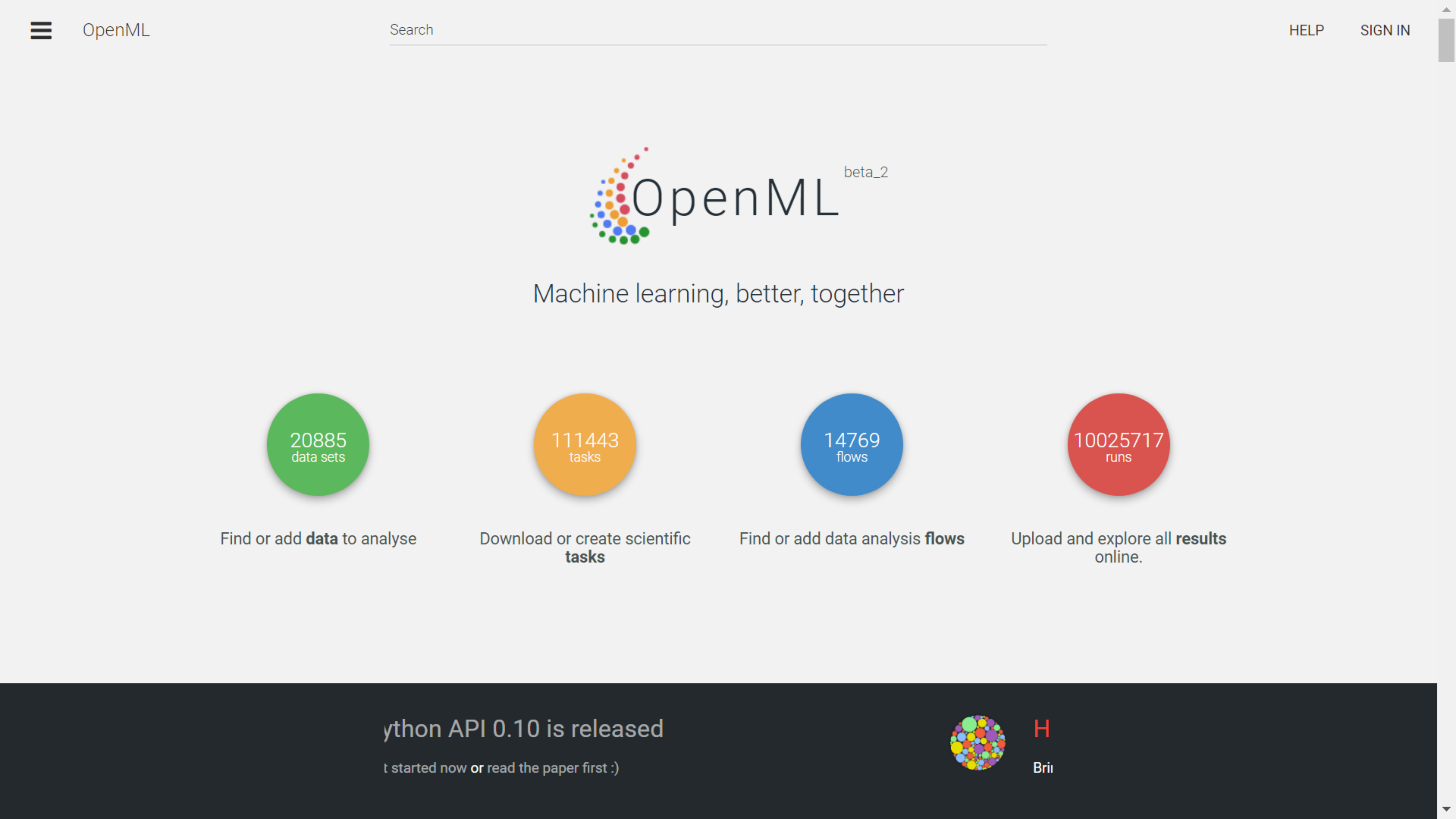
OpenML, metafeatures and analyzing the results of experiments with filtering anomalies

Katarína Švecová

PV115

2020

OpenML



OpenML beta_2

Machine learning, better, together

20885 data sets

Find or add **data** to analyse

111443 tasks

Download or create scientific **tasks**

14769 flows

Find or add data analysis **flows**

10025717 runs

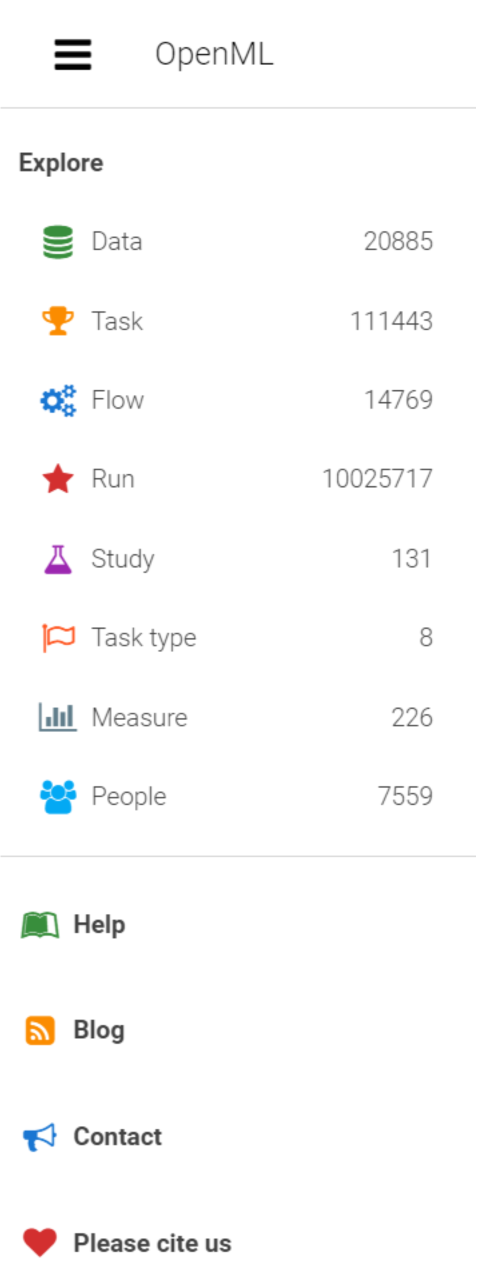
Upload and explore all **results** online.

ython API 0.10 is released

t started now **or** read the paper first :)

H

Bri



OpenML

Explore

Data	20885
Task	111443
Flow	14769
Run	10025717
Study	131
Task type	8
Measure	226
People	7559

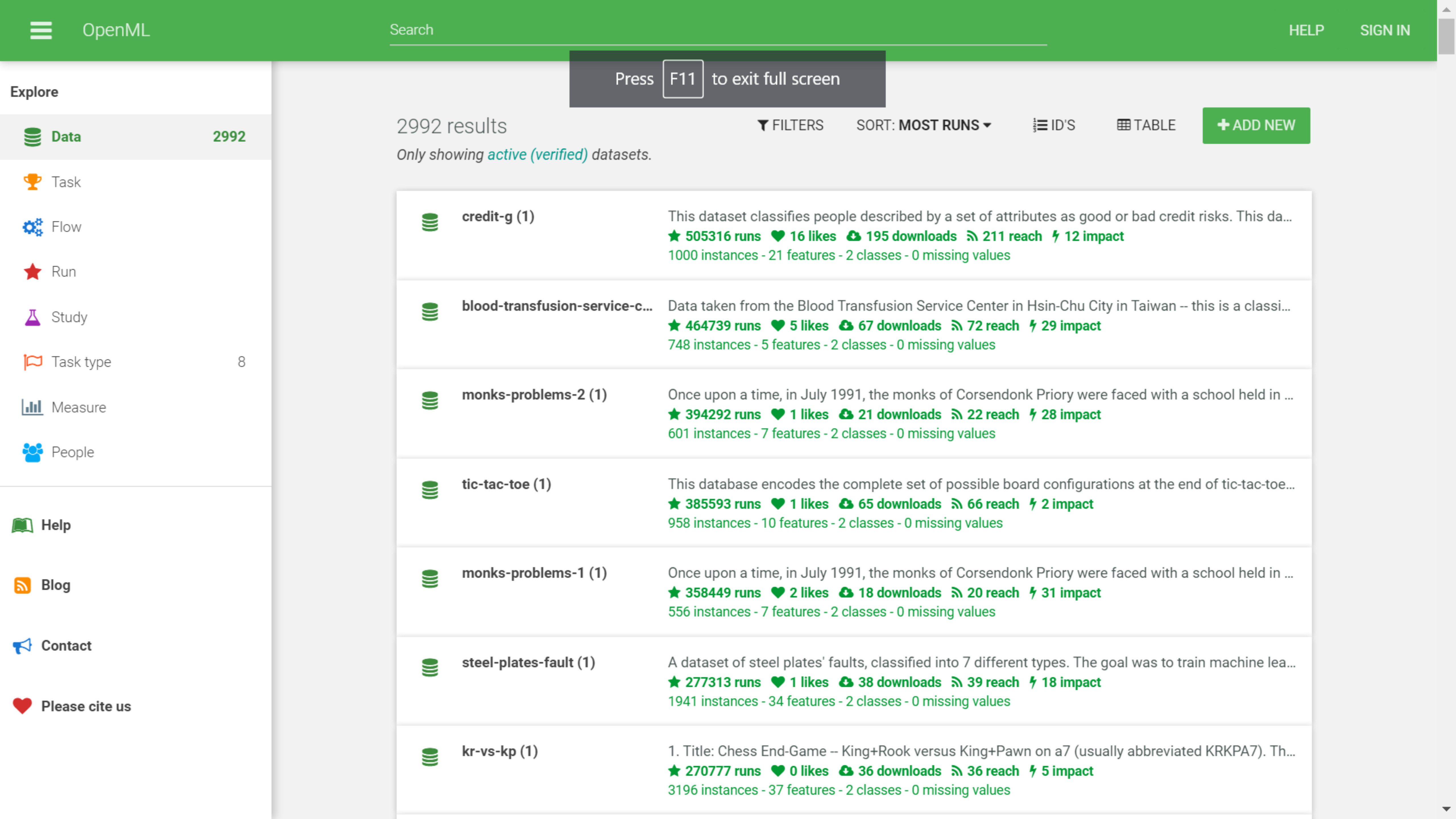
Help

Blog

Contact

Please cite us

- Main goal – to make machine learning accesible
- open source project on GitHub
- Datasets, Tasks, Flows, Runs
- Study, task type, measure
- API (REST, Python, R, Java, C#) – enable downloading datasets, tasks and sharing results



OpenML Search HELP SIGN IN

Press F11 to exit full screen

Explore

Data 2992

Task

Flow

Run

Study

Task type 8

Measure

People

Help

Blog

Contact

Please cite us

2992 results

FILTERS SORT: MOST RUNS ID'S TABLE + ADD NEW

Only showing active (verified) datasets.

credit-g (1)
This dataset classifies people described by a set of attributes as good or bad credit risks. This da...
505316 runs 16 likes 195 downloads 211 reach 12 impact
1000 instances - 21 features - 2 classes - 0 missing values

blood-transfusion-service-c...
Data taken from the Blood Transfusion Service Center in Hsin-Chu City in Taiwan -- this is a classi...
464739 runs 5 likes 67 downloads 72 reach 29 impact
748 instances - 5 features - 2 classes - 0 missing values

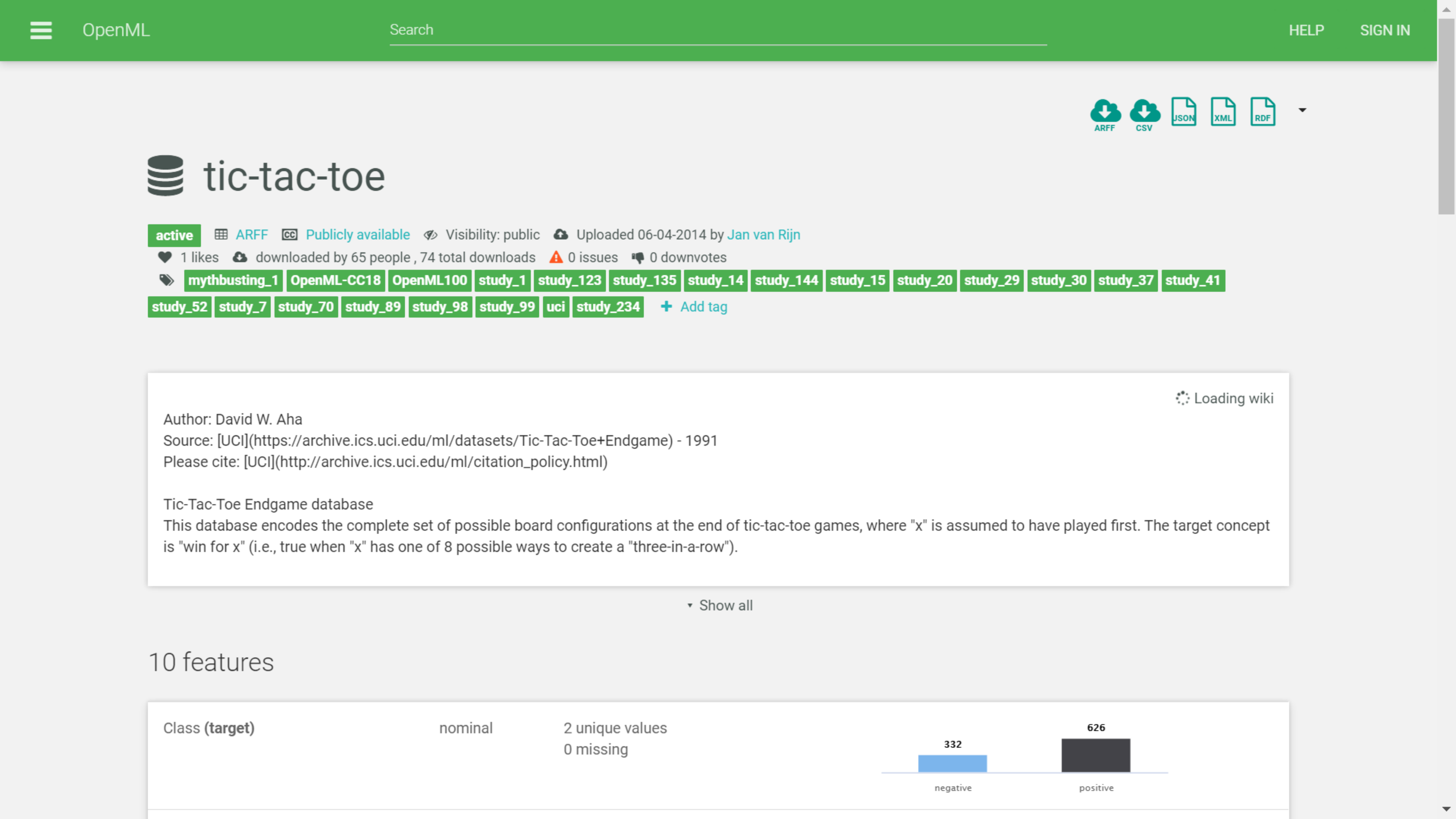
monks-problems-2 (1)
Once upon a time, in July 1991, the monks of Corsendonk Priory were faced with a school held in ...
394292 runs 1 likes 21 downloads 22 reach 28 impact
601 instances - 7 features - 2 classes - 0 missing values

tic-tac-toe (1)
This database encodes the complete set of possible board configurations at the end of tic-tac-toe...
385593 runs 1 likes 65 downloads 66 reach 2 impact
958 instances - 10 features - 2 classes - 0 missing values

monks-problems-1 (1)
Once upon a time, in July 1991, the monks of Corsendonk Priory were faced with a school held in ...
358449 runs 2 likes 18 downloads 20 reach 31 impact
556 instances - 7 features - 2 classes - 0 missing values

steel-plates-fault (1)
A dataset of steel plates' faults, classified into 7 different types. The goal was to train machine lea...
277313 runs 1 likes 38 downloads 39 reach 18 impact
1941 instances - 34 features - 2 classes - 0 missing values

kr-vs-kp (1)
1. Title: Chess End-Game -- King+Rook versus King+Pawn on a7 (usually abbreviated KRKPA7). Th...
270777 runs 0 likes 36 downloads 36 reach 5 impact
3196 instances - 37 features - 2 classes - 0 missing values



tic-tac-toe

active ARFF Publicly available Visibility: public Uploaded 06-04-2014 by Jan van Rijn

