range (IQR) of each attribute.	statistical
kurtosis.mean	Kurtosis of each attribute.	statistical
kurtosis.sd	Kurtosis of each attribute.	statistical
lh_trace	Lawley-Hotelling trace.	statistical
mad.mean	Median Absolute Deviation (MAD) adjusted by a factor.	statistical
mad.sd	Median Absolute Deviation (MAD) adjusted by a factor.	statistical
max.mean	Maximum value from each attribute.	statistical
max.sd	Maximum value from each attribute.	statistical
mean.mean	Mean value of each attribute.	statistical
mean.sd	Median value from each attribute.	statistical
median.mean	Median value from each attribute.	statistical
median.sd	Median value from each attribute.	statistical
min.mean	Minimum value from each attribute.	statistical
min.sd	Minimum value from each attribute.	statistical
nr_cor_attr	Number of distinct highly correlated pair of attributes.	statistical
nr_disc	Number of canonical correlation between each attribute and class.	statistical
nr_norm	Number of attributes normally distributed based in a given method.	statistical
nr_outliers	Number of attributes with at least one outlier value.	statistical
p_trace	Pillai's trace.	statistical
range.mean	Range (max - min) of each attribute.	statistical
range.sd	Range (max - min) of each attribute.	statistical
roy_root	Roy's largest root.	statistical
sd.mean	Standard deviation of each attribute.	statistical
sd.sd	Standard deviation of each attribute.	statistical
sd_ratio	Statistical test for homogeneity of covariances.	statistical
skewness.mean	Skewness for each attribute.	statistical
skewness.sd	Skewness for each attribute.	statistical
sparsity.mean	Sparsity metric for each attribute.	statistical
sparsity.sd	Sparsity metric for each attribute.	statistical
t_mean.mean	Trimmed mean of each attribute.	statistical
t_mean.sd	Trimmed mean of each attribute.	statistical
var.mean	variance of each attribute.	statistical
var.sd	Variance of each attribute.	statistical
w_lambda	Wilks' Lambda value.	statistical
leaves	Number of leaf nodes in the DT model.	model-based

Meta-feature	Description	Group
leaves_branch.mean	Size of branches in the DT model.	model-based
leaves_branch.sd	Size of branches in the DT model.	model-based
leaves_corrob.mean	Leaves corroboration of the DT model.	model-based
leaves_corrob.sd	Leaves corroboration of the DT model.	model-based
leaves_homo.mean	DT model Homogeneity for every leaf node.	model-based
leaves_homo.sd	DT model Homogeneity for every leaf node.	model-based
leaves_per_class.mean	Proportion of leaves per class in DT model.	model-based
leaves_per_class.sd	Proportion of leaves per class in DT model.	model-based
nodes	Number of non-leaf nodes in DT model.	model-based
nodes_per_attr	Ratio of nodes per number of attributes in DT model.	model-based
nodes_per_inst	Ratio of non-leaf nodes per number of instances in DT model.	model-based
nodes_per_level.mean	Ratio of number of nodes per tree level in DT model.	model-based
nodes_per_level.sd	Ratio of number of nodes per tree level in DT model.	model-based
nodes_repeated.mean	Number of repeated nodes in DT model.	model-based
nodes_repeated.sd	Number of repeated nodes in DT model.	model-based
tree_depth.mean	Depth of every node in the DT model.	model-based
tree_depth.sd	Depth of every node in the DT model.	model-based
tree_imbalance.mean	Tree imbalance for each leaf node.	model-based
tree_imbalance.sd	Tree imbalance for each leaf node.	model-based
tree_shape.mean	Tree shape for every leaf node.	model-based
tree_shape.sd	Tree shape for every leaf node.	model-based
var_importance.mean	Features importance of the DT model for each attribute.	model-based
var_importance.sd	Features importance of the DT model for each attribute.	model-based
best_node.mean	Performance of a the best single decision tree node.	landmarking
best_node.sd	Performance of a the best single decision tree node.	landmarking
elite_nn.mean	Performance of Elite Nearest Neighbor.	landmarking
elite_nn.sd	Performance of Elite Nearest Neighbor.	landmarking
linear_discr.mean	Performance of the Linear Discriminant classifier.	landmarking
linear_discr.sd	Performance of the Linear Discriminant classifier.	landmarking
naive_bayes.mean	Performance of the Naive Bayes classifier.	landmarking
naive_bayes.sd	Performance of the Naive Bayes classifier.	landmarking
one_nn.mean	Performance of the 1-Nearest Neighbor classifier.	landmarking
one_nn.sd	Performance of the 1-Nearest Neighbor classifier.	landmarking
random_node.mean	Performance of the single decision tree node model induced by a random attribute.	landmarking
random_node.sd	Performance of the single decision tree node model induced by a random attribute.	landmarking
worst_node.mean	Performance of the single decision tree node model induced by the worst informative attribute.	landmarking
worst_node.sd	Performance of the single decision tree node model induced by the worst informative attribute.	landmarking
c1	Entropy of class proportions.	complexity
c2	Imbalance ratio.	complexity
cls_coef	Clustering coefficient.	complexity
density	Average density of the network.	complexity
f1.mean	Maximum Fisher's discriminant ratio.	complexity
f1.sd	Maximum Fisher's discriminant ratio.	complexity
f1v.mean	Directional-vector maximum Fisher's discriminant ratio.	complexity