1 likes downloaded by 65 people , 74 total downloads 0 issues 0 downvotes

mythbusting_1 OpenML-CC18 OpenML100 study_1 study_123 study_135 study_14 study_144 study_15 study_20 study_29 study_30 study_37 study_41 study_52 study_7 study_70 study_89 study_98 study_99 uci study_234 + Add tag

Loading wiki

Author: David W. Aha
Source: [UCI](https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame) - 1991
Please cite: [UCI](http://archive.ics.uci.edu/ml/citation_policy.html)

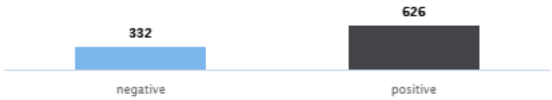
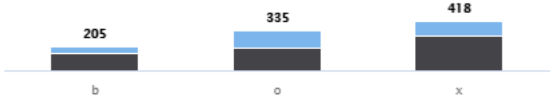
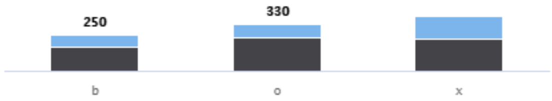
Tic-Tac-Toe Endgame database
This database encodes the complete set of possible board configurations at the end of tic-tac-toe games, where "x" is assumed to have played first. The target concept is "win for x" (i.e., true when "x" has one of 8 possible ways to create a "three-in-a-row").

▾ Show all

10 features

Class **(target)** nominal 2 unique values
0 missing

10 features

Class **(target)**	nominal	2 unique values 0 missing	
top-left-square	nominal	3 unique values 0 missing	
top-middle-square	nominal	3 unique values 0 missing	

Show all 10 features

107 properties

NumberOfInstances	**958**	Number of instances (rows) of the dataset.
NumberOfFeatures	**10**	Number of attributes (columns) of the dataset.
NumberOfClasses	**2**	Number of distinct values of the target attribute (if it is nominal).
NumberOfMissingVal...	**0**	Number of missing values in the dataset.
NumberOfInstancesW...	**0**	Number of instances with at least one value missing.

Show all 107 properties

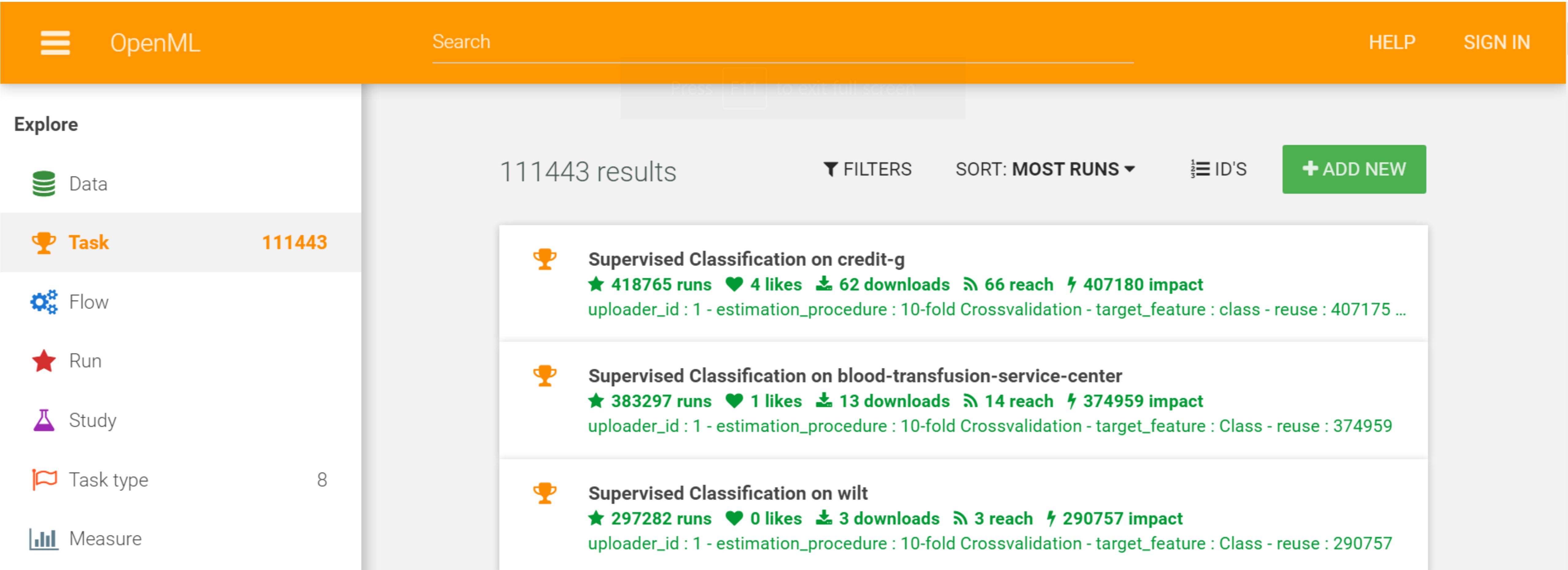
23 tasks

Supervised Classification on tic-tac-toe
270253 runs - estimation_procedure: 10-fold Crossvalidation - target_feature: Class

Supervised Classification on tic-tac-toe

Tasks

- Dataset with a specific task – clustering/classification and a method of evaluation



OpenML

Search

HELP

SIGN IN

Explore

Data

Task 111443

Flow

Run

Study

Task type 8

Measure

111443 results

FILTERS

SORT: **MOST RUNS**

ID'S

ADD NEW

Supervised Classification on credit-g

418765 runs 4 likes 62 downloads 66 reach 407180 impact

uploader_id : 1 - estimation_procedure : 10-fold Crossvalidation - target_feature : class - reuse : 407175 ...

Supervised Classification on blood-transfusion-service-center

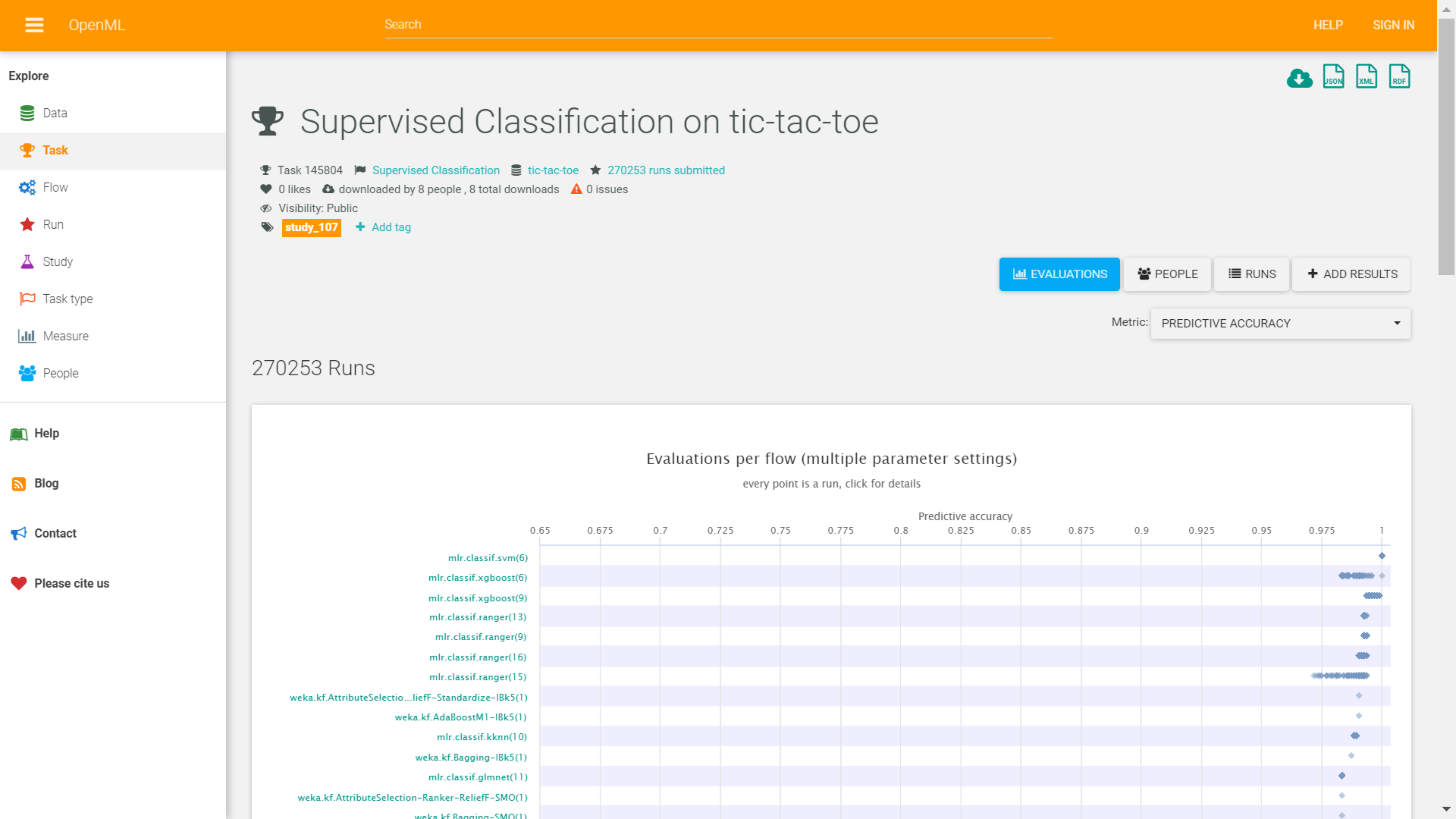
383297 runs 1 likes 13 downloads 14 reach 374959 impact

uploader_id : 1 - estimation_procedure : 10-fold Crossvalidation - target_feature : Class - reuse : 374959

Supervised Classification on wilt

297282 runs 0 likes 3 downloads 3 reach 290757 impact

uploader_id : 1 - estimation_procedure : 10-fold Crossvalidation - target_feature : Class - reuse : 290757



OpenML Search HELP SIGN IN

Explore

Data

Task

Flow

Run

Study

Task type

Measure

People

Help

Blog

Contact

Please cite us

JSON XML RDF

Supervised Classification on tic-tac-toe

Task 145804 Supervised Classification tic-tac-toe 270253 runs submitted

0 likes downloaded by 8 people , 8 total downloads 0 issues

Visibility: Public

study_107 + Add tag

EVALUATIONS PEOPLE RUNS + ADD RESULTS

Metric: PREDICTIVE ACCURACY

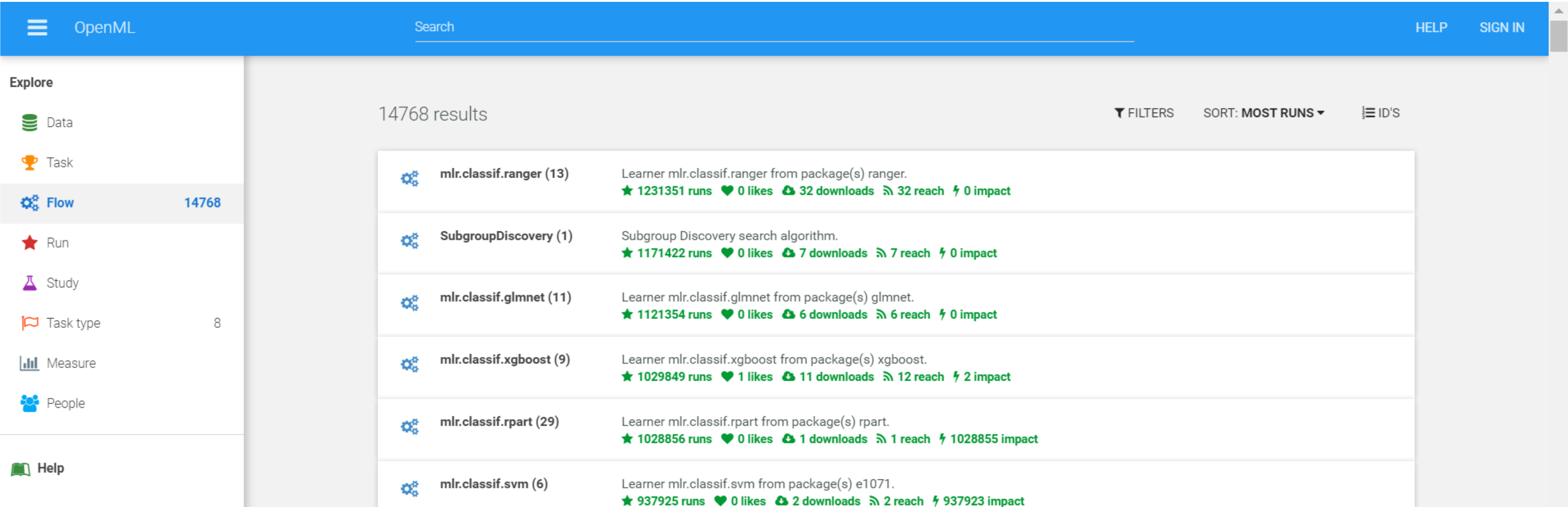
270253 Runs

Evaluations per flow (multiple parameter settings)

every point is a run, click for details

Flows

- Flow – a specific algorithm in a specific implementation



14768 results

mlr.classif.ranger (13) Learner mlr.classif.ranger from package(s) ranger.
1231351 runs 0 likes 32 downloads 32 reach 0 impact

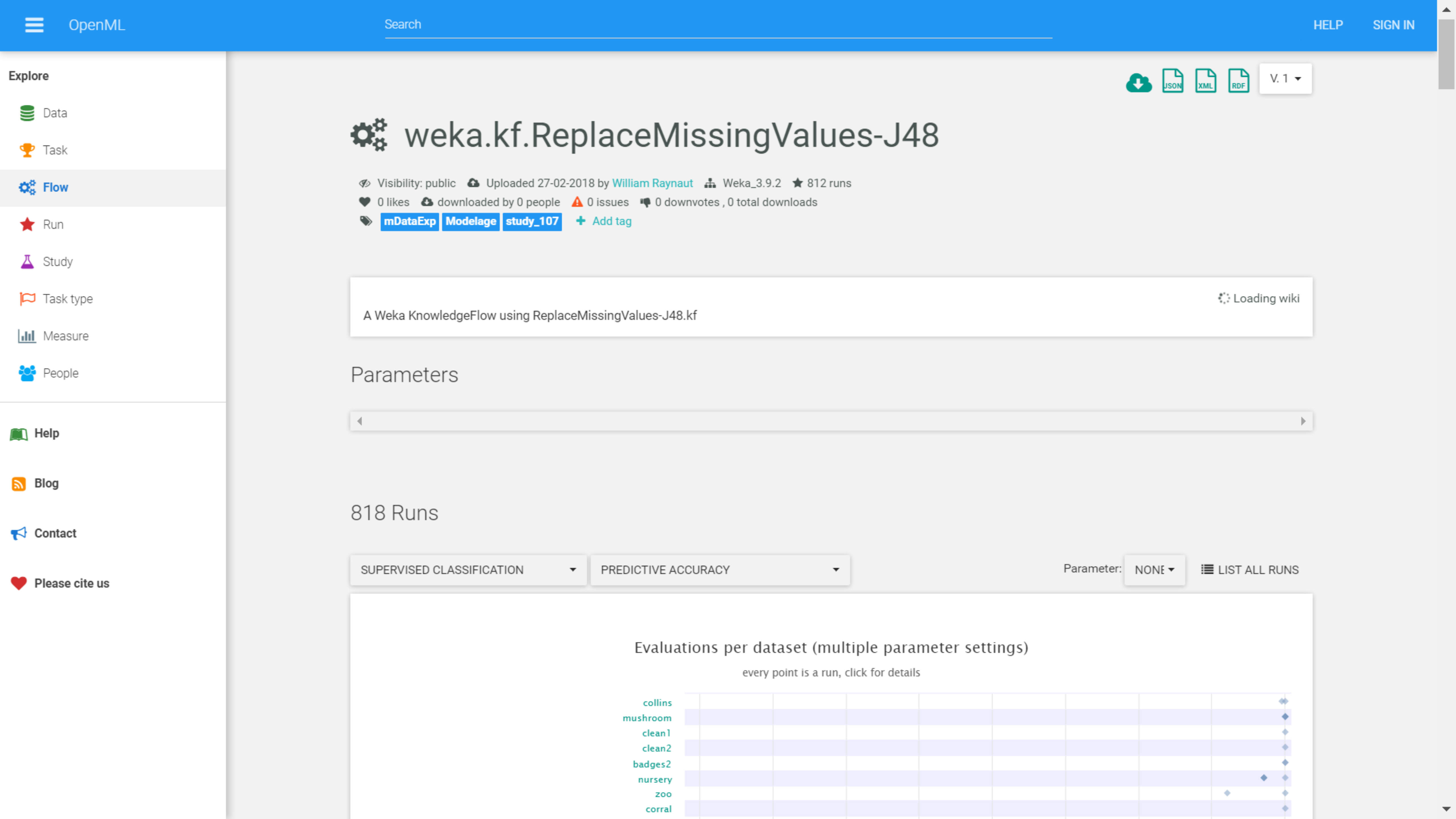
SubgroupDiscovery (1) Subgroup Discovery search algorithm.
1171422 runs 0 likes 7 downloads 7 reach 0 impact

mlr.classif.glmnet (11) Learner mlr.classif.glmnet from package(s) glmnet.
1121354 runs 0 likes 6 downloads 6 reach 0 impact

mlr.classif.xgboost (9) Learner mlr.classif.xgboost from package(s) xgboost.
1029849 runs 1 likes 11 downloads 12 reach 2 impact

mlr.classif.rpart (29) Learner mlr.classif.rpart from package(s) rpart.
1028856 runs 0 likes 1 downloads 1 reach 1028855 impact

mlr.classif.svm (6) Learner mlr.classif.svm from package(s) e1071.
937925 runs 0 likes 2 downloads 2 reach 937923 impact



weka.kf.ReplaceMissingValues-J48

Visibility: public Uploaded 27-02-2018 by William Raynaut Weka_3.9.2 812 runs

0 likes downloaded by 0 people 0 issues 0 downvotes , 0 total downloads

mDataExp Modelage study_107 + Add tag

Loading wiki

A Weka KnowledgeFlow using ReplaceMissingValues-J48.kf

Parameters

818 Runs

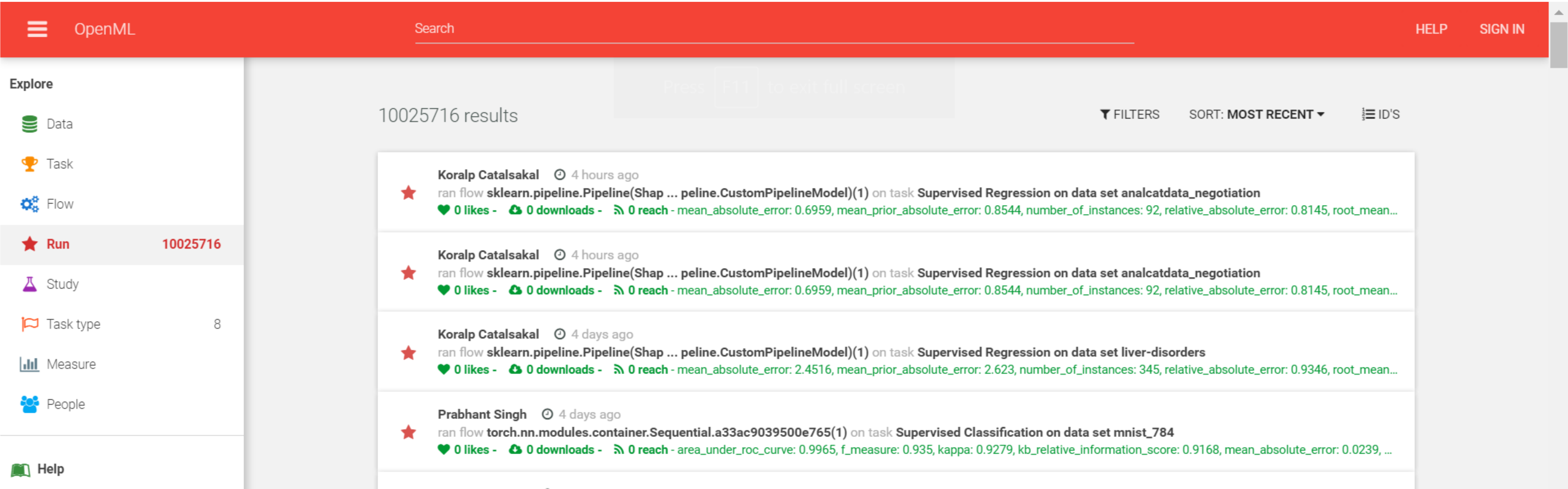
SUPERVISED CLASSIFICATION ▾ PREDICTIVE ACCURACY ▾ Parameter: NONE ▾ LIST ALL RUNS

Evaluations per dataset (multiple parameter settings)

every point is a run, click for details

Run

- A specific run with specific settings



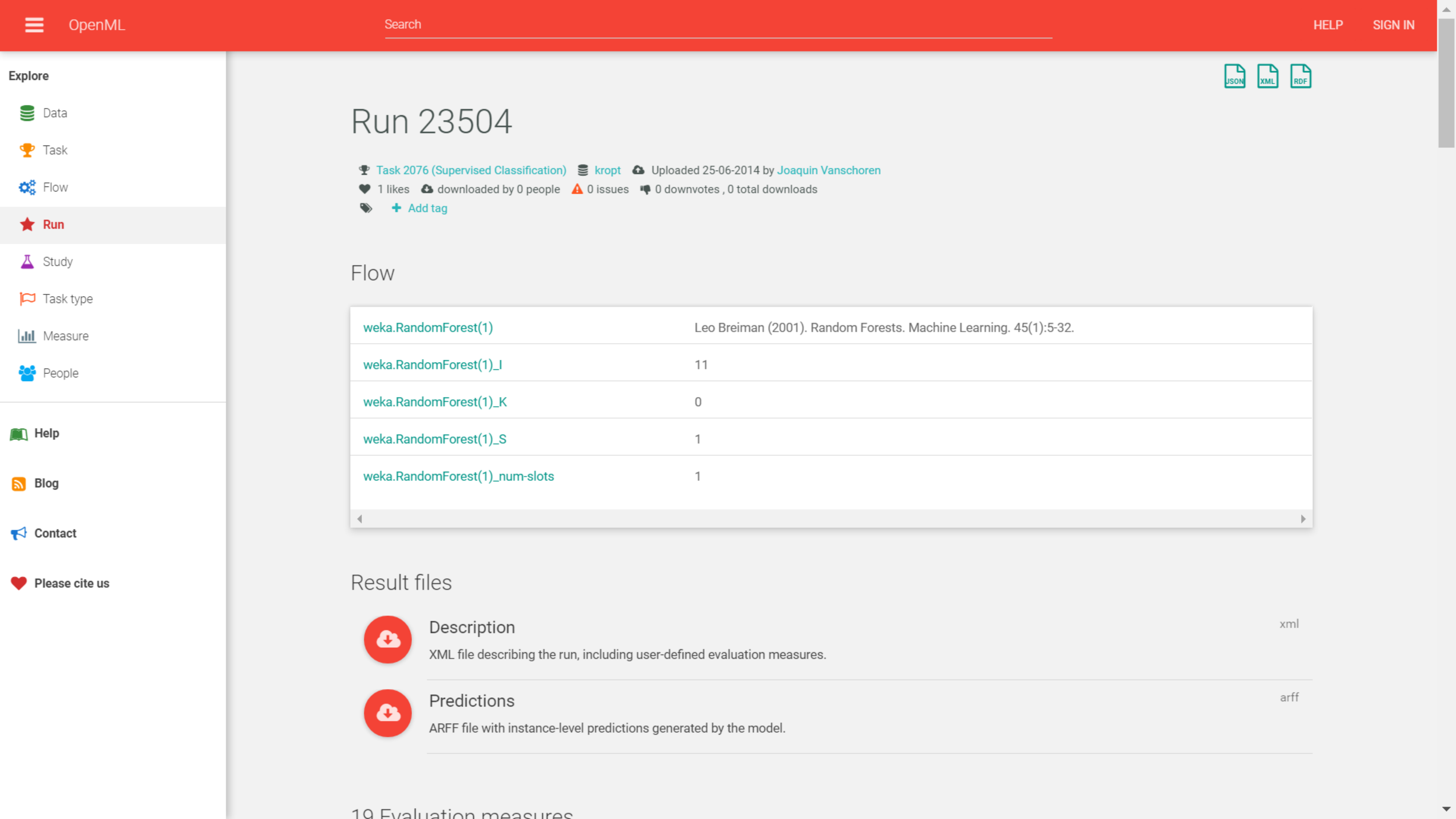
10025716 results

Koralp Catalsakal 4 hours ago
ran flow sklearn.pipeline.Pipeline(Shap ... peline.CustomPipelineModel)(1) on task Supervised Regression on data set analcatdata_negotiation
0 likes - 0 downloads - 0 reach - mean_absolute_error: 0.6959, mean_prior_absolute_error: 0.8544, number_of_instances: 92, relative_absolute_error: 0.8145, root_mean...

Koralp Catalsakal 4 hours ago
ran flow sklearn.pipeline.Pipeline(Shap ... peline.CustomPipelineModel)(1) on task Supervised Regression on data set analcatdata_negotiation
0 likes - 0 downloads - 0 reach - mean_absolute_error: 0.6959, mean_prior_absolute_error: 0.8544, number_of_instances: 92, relative_absolute_error: 0.8145, root_mean...

Koralp Catalsakal 4 days ago
ran flow sklearn.pipeline.Pipeline(Shap ... peline.CustomPipelineModel)(1) on task Supervised Regression on data set liver-disorders
0 likes - 0 downloads - 0 reach - mean_absolute_error: 2.4516, mean_prior_absolute_error: 2.623, number_of_instances: 345, relative_absolute_error: 0.9346, root_mean...

Prabhant Singh 4 days ago
ran flow torch.nn.modules.container.Sequential.a33ac9039500e765(1) on task Supervised Classification on data set mnist_784
0 likes - 0 downloads - 0 reach - area_under_roc_curve: 0.9965, f_measure: 0.935, kappa: 0.9279, kb_relative_information_score: 0.9168, mean_absolute_error: 0.0239, ...