Meta-feature	Description	Group
f1v.sd	Directional-vector maximum Fisher's discriminant ratio.	complexity
f2.mean	Volume of the overlapping region.	complexity
f2.sd	Volume of the overlapping region.	complexity
f3.mean	Feature maximum individual efficiency.	complexity
f3.sd	Feature maximum individual efficiency.	complexity
f4.mean	Collective feature efficiency.	complexity
f4.sd	Collective feature efficiency	complexity
hubs.mean	Hub score.	complexity
hubs.sd	Hub score.	complexity
l1.mean	Sum of error distance by linear programming.	complexity
l1.sd	Sum of error distance by linear programming.	complexity
l2.mean	OVO subsets error rate of linear classifier.	complexity
l2.sd	OVO subsets error rate of linear classifier.	complexity
l3.mean	Non-Linearity of a linear classifier.	complexity
l3.sd	Non-Linearity of a linear classifier.	complexity
lsc	Local set average cardinality	complexity
n1	Fraction of borderline points	complexity
n2.mean	Ratio of intra and extra class nearest neighbor distance.	complexity
n2.sd	Ratio of intra and extra class nearest neighbor distance.	complexity
n3.mean	Error rate of the nearest neighbor classifier.	complexity
n3.sd	Error rate of the nearest neighbor classifier.	complexity
n4.mean	Non-linearity of the k-NN Classifier.	complexity
n4.sd	Non-linearity of the k-NN Classifier.	complexity
t1.mean	Fraction of hyperspheres covering data.	complexity
t1.sd	Fraction of hyperspheres covering data.	complexity
t2	Average number of features per dimension.	complexity
t3	Average number of PCA dimensions per points.	complexity
t4	Ratio of the PCA dimension to the original dimension.	complexity

A. Datasets used analysis

Table III lists 204 datasets used in Section V's analyses.

TABLE III: Datasets that were used in analysis.

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
40981	Australian	2	15	690
1447	CastMetal1	2	38	327
1446	CostaMadre1	2	38	296
41430	DiabeticMellitus	2	98	281
4340	Engine1	3	6	383
46116	German-Credit-Risk-with-Target	2	10	1000
1448	KnuggetChase3	2	40	194
1441	KungChi3	2	40	123
40496	LED-display-domain-7digit	10	8	500
43595	Loan-Predication	2	12	614
45557	Mammographic-Mass-Data-Set	2	5	961
1449	MeanWhile1	2	38	253
1442	MegaWatt1	2	38	253
1450	MindCave2	2	40	125
4534	PhishingWebsites	2	31	11055