Run 23504

Task 2076 (Supervised Classification) kropt Uploaded 25-06-2014 by Joaquin Vanschoren
1 likes downloaded by 0 people 0 issues 0 downvotes , 0 total downloads
Add tag

Flow

weka.RandomForest(1)	Leo Breiman (2001). Random Forests. Machine Learning. 45(1):5-32.
weka.RandomForest(1)_I	11
weka.RandomForest(1)_K	0
weka.RandomForest(1)_S	1
weka.RandomForest(1)_num-slots	1

Result files

Description

xml

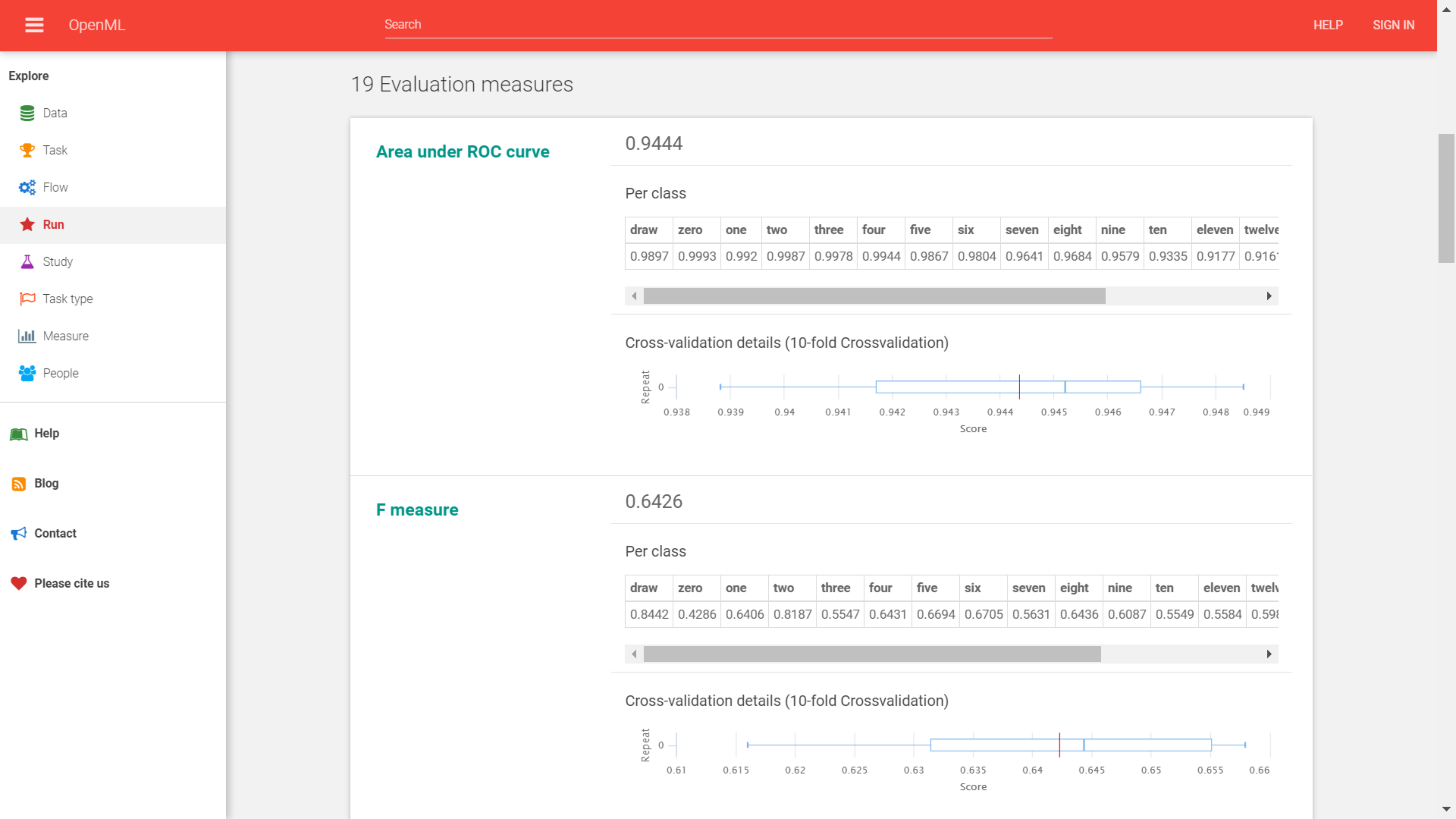
XML file describing the run, including user-defined evaluation measures.

Predictions

arff

ARFF file with instance-level predictions generated by the model.

19 Evaluation measures



19 Evaluation measures

Area under ROC curve

0.9444

Per class

draw	zero	one	two	three	four	five	six	seven	eight	nine	ten	eleven	twelve
0.9897	0.9993	0.992	0.9987	0.9978	0.9944	0.9867	0.9804	0.9641	0.9684	0.9579	0.9335	0.9177	0.916

Cross-validation details (10-fold Crossvalidation)

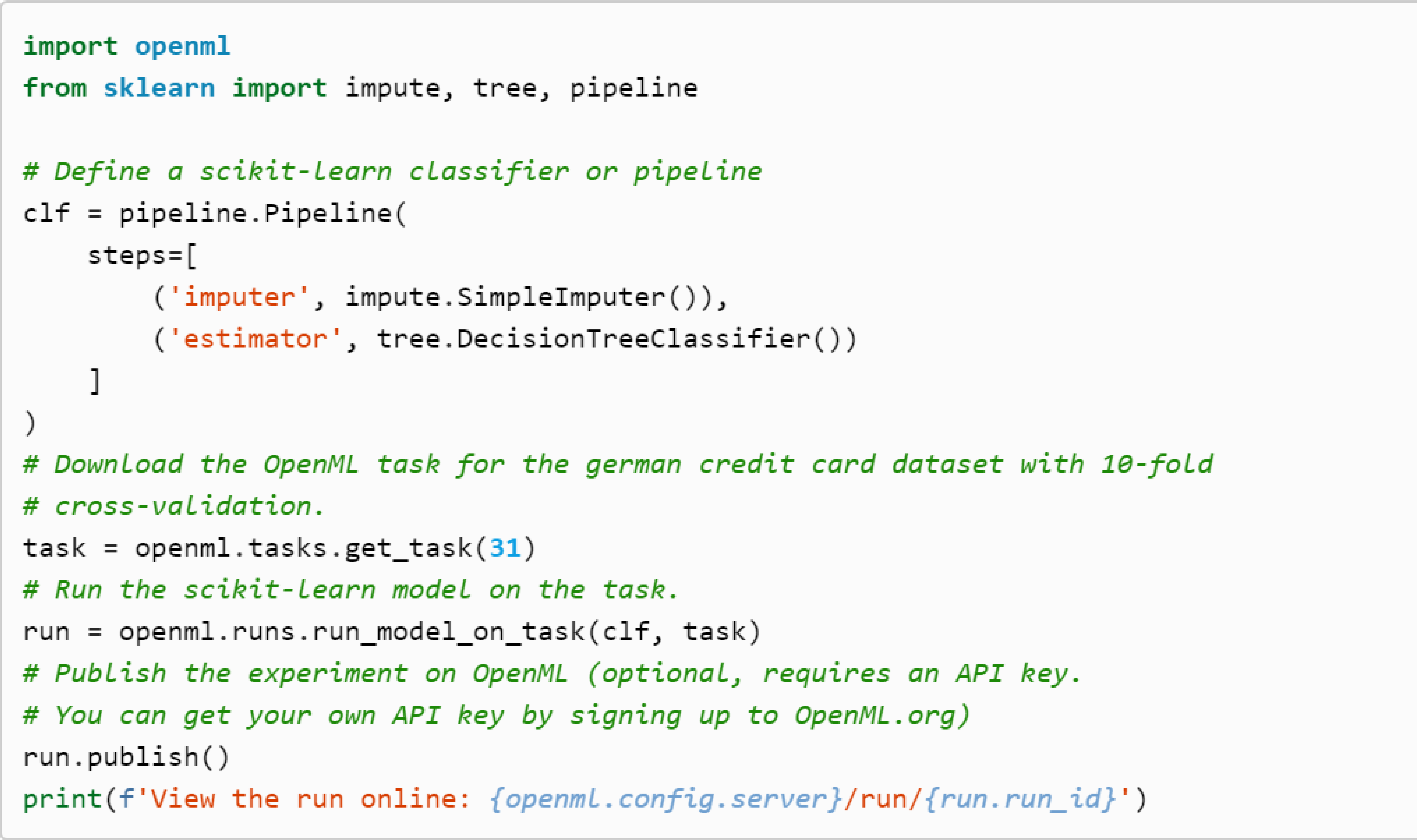
F measure

0.6426

Per class

draw	zero	one	two	three	four	five	six	seven	eight	nine	ten	eleven	twel
0.8442	0.4286	0.6406	0.8187	0.5547	0.6431	0.6694	0.6705	0.5631	0.6436	0.6087	0.5549	0.5584	0.59

Cross-validation details (10-fold Crossvalidation)



```python
import openml
from sklearn import impute, tree, pipeline

# Define a scikit-learn classifier or pipeline
clf = pipeline.Pipeline(
    steps=[
        ('imputer', impute.SimpleImputer()),
        ('estimator', tree.DecisionTreeClassifier())
    ]
)
# Download the OpenML task for the german credit card dataset with 10-fold
# cross-validation.
task = openml.tasks.get_task(31)
# Run the scikit-learn model on the task.
run = openml.runs.run_model_on_task(clf, task)
# Publish the experiment on OpenML (optional, requires an API key.
# You can get your own API key by signing up to OpenML.org)
run.publish()
print(f'View the run online: {openml.config.server}/run/{run.run_id}')
```

Metafeatures

Types of metafeatures

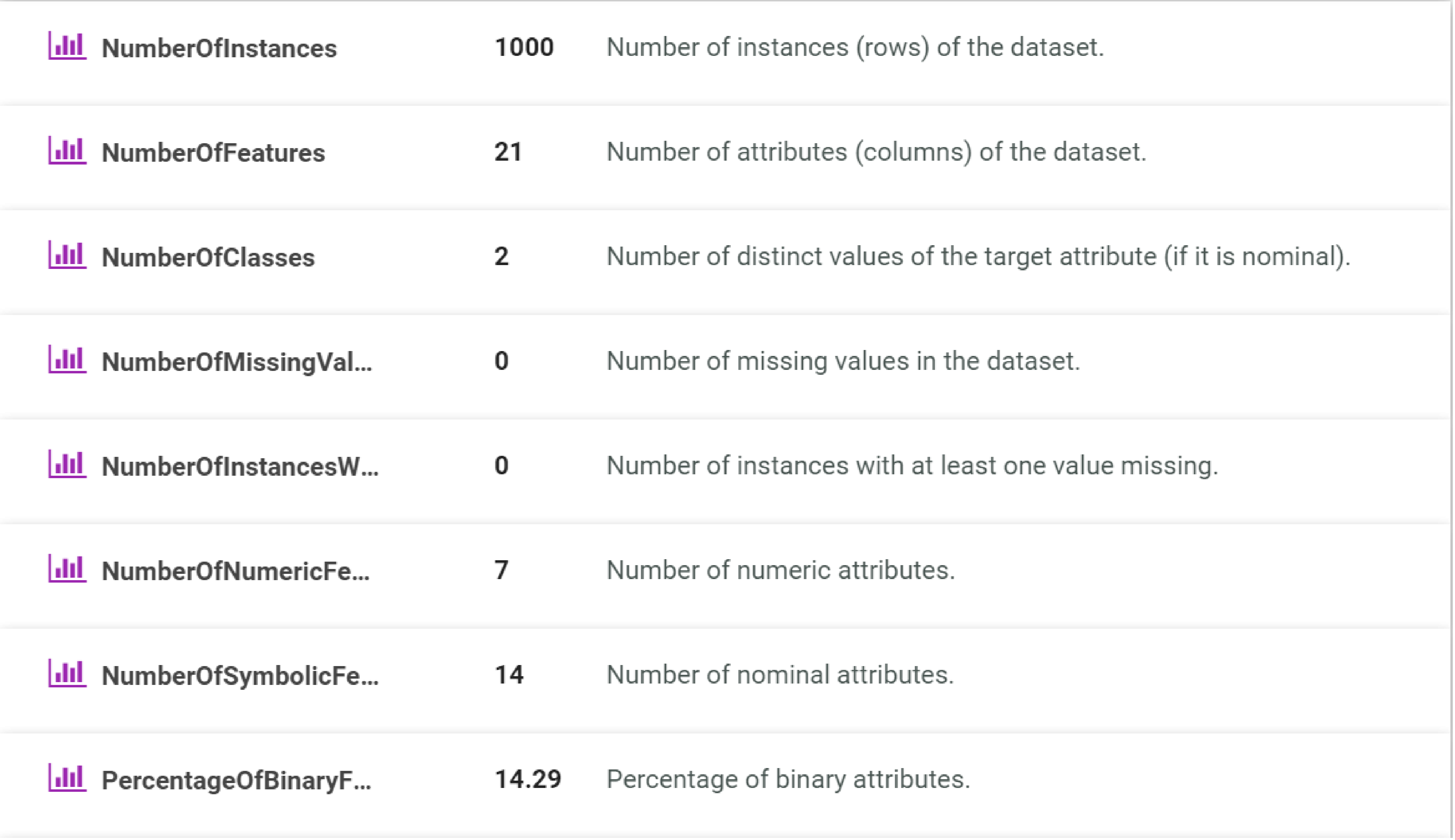
- Basic data description (number of instances, dimensionality, number of classes, majority class percentage...)
- Statistical methods (mean, median, skewness, kurtosis...)
- Landmarking – results with (quite a few) selected methods (

  random tree **depth**,

  DecisionStumpKappa, NaiveBayes**Kappa -** the agreement between two raters, similar to accuracy, but considering the probability of a chance agreement,

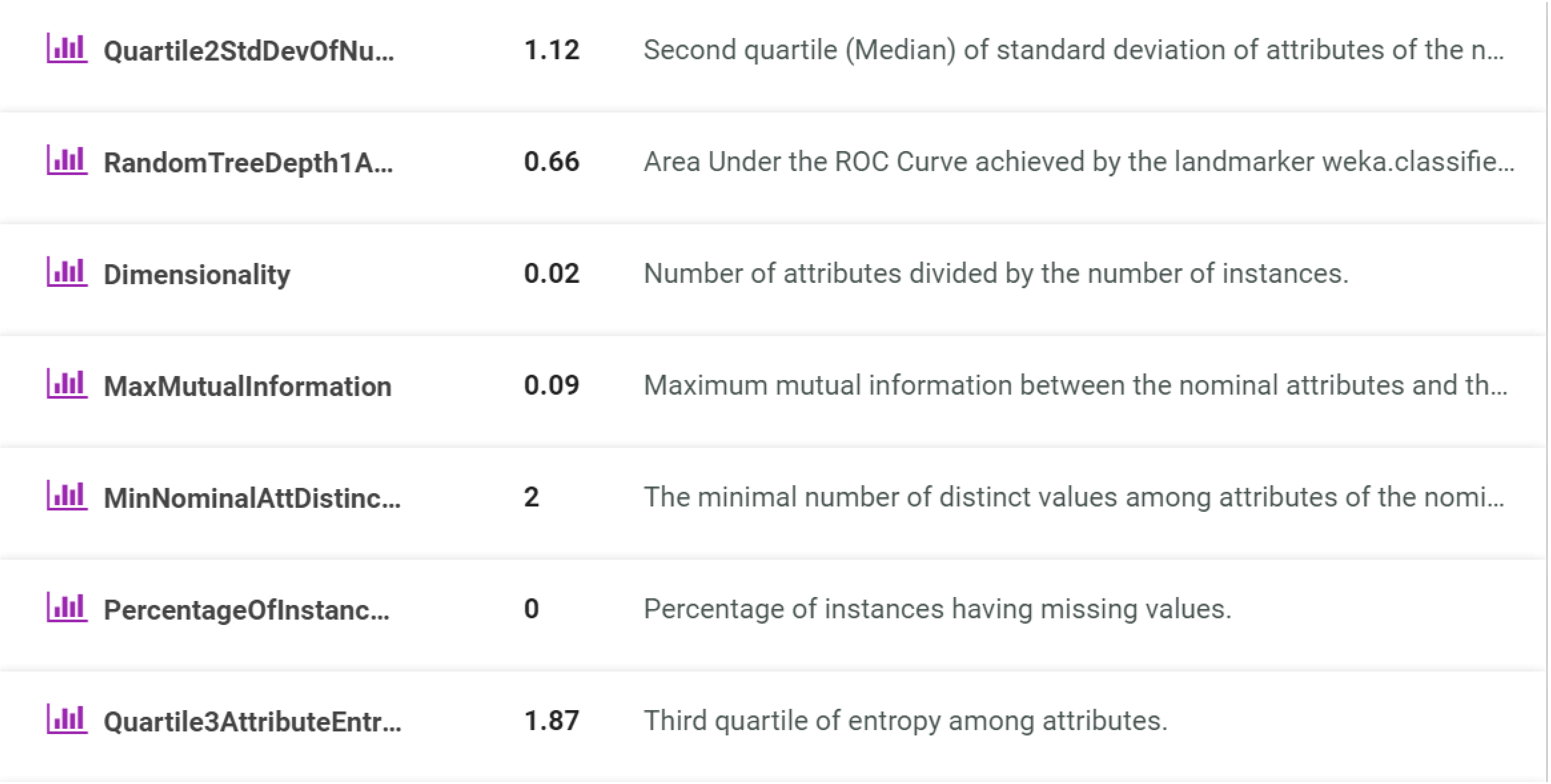
  J48 **error rate**,

  Area **U**nder the ROC **C**urve, which is made by plotting true positive rate and false positive rate)

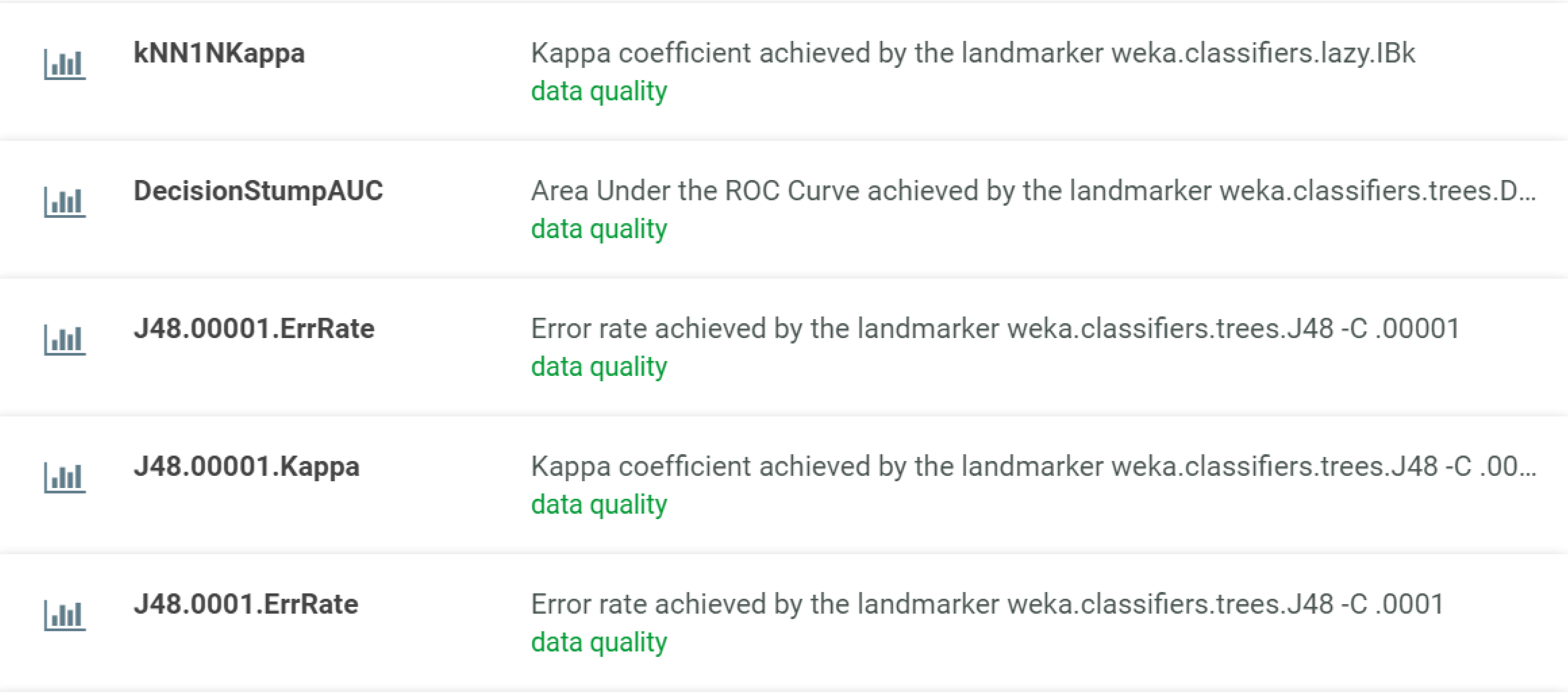


NumberOfInstances	1000	Number of instances (rows) of the dataset.
NumberOfFeatures	21	Number of attributes (columns) of the dataset.
NumberOfClasses	2	Number of distinct values of the target attribute (if it is nominal).
NumberOfMissingVal...	0	Number of missing values in the dataset.
NumberOfInstancesW...	0	Number of instances with at least one value missing.
NumberOfNumericFe...	7	Number of numeric attributes.
NumberOfSymbolicFe...	14	Number of nominal attributes.
PercentageOfBinaryF...	14.29	Percentage of binary attributes.

Number of different things



Quartile2StdDevOfNu...	**1.12**	Second quartile (Median) of standard deviation of attributes of the n...
RandomTreeDepth1A...	**0.66**	Area Under the ROC Curve achieved by the landmarker weka.classifie...
Dimensionality	**0.02**	Number of attributes divided by the number of instances.
MaxMutualInformation	**0.09**	Maximum mutual information between the nominal attributes and th...
MinNominalAttDistinc...	**2**	The minimal number of distinct values among attributes of the nomi...
PercentageOfInstanc...	**0**	Percentage of instances having missing values.
Quartile3AttributeEntr...	**1.87**	Third quartile of entropy among attributes.

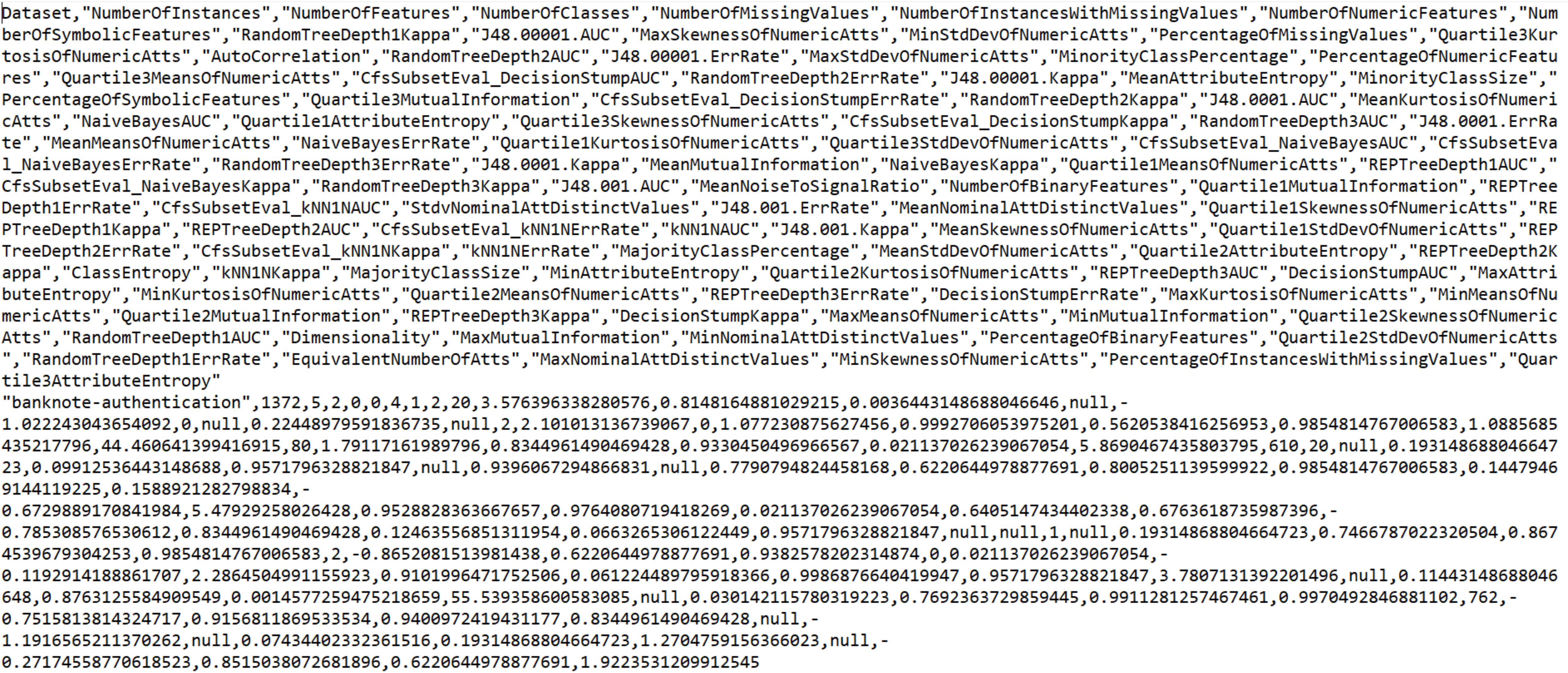


kNN1NKappa	Kappa coefficient achieved by the landmarker weka.classifiers.lazy.IBk data quality
DecisionStumpAUC	Area Under the ROC Curve achieved by the landmarker weka.classifiers.trees.D... data quality
J48.00001.ErrRate	Error rate achieved by the landmarker weka.classifiers.trees.J48 -C .00001 data quality
J48.00001.Kappa	Kappa coefficient achieved by the landmarker weka.classifiers.trees.J48 -C .00... data quality
J48.0001.ErrRate	Error rate achieved by the landmarker weka.classifiers.trees.J48 -C .0001 data quality

On the predictive power of meta-features in OpenML

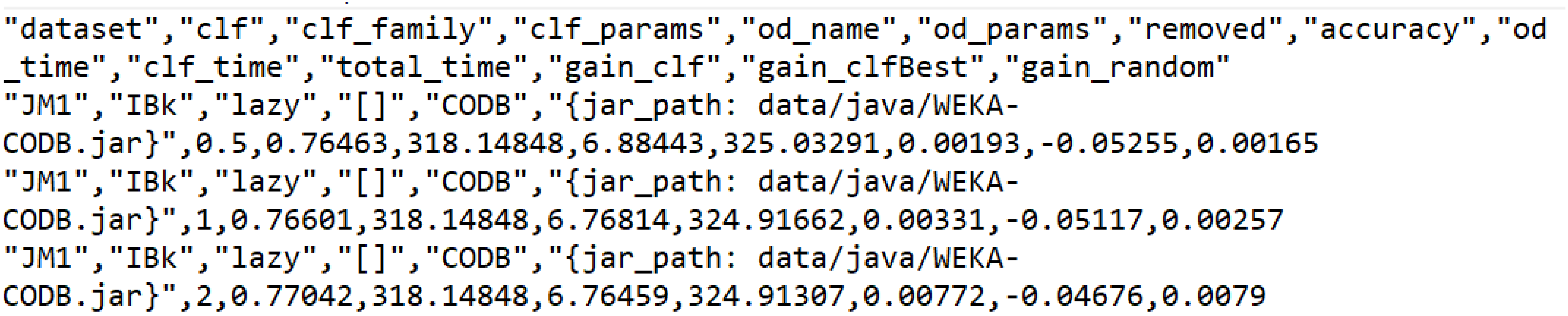
- 2017
- Study on 61 metafeatures and 720 datasets
- Feature selection gains better results
- Information on response, mutual information, noise to signal, shape of extremes, **dimensionality,** minimum variability of numeric attributes, information of categorical attributes and mutual information