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
1451	PieChart1	2	38	705
1452	PieChart2	2	37	745
1453	PieChart3	2	38	1077
1443	PizzaCutter1	2	38	661
1444	PizzaCutter3	2	38	1043
1100	PopularKids	3	11	478
45558	Pulsar-Dataset-HTRU2	2	9	17898
336	SPECT	2	23	267
40900	Satellite	2	37	5100
41946	Sick_numeric	2	30	3772
4153	Smartphone-Based_Recognition_of _Human_Activities	6	68	180
183	abalone	28	9	4177
1455	acute-inflammations	2	7	120
46279	algerian_forest_fires	2	14	243
458	analcatdata_authorship	4	71	841
452	analcatdata_broadwaymult	7	8	285
461	analcatdata_creditscore	2	7	100
1014	analcatdata_dmft	2	5	797
984	analcatdata_draft	2	5	366
1025	analcatdata_germangss	2	6	400
852	analcatdata_gsssexsurvey	2	10	159
966	analcatdata_halloffame	2	17	1340
450	analcatdata_lawsuit	2	5	264
986	analcatdata_marketing	2	33	364
1008	analcatdata_reviewer	2	8	379
728	analcatdata_supreme	2	8	4052
748	analcatdata_wildcat	2	6	163
45613	appendicitis_test_edsa	2	8	106
1059	ar1	2	30	121
1061	ar4	2	30	107
1064	ar6	2	30	101
949	arsenic-female-bladder	2	5	559
950	arsenic-female-lung	2	5	559
947	arsenic-male-bladder	2	5	559
951	arsenic-male-lung	2	5	559
756	autoPrice	2	16	159
1547	autoUniv-au1-1000	2	21	1000
1555	autoUniv-au6-1000	8	41	1000
1551	autoUniv-au6-400	8	41	400
1549	autoUniv-au6-750	8	41	750
1552	autoUniv-au7-1100	5	13	1100
1554	autoUniv-au7-500	5	13	500
1553	autoUniv-au7-700	3	13	700
463	backache	2	32	180
1121	badges2	2	11	294
11	balance-scale	3	5	625
481	biomed	2	9	209
1463	blogger	2	6	100
1464	blood-transfusion-service-center	2	5	748
778	bodyfat	2	15	252
853	boston	2	14	506
15	breast-w	2	10	699
844	breastTumor	2	10	286
327	bridges	6	12	105

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
45717	bwin_amlb	2	14	530
40663	calendarDOW	5	33	399
820	chatfield_4	2	13	235
798	cholesterol	2	14	303
909	chscase_census2	2	8	400
908	chscase_census3	2	8	400
907	chscase_census4	2	8	400
906	chscase_census5	2	8	400
900	chscase_census6	2	7	400
939	chscase_whale	2	9	228
786	cleveland	2	14	303
1467	climate-model-simulation-crashes	2	21	540
890	cloud	2	8	108
27	colic	2	23	368
897	colleges_aaup	2	16	1161
987	collins	2	23	500
41538	conference_attendance	2	7	246
40669	corral	2	7	160
796	cpu	2	8	209
31	credit-g	2	21	1000
1075	datatrieve	2	9	130
35	dermatology	6	35	366
37	diabetes	2	9	768
818	diggle_table_a2	2	9	310
45712	doa_bwin_balanced	2	14	708
45604	dummy	2	7	1000
944	echoMonths	2	10	130
1011	ecoli	2	8	336
1472	energy-efficiency	37	10	768
188	eucalyptus	5	20	736
1473	fertility	2	10	100
285	flags	8	29	194
904	fri_c0_1000_50	2	51	1000
888	fri_c0_500_50	2	51	500
837	fri_c1_1000_50	2	51	1000
766	fri_c1_500_50	2	51	500
866	fri_c2_1000_50	2	51	1000
920	fri_c2_500_50	2	51	500
806	fri_c3_1000_50	2	51	1000
937	fri_c3_500_50	2	51	500
797	fri_c4_1000_50	2	51	1000
805	fri_c4_500_50	2	51	500
714	fruitfly	2	5	125
1005	glass	2	10	214
1026	grub-damage	2	9	155
329	hayes-roth	3	5	160
49	heart-c	2	14	303
51	heart-h	2	14	294
1512	heart-long-beach	5	14	200
53	heart-statlog	2	14	270
1513	heart-switzerland	5	13	123
55	hepatitis	2	20	155
41945	ilpd-numeric	2	11	583
59	ionosphere	2	35	351
61	iris	3	5	150