Analyzing the results of experiments with filtering anomalies

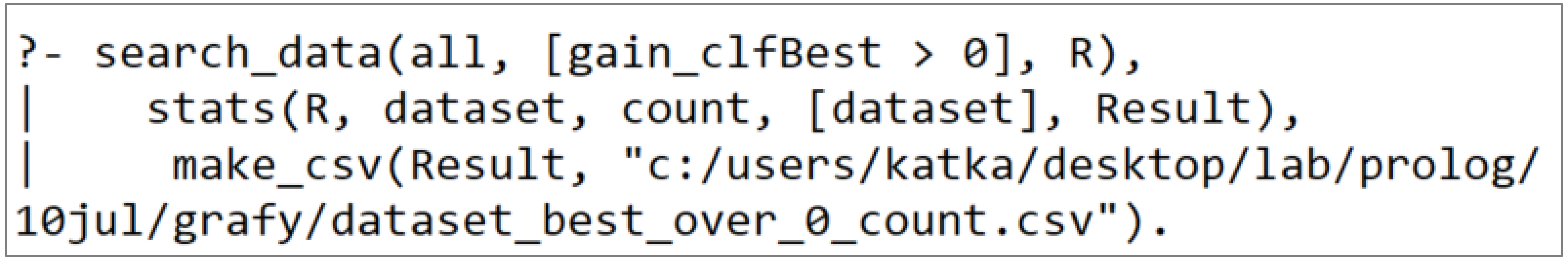


```
Dataset,"NumberOfInstances","NumberOfFeatures","NumberOfClasses","NumberOfMissingValues","NumberOfInstancesWithMissingValues","NumberOfNumericFeatures","Num
berOfSymbolicFeatures","RandomTreeDepth1Kappa","J48.00001.AUC","MaxSkewnessOfNumericAtts","MinStdDevOfNumericAtts","PercentageOfMissingValues","Quartile3Kur
tosisOfNumericAtts","AutoCorrelation","RandomTreeDepth2AUC","J48.00001.ErrRate","MaxStdDevOfNumericAtts","MinorityClassPercentage","PercentageOfNumericFeatu
res","Quartile3MeansOfNumericAtts","CfsSubsetEval_DecisionStumpAUC","RandomTreeDepth2ErrRate","J48.00001.Kappa","MeanAttributeEntropy","MinorityClassSize","
PercentageOfSymbolicFeatures","Quartile3MutualInformation","CfsSubsetEval_DecisionStumpErrRate","RandomTreeDepth2Kappa","J48.0001.AUC","MeanKurtosisOfNumeri
cAtts","NaiveBayesAUC","Quartile1AttributeEntropy","Quartile3SkewnessOfNumericAtts","CfsSubsetEval_DecisionStumpKappa","RandomTreeDepth3AUC","J48.0001.ErrRa
te","MeanMeansOfNumericAtts","NaiveBayesErrRate","Quartile1KurtosisOfNumericAtts","Quartile3StdDevOfNumericAtts","CfsSubsetEval_NaiveBayesAUC","CfsSubsetEva
l_NaiveBayesErrRate","RandomTreeDepth3ErrRate","J48.0001.Kappa","MeanMutualInformation","NaiveBayesKappa","Quartile1MeansOfNumericAtts","REPTreeDepth1AUC","
CfsSubsetEval_NaiveBayesKappa","RandomTreeDepth3Kappa","J48.001.AUC","MeanNoiseToSignalRatio","NumberOfBinaryFeatures","Quartile1MutualInformation","REPTree
Depth1ErrRate","CfsSubsetEval_kNN1NAUC","StdvNominalAttDistinctValues","J48.001.ErrRate","MeanNominalAttDistinctValues","Quartile1SkewnessOfNumericAtts","RE
PTreeDepth1Kappa","REPTreeDepth2AUC","CfsSubsetEval_kNN1NErrRate","kNN1NAUC","J48.001.Kappa","MeanSkewnessOfNumericAtts","Quartile1StdDevOfNumericAtts","REP
TreeDepth2ErrRate","CfsSubsetEval_kNN1NKappa","kNN1NErrRate","MajorityClassPercentage","MeanStdDevOfNumericAtts","Quartile2AttributeEntropy","REPTreeDepth2K
appa","ClassEntropy","kNN1NKappa","MajorityClassSize","MinAttributeEntropy","Quartile2KurtosisOfNumericAtts","REPTreeDepth3AUC","DecisionStumpAUC","MaxAttri
buteEntropy","MinKurtosisOfNumericAtts","Quartile2MeansOfNumericAtts","REPTreeDepth3ErrRate","DecisionStumpErrRate","MaxKurtosisOfNumericAtts","MinMeansOfNu
mericAtts","Quartile2MutualInformation","REPTreeDepth3Kappa","DecisionStumpKappa","MaxMeansOfNumericAtts","MinMutualInformation","Quartile2SkewnessOfNumeric
Atts","RandomTreeDepth1AUC","Dimensionality","MaxMutualInformation","MinNominalAttDistinctValues","PercentageOfBinaryFeatures","Quartile2StdDevOfNumericAtts
","RandomTreeDepth1ErrRate","EquivalentNumberOfAtts","MaxNominalAttDistinctValues","MinSkewnessOfNumericAtts","PercentageOfInstancesWithMissingValues","Quar
tile3AttributeEntropy"
"banknote-authentication",1372,5,2,0,0,4,1,2,20,3.576396338280576,0.8148164881029215,0.003644314868804664 6,null,-
1.02224304365 4092,0,null,0.2244897959183673 5,null,2,2.10101313673906 7,0,1.077230875627456,0.9992706053975201,0.562053841625695 3,0.985481476700658 3,1.0885685
435217796,44.46064139941691 5,80,1.7911716198979 6,0.8344961490469428,0.9330450496966567,0.0211370262390670 54,5.869046743580379 5,610,20,null,0.193148688046647
23,0.0991253644314868 8,0.9571796328821847,null,0.9396067294866831,null,0.7790794824458168,0.6220644978877691,0.8005251139599922,0.9854814767006583,0.1447946
9144119225,0.158892128279 8834,-
0.6729889170841984,5.47929258026428,0.9528828363667657,0.9764080719418269,0.021137026239067054,0.6405147434402338,0.6763618735987396,-
0.785308576530612,0.8344961490469428,0.1246355685131195 4,0.066326530612244 9,0.9571796328821847,null,null,1,null,0.19314868804664723,0.7466787022320504,0.867
4539679304253,0.9854814767006583,2,-0.8652081513981438,0.6220644978877691,0.9382578202314874,0,0.021137026239067054,-
0.1192914188861707,2.286450499115592 3,0.910199647175250 6,0.0612244897959183 66,0.9986876640419947,0.9571796328821847,3.780713139220149 6,null,0.11443148688046
648,0.876312558490954 9,0.001457725947521865 9,55.53935860058308 5,null,0.030142115780319223,0.769236372985944 5,0.9911281257467461,0.997049284688110 2,762,-
0.751581381432471 7,0.915681186953353 4,0.940097241943117 7,0.8344961490469428,null,-
1.191656521137026 2,null,0.0743440233236151 6,0.19314868804664723,1.270475915636602 3,null,-
0.2717455877061852 3,0.851503807268189 6,0.6220644978877691,1.922353120991254 5
```

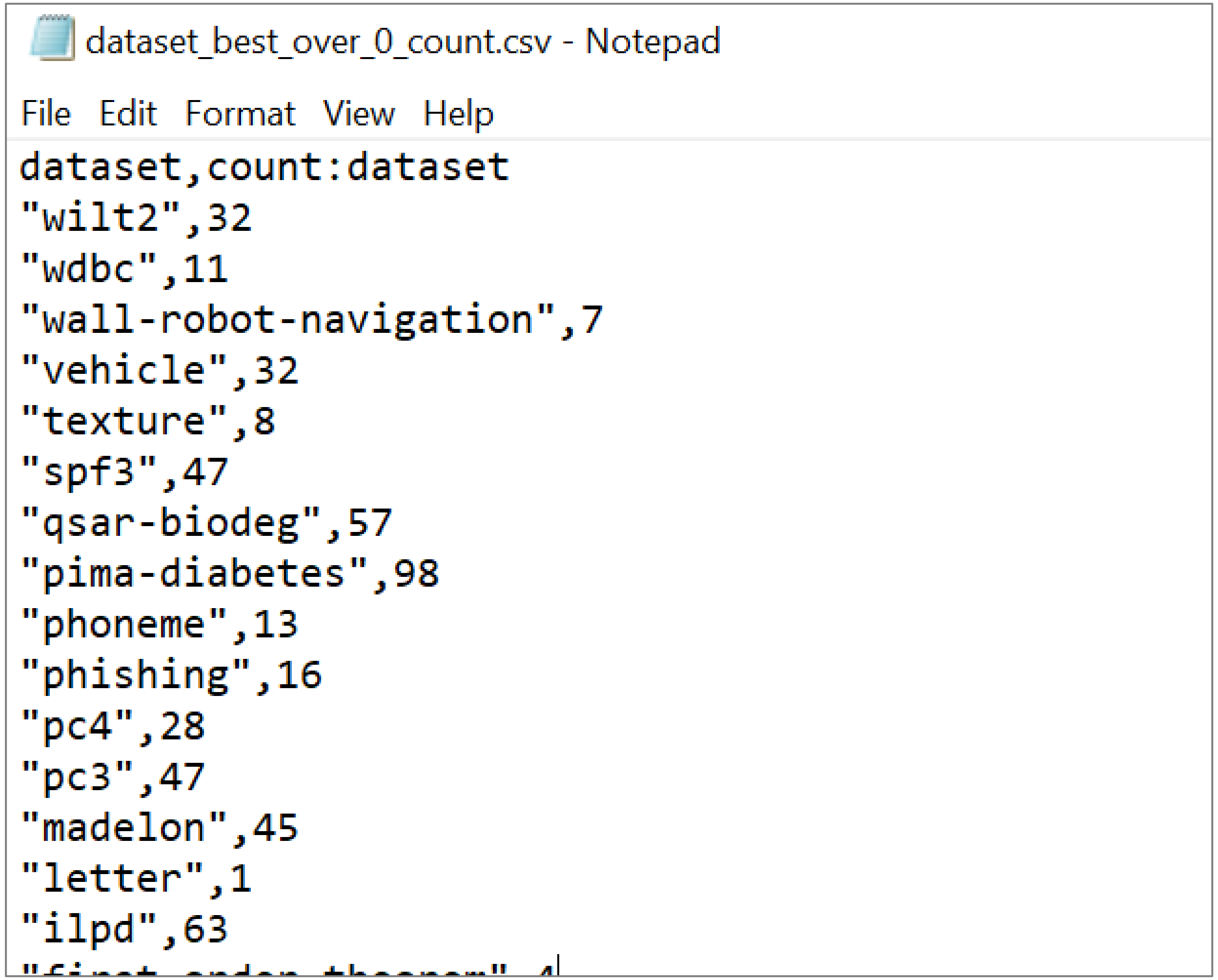
Example of data



```
"dataset","clf","clf_family","clf_params","od_name","od_params","removed","accuracy","od
_time","clf_time","total_time","gain_clf","gain_clfBest","gain_random"
"JM1","IBk","lazy","[]","CODB","{jar_path: data/java/WEKA-
CODB.jar}",0.5,0.76463,318.14848,6.88443,325.03291,0.00193,-0.05255,0.00165
"JM1","IBk","lazy","[]","CODB","{jar_path: data/java/WEKA-
CODB.jar}",1,0.76601,318.14848,6.76814,324.91662,0.00331,-0.05117,0.00257
"JM1","IBk","lazy","[]","CODB","{jar_path: data/java/WEKA-
CODB.jar}",2,0.77042,318.14848,6.76459,324.91307,0.00772,-0.04676,0.0079
```

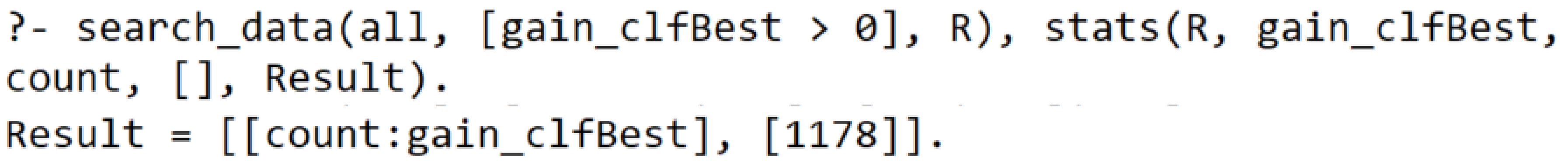


```
?- search_data(all, [gain_clfBest > 0], R),
|    stats(R, dataset, count, [dataset], Result),
|     make_csv(Result, "c:/users/katka/desktop/lab/prolog/
10jul/grafy/dataset_best_over_0_count.csv").
```

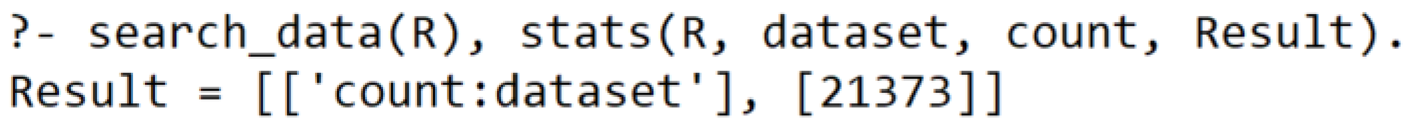


dataset_best_over_0_count.csv - Notepad
File Edit Format View Help

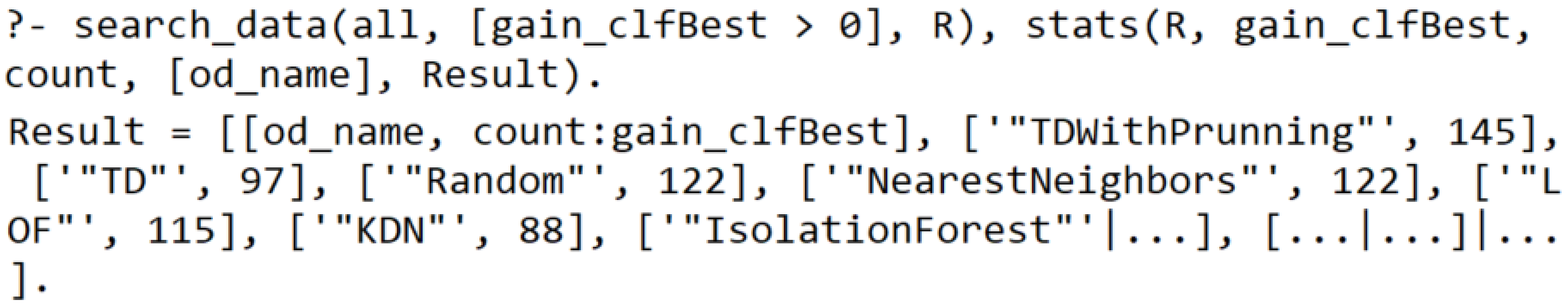
```
dataset,count:dataset
"wilt2",32
"wdbc",11
"wall-robot-navigation",7
"vehicle",32
"texture",8
"spf3",47
"qsar-biodeg",57
"pima-diabetes",98
"phoneme",13
"phishing",16
"pc4",28
"pc3",47
"madelon",45
"letter",1
"ilpd",63
"first-order-theorem",4
```



```
?- search_data(all, [gain_clfBest > 0], R), stats(R, gain_clfBest,
count, [], Result).
Result = [[count:gain_clfBest], [1178]].
```



```
?- search_data(R), stats(R, dataset, count, Result).
Result = [['count:dataset'], [21373]]
```