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
1073	jEdit_4.0_4.2	2	9	274
1066	kc1-binary	2	95	145
1045	kc1-top5	2	95	145
1067	kc1	2	22	2109
1063	kc2	2	22	522
1065	kc3	2	40	458
807	kin8nm	2	9	8192
1481	kr-vs-k	18	7	28056
184	kropt	18	7	28056
1056	mc1	2	39	9466
1054	mc2	2	40	161
1488	parkinsons	2	23	195
750	pm10	2	8	500
871	pollen	2	6	3848
722	pol	2	49	15000
1003	primary-tumor	2	18	339
446	prnn_crabs	2	8	200
952	prnn_fglass	6	10	214
470	profb	2	10	672
752	puma32H	2	33	8192
816	puma8NH	2	9	8192
721	pwLinear	2	11	200
1495	qualitative-bankruptcy	2	7	250
42172	regime_alimentaire	2	20	202
42665	ricci_vs_destefano	2	6	118
717	rmftsa_ladata	2	11	508
1519	robot-failures-lp4	3	91	117
1520	robot-failures-lp5	5	91	164
1498	sa-heart	2	10	462
182	satimage	6	37	6430
466	schizo	2	15	340
1499	seeds	3	8	210
36	segment	7	20	2310
1500	seismic-bumps	3	8	210
826	sensory	2	12	576
747	servo	2	5	167
38	sick	2	30	3772
902	sleuth_case2002	2	7	147
934	socmob	2	6	1156
40687	solar-flare	6	13	1066
40	sonar	2	61	208
42	soybean	19	36	683
737	space_ga	2	7	3107
44	spambase	2	58	4601
46	splice	3	61	3190
40982	steel-plates-fault	7	28	1941
841	stock	2	10	950
770	strikes	2	7	625
43097	students_scores	2	8	1000
41146	sylvine	2	21	5124
1004	synthetic_control	2	61	600
48	tae	3	6	151
1115	teachingAssistant	3	7	151
41526	test_dataset	2	61	15547
4329	thoracic_surgery	2	17	470

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
40690	threeOf9	2	10	512
40705	tokyo1	2	45	959
788	triazines	2	61	186
1507	twonorm	2	21	7400
1508	user-knowledge	5	6	403
1047	usp05	11	17	203
54	vehicle	4	19	846
44153	vehicle_reproduced	4	19	846
719	veteran	2	8	137
925	visualizing_galaxy	2	5	323
923	visualizing_soil	2	5	8641
56	vote	2	17	435
1016	vowel	2	14	990
1497	wall-robot-navigation	4	25	5456
940	water-treatment	2	37	527
60	waveform-5000	3	41	5000
1510	wdbc	2	31	569
1511	wholesale-customers	2	9	440
40983	wilt	2	6	4839
847	wind	2	15	6574
187	wine	3	14	178
753	wisconsin	2	33	194
40693	xd6	2	10	973
181	yeast	10	9	1484
62	zoo	7	17	101

B. Datasets comparative OpenML with PIPES

Table IV lists 153 datasets used in Section V's analyses.

TABLE IV: Datasets that were used comparative - OpenML versus PIPES

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
40981	Australian	2	15	690
1447	CastMetal1	2	38	327
1446	CostaMadre1	2	38	296
40496	LED-display-domain-7digit	10	8	500
1442	MegaWatt1	2	38	253
4534	PhishingWebsites	2	31	11055
1451	PieChart1	2	38	705
1452	PieChart2	2	37	745
1453	PieChart3	2	38	1077
1443	PizzaCutter1	2	38	661
1444	PizzaCutter3	2	38	1043
1100	PopularKids	3	11	478
336	SPECT	2	23	267
40900	Satellite	2	37	5100
183	abalone	28	9	4177
1455	acute-inflammations	2	7	120
458	analcatdata_authorship	4	71	841
461	analcatdata_creditscore	2	7	100
1014	analcatdata_dmft	2	5	797
1025	analcatdata_germangss	2	6	400
450	analcatdata_lawsuit	2	5	264