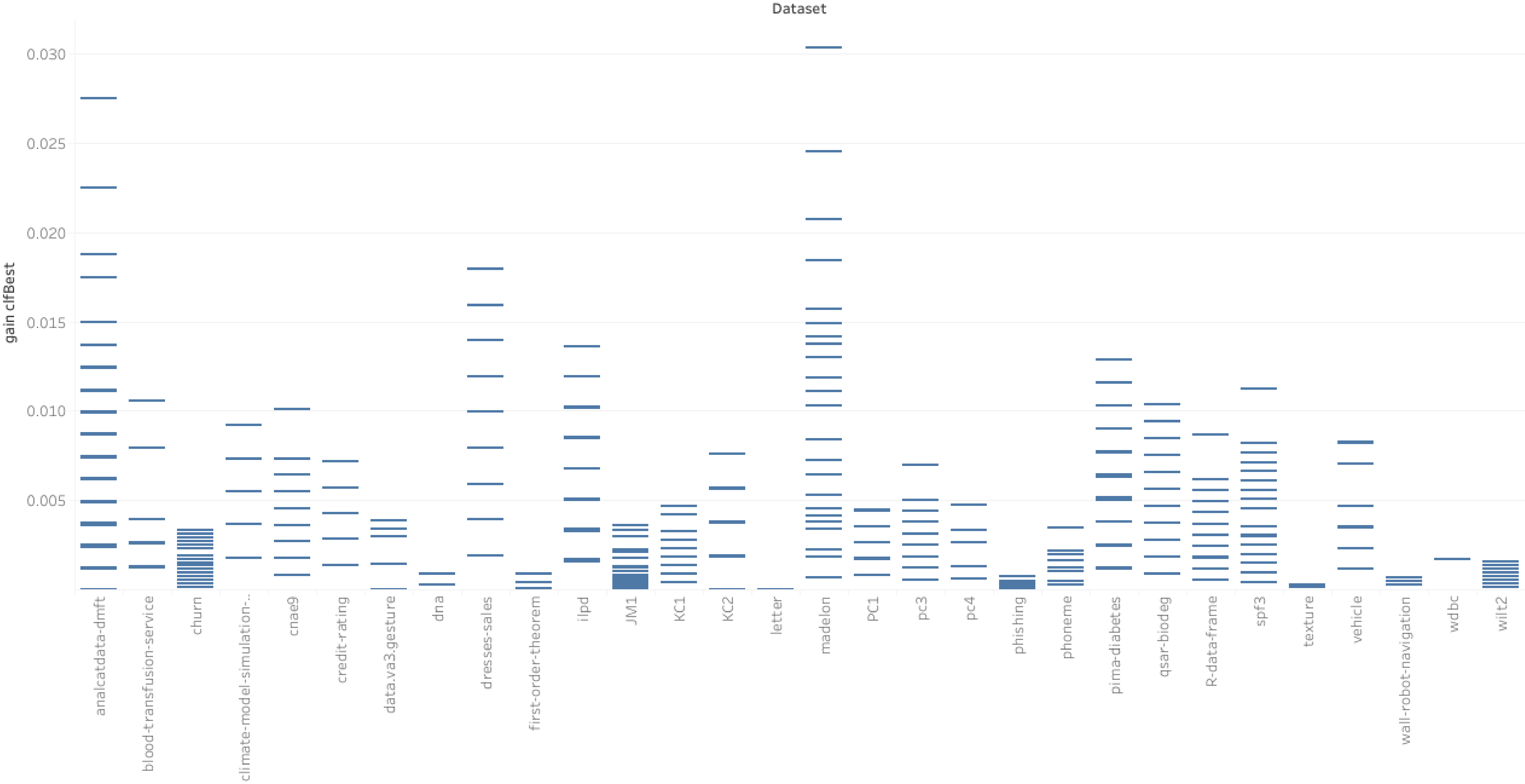
```
?- search_data(all, [gain_clfBest > 0], R), stats(R, gain_clfBest,
count, [od_name], Result).
Result = [[od_name, count:gain_clfBest], ['"TDWithPrunning"', 145],
 ['"TD"', 97], ['"Random"', 122], ['"NearestNeighbors"', 122], ['"L
OF"', 115], ['"KDN"', 88], ['"IsolationForest"'|...], [...|...]|...
].
```

```
?- search_data(all, [gain_clfBest > 0], R1), stats(R1, gain_clfBest
, count, [od_name], R2), condi(R2, [_, BestCount], BestCount > 100,
 Result).
Result = [[od_name, 'count:gain_clfBest'], ['"TDWithPrunning"', 145
], ['"Random"', 122], ['"NearestNeighbors"', 122], ['"LOF"', 115],
['"DS"', 102], ['"ClassLikelihood"', 108]].
```

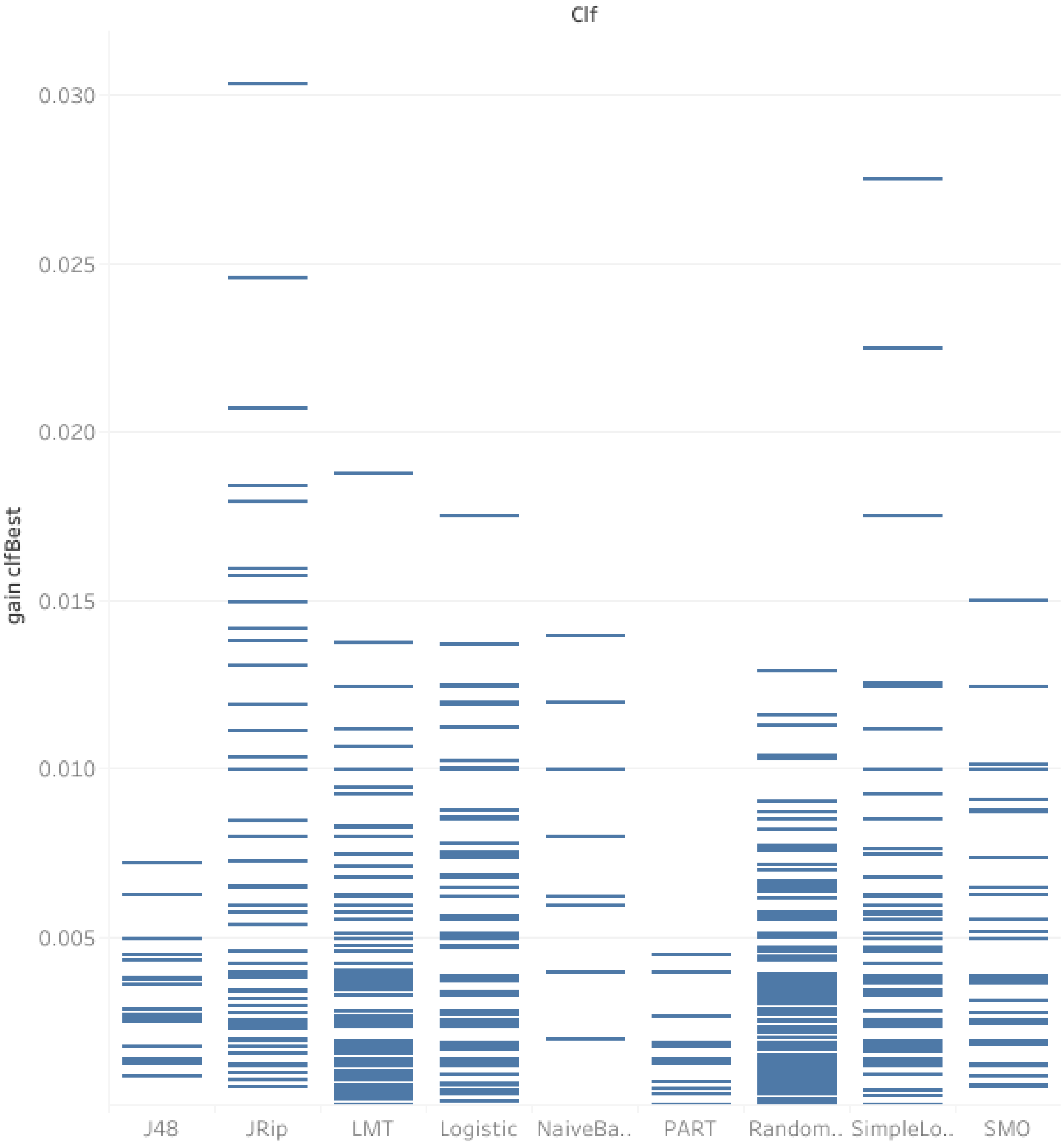
Graphs

Positive gain_clfBest

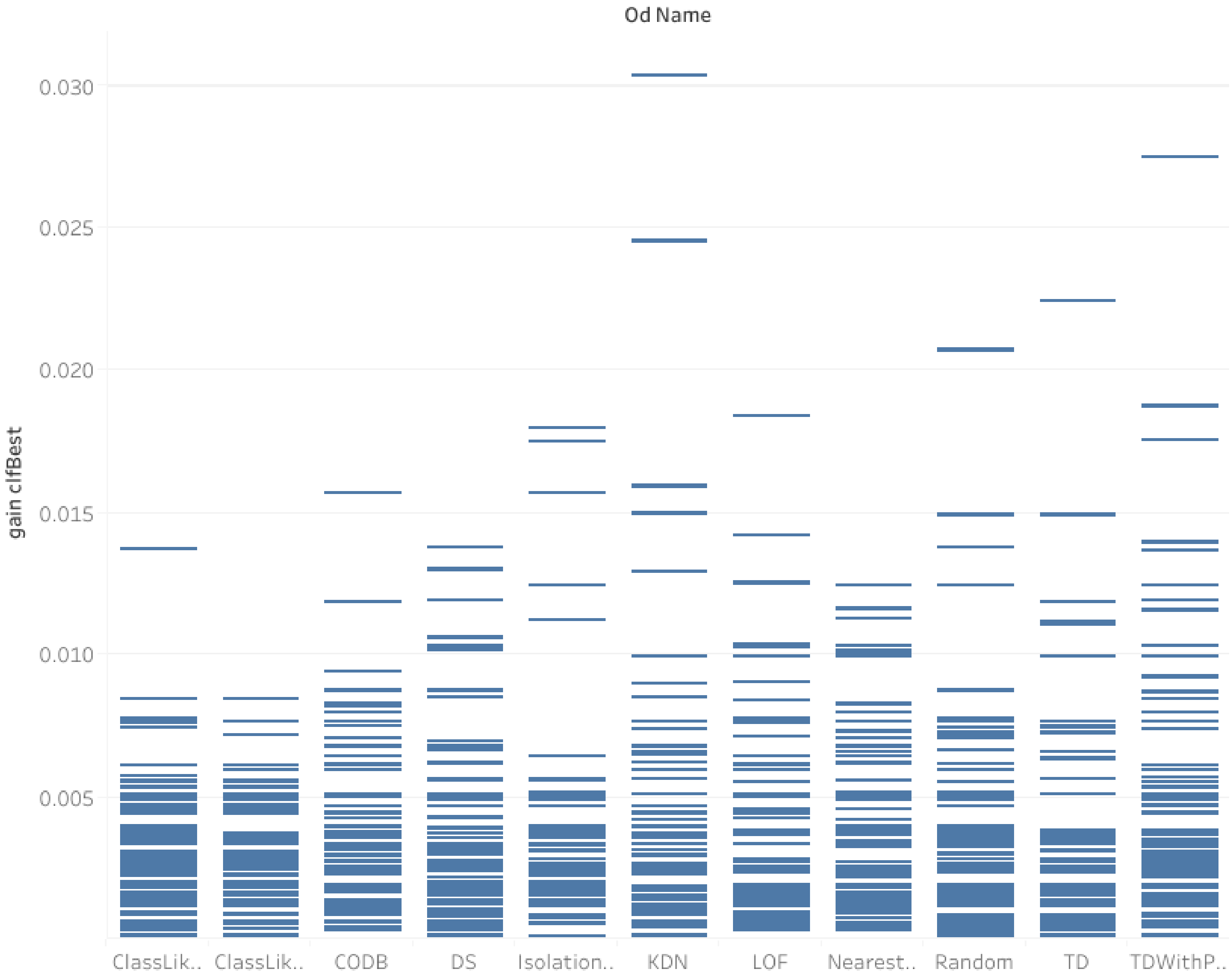
Sheet 3


Dataset
gain clfBest
0.030
0.025
0.020
0.015
0.010
0.005
analcatdata-dmft
blood-transfusion-service
churn
climate-model-simulation-..
cnae9
credit-rating
data.va3.gesture
dna
dresses-sales
first-order-theorem
ilpd
JM1
KC1
KC2
letter
madelon
PC1
pc3
pc4
phishing
phoneme
pima-diabetes
qsar-biodeg
R-data-frame
spf3
texture
vehicle
wall-robot-navigation
wdbc
wilt2