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
728	analcatdata_supreme	2	8	4052
748	analcatdata_wildcat	2	6	163
1059	ar1	2	30	121
1061	ar4	2	30	107
1064	ar6	2	30	101
949	arsenic-female-bladder	2	5	559
950	arsenic-female-lung	2	5	559
947	arsenic-male-bladder	2	5	559
951	arsenic-male-lung	2	5	559
756	autoPrice	2	16	159
1547	autoUniv-au1-1000	2	21	1000
1555	autoUniv-au6-1000	8	41	1000
1551	autoUniv-au6-400	8	41	400
1549	autoUniv-au6-750	8	41	750
1552	autoUniv-au7-1100	5	13	1100
1554	autoUniv-au7-500	5	13	500
1553	autoUniv-au7-700	3	13	700
463	backache	2	32	180
1121	badges2	2	11	294
11	balance-scale	3	5	625
481	biomed	2	9	209
1463	blogger	2	6	100
1464	blood-transfusion-service-center	2	5	748
778	bodyfat	2	15	252
853	boston	2	14	506
15	breast-w	2	10	699
844	breastTumor	2	10	286
820	chatfield_4	2	13	235
798	cholesterol	2	14	303
909	chscase_census2	2	8	400
908	chscase_census3	2	8	400
907	chscase_census4	2	8	400
906	chscase_census5	2	8	400
900	chscase_census6	2	7	400
786	cleveland	2	14	303
1467	climate-model-simulation-crashes	2	21	540
890	cloud	2	8	108
987	collins	2	23	500
796	cpu	2	8	209
31	credit-g	2	21	1000
1075	datatrieve	2	9	130
35	dermatology	6	35	366
37	diabetes	2	9	768
818	diggle_table_a2	2	9	310
944	echoMonths	2	10	130
1011	ecoli	2	8	336
188	eucalyptus	5	20	736
1473	fertility	2	10	100
285	flags	8	29	194
904	fri_c0_1000_50	2	51	1000
888	fri_c0_500_50	2	51	500
837	fri_c1_1000_50	2	51	1000
766	fri_c1_500_50	2	51	500
866	fri_c2_1000_50	2	51	1000
920	fri_c2_500_50	2	51	500

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
806	fri_c3_1000_50	2	51	1000
937	fri_c3_500_50	2	51	500
797	fri_c4_1000_50	2	51	1000
805	fri_c4_500_50	2	51	500
714	fruitfly	2	5	125
1005	glass	2	10	214
1026	grub-damage	2	9	155
1512	heart-long-beach	5	14	200
53	heart-statlog	2	14	270
1513	heart-switzerland	5	13	123
55	hepatitis	2	20	155
59	ionosphere	2	35	351
61	iris	3	5	150
1073	jEdit_4.0_4.2	2	9	274
1066	kc1-binary	2	95	145
1045	kc1-top5	2	95	145
1067	kc1	2	22	2109
1063	kc2	2	22	522
1065	kc3	2	40	458
807	kin8nm	2	9	8192
1056	mc1	2	39	9466
1054	mc2	2	40	161
1488	parkinsons	2	23	195
750	pm10	2	8	500
871	pollen	2	6	3848
446	prnn_crabs	2	8	200
952	prnn_fglass	6	10	214
470	profb	2	10	672
752	puma32H	2	33	8192
816	puma8NH	2	9	8192
721	pwLinear	2	11	200
1495	qualitative-bankruptcy	2	7	250
717	rmftsa_ladata	2	11	508
1519	robot-failures-lp4	3	91	117
1520	robot-failures-lp5	5	91	164
1498	sa-heart	2	10	462
182	satimage	6	37	6430
466	schizo	2	15	340
1499	seeds	3	8	210
36	segment	7	20	2310
1500	seismic-bumps	3	8	210
826	sensory	2	12	576
747	servo	2	5	167
902	sleuth_case2002	2	7	147
934	socmob	2	6	1156
40	sonar	2	61	208
42	soybean	19	36	683
737	space_ga	2	7	3107
44	spambase	2	58	4601
46	splice	3	61	3190
40982	steel-plates-fault	7	28	1941
841	stock	2	10	950
770	strikes	2	7	625
41146	sylvine	2	21	5124
1004	synthetic_control	2	61	600

ID	Name	Nº of Classes	Nº of Attribute	Nº of Instances
48	tae	3	6	151
1115	teachingAssistant	3	7	151
41526	test_dataset	2	61	15547
788	triazines	2	61	186
1507	twonorm	2	21	7400
1508	user-knowledge	5	6	403
54	vehicle	4	19	846
719	veteran	2	8	137
925	visualizing_galaxy	2	5	323
923	visualizing_soil	2	5	8641
56	vote	2	17	435
1016	vowel	2	14	990
1497	wall-robot-navigation	4	25	5456
60	waveform-5000	3	41	5000
1510	wdbc	2	31	569
1511	wholesale-customers	2	9	440
40983	wilt	2	6	4839
847	wind	2	15	6574
187	wine	3	14	178
753	wisconsin	2	33	194
181	yeast	10	9	1484
62	zoo	7	17	101