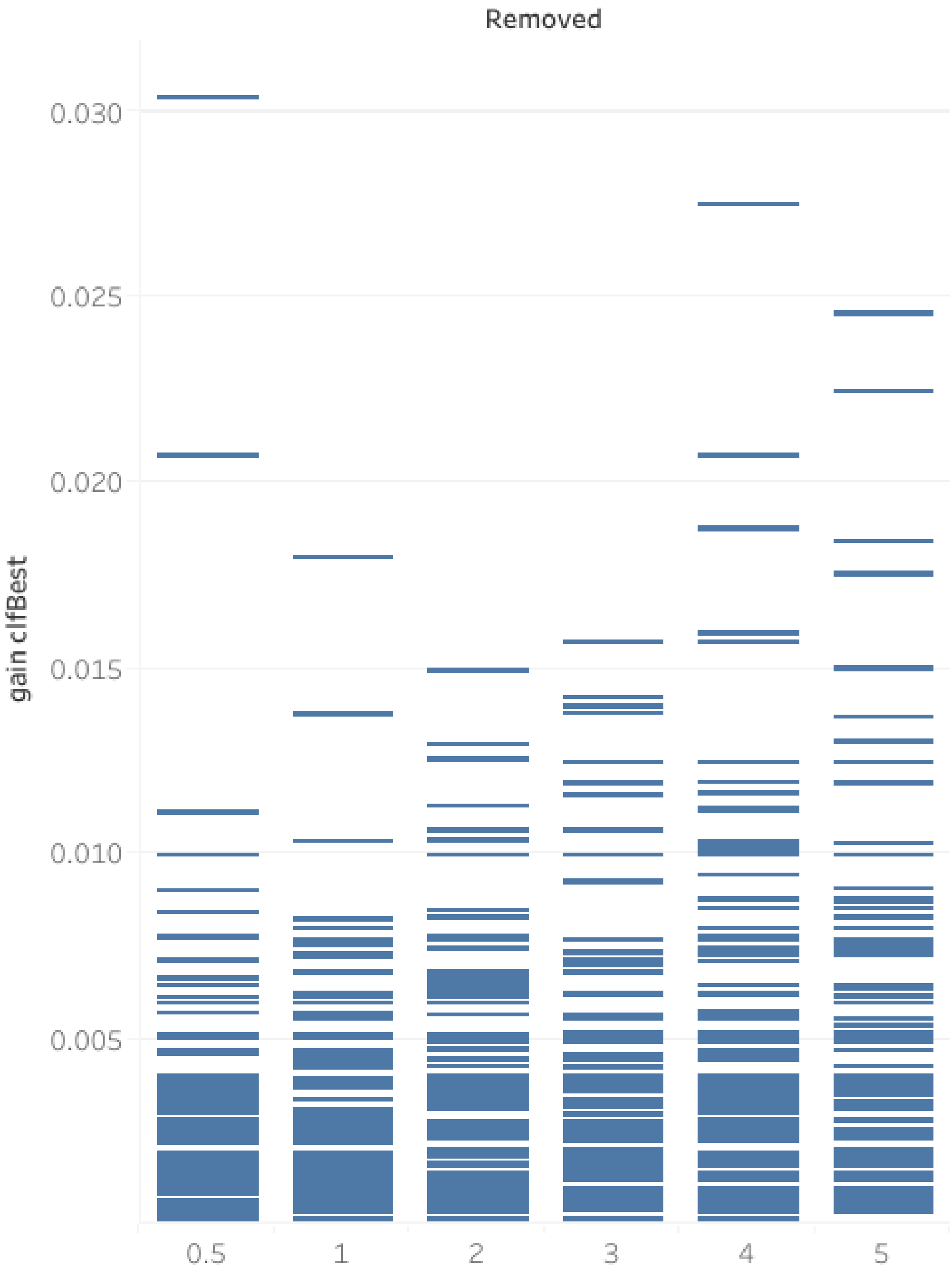
Sheet 3 (12)


Clf
gain clfBest
0.030
0.025
0.020
0.015
0.010
0.005
J48
JRip
LMT
Logistic
NaiveBa..
PART
Random..
SimpleLo..
SMO

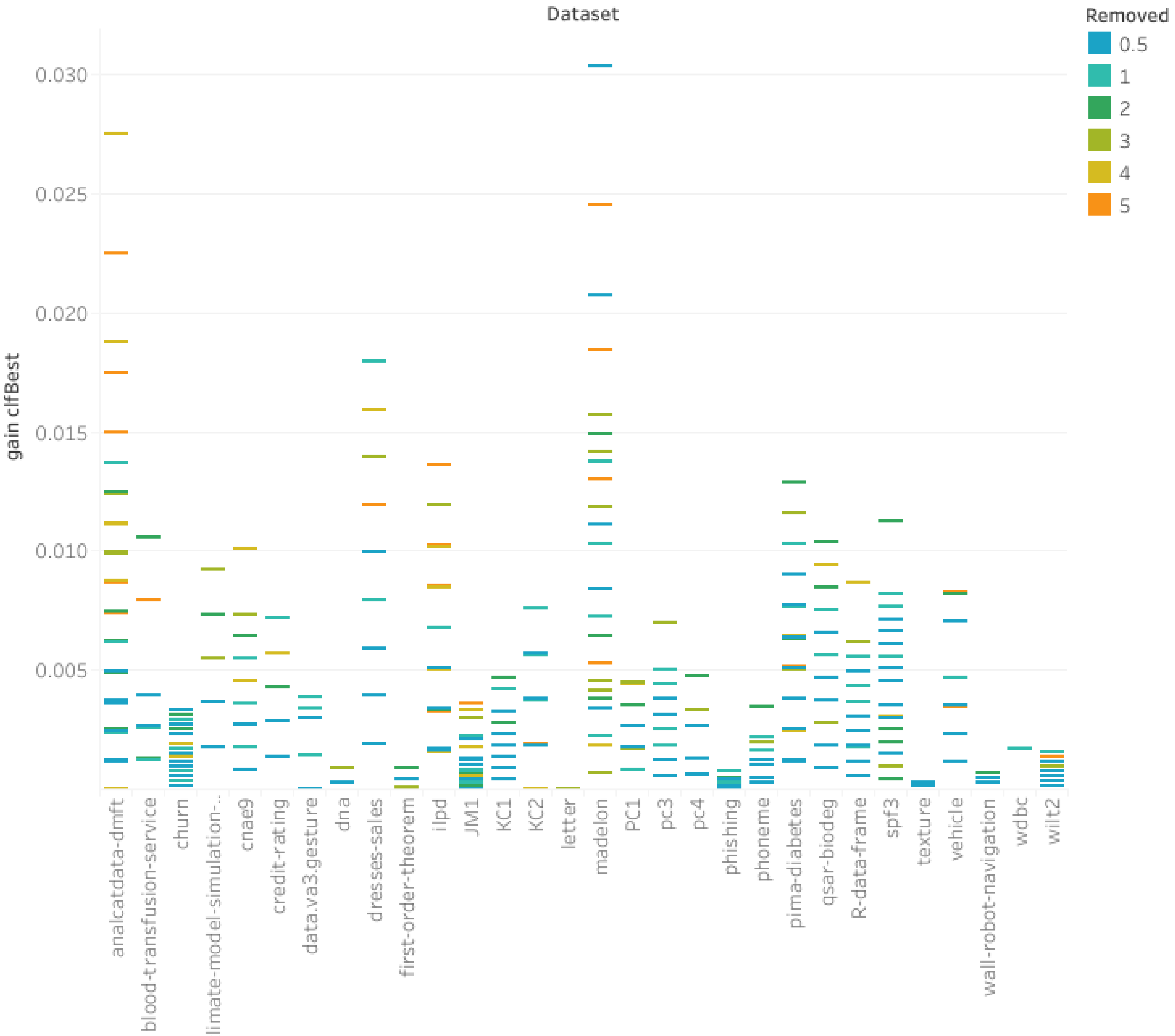
Sheet 3 (13)


Od Name
gain clfBest
0.030
0.025
0.020
0.015
0.010
0.005
ClassLik..
ClassLik..
CODB
DS
Isolation..
KDN
LOF
Nearest..
Random
TD
TDWithP..

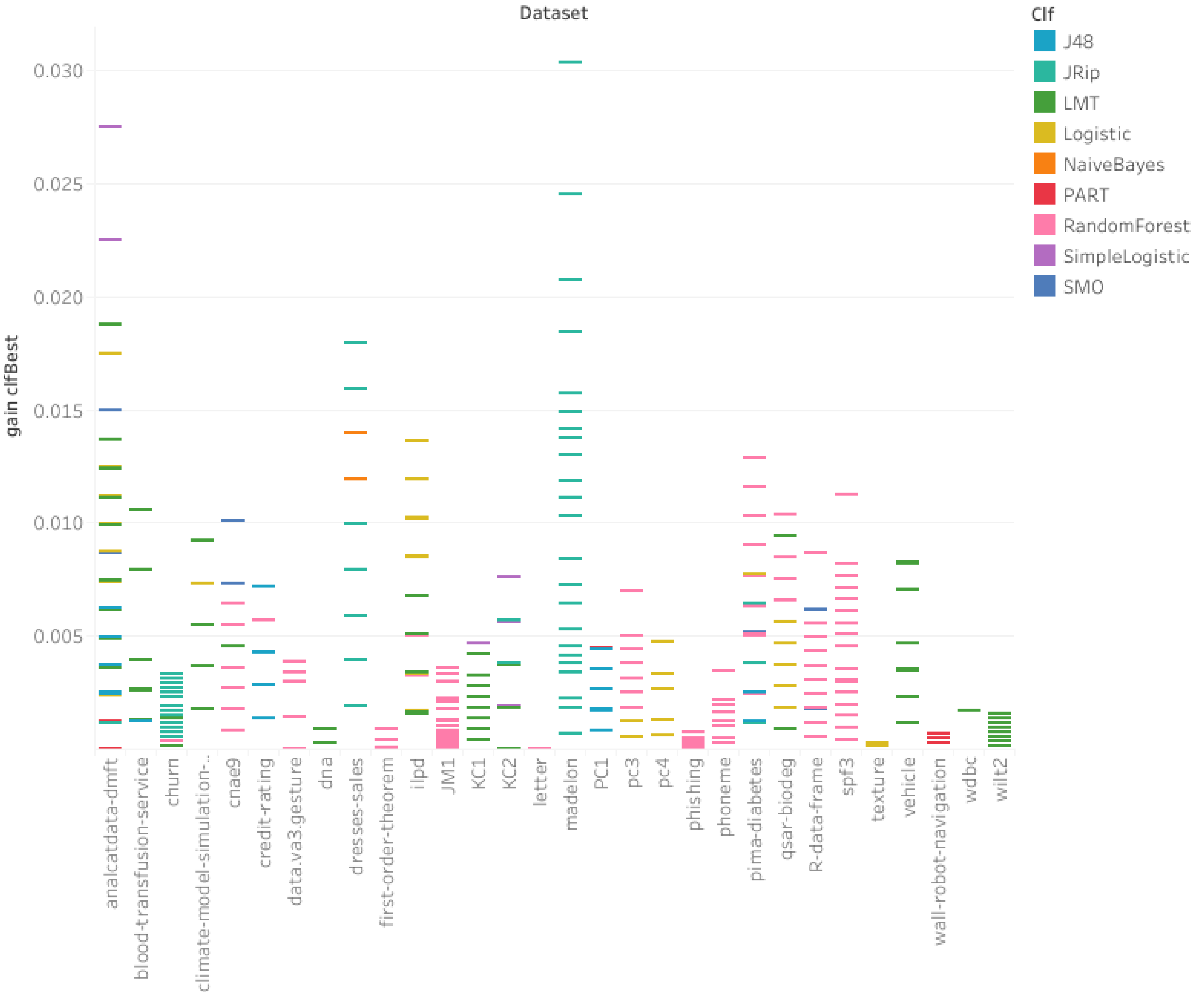
Sheet 3 (6)



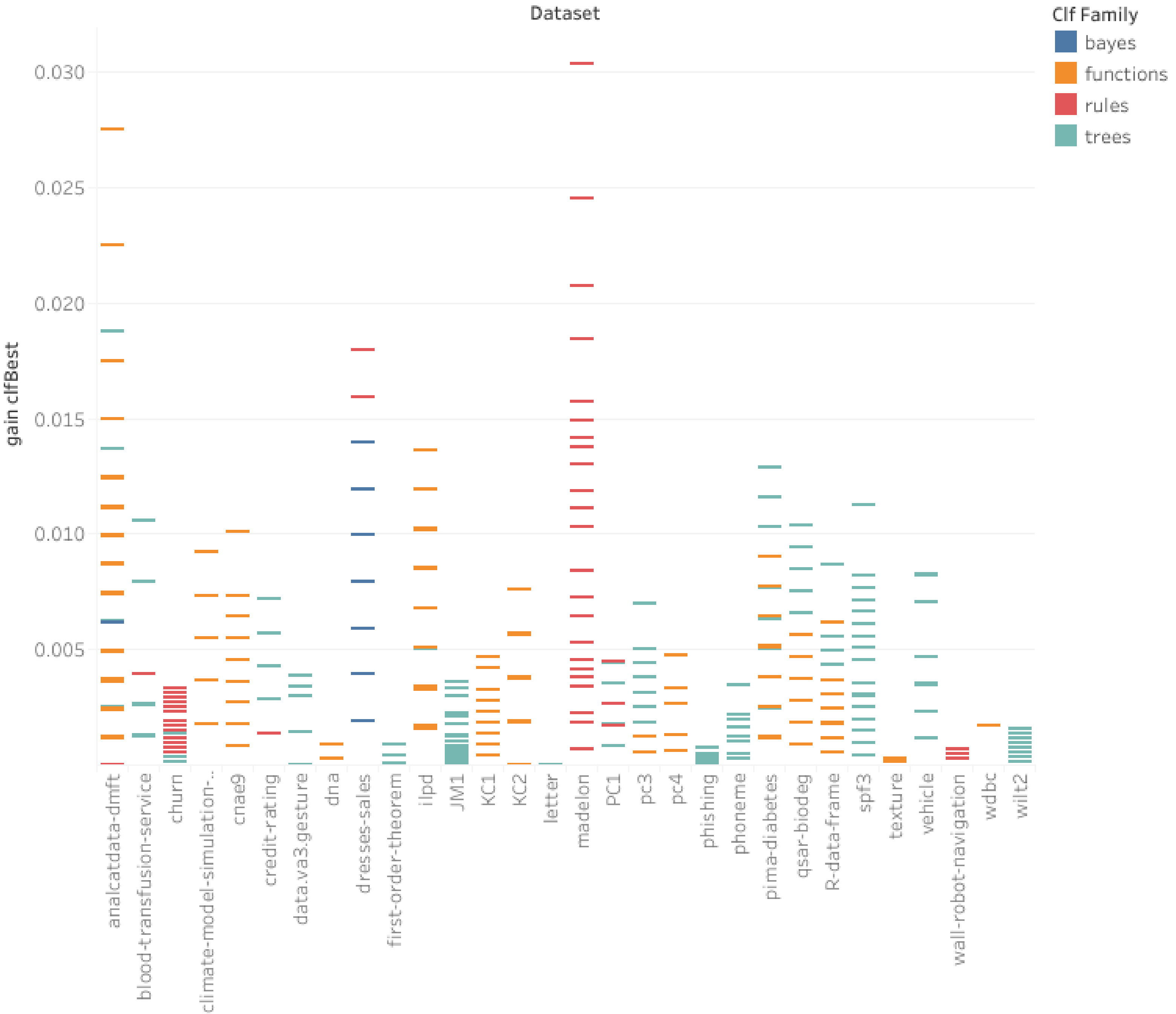
Sheet 3 (14)


Dataset
Removed
0.5
1
2
3
4
5
gain clfBest
0.030
0.025
0.020
0.015
0.010
0.005
analcatdata-dmft
blood-transfusion-service
churn
limate-model-simulation-..
cnae9
credit-rating
data.va3.gesture
dna
dresses-sales
first-order-theorem
ilpd
JM1
KC1
KC2
letter
madelon
PC1
pc3
pc4
phishing
phoneme
pima-diabetes
qsar-biodeg
R-data-frame
spf3
texture
vehicle
wall-robot-navigation
wdbc
wilt2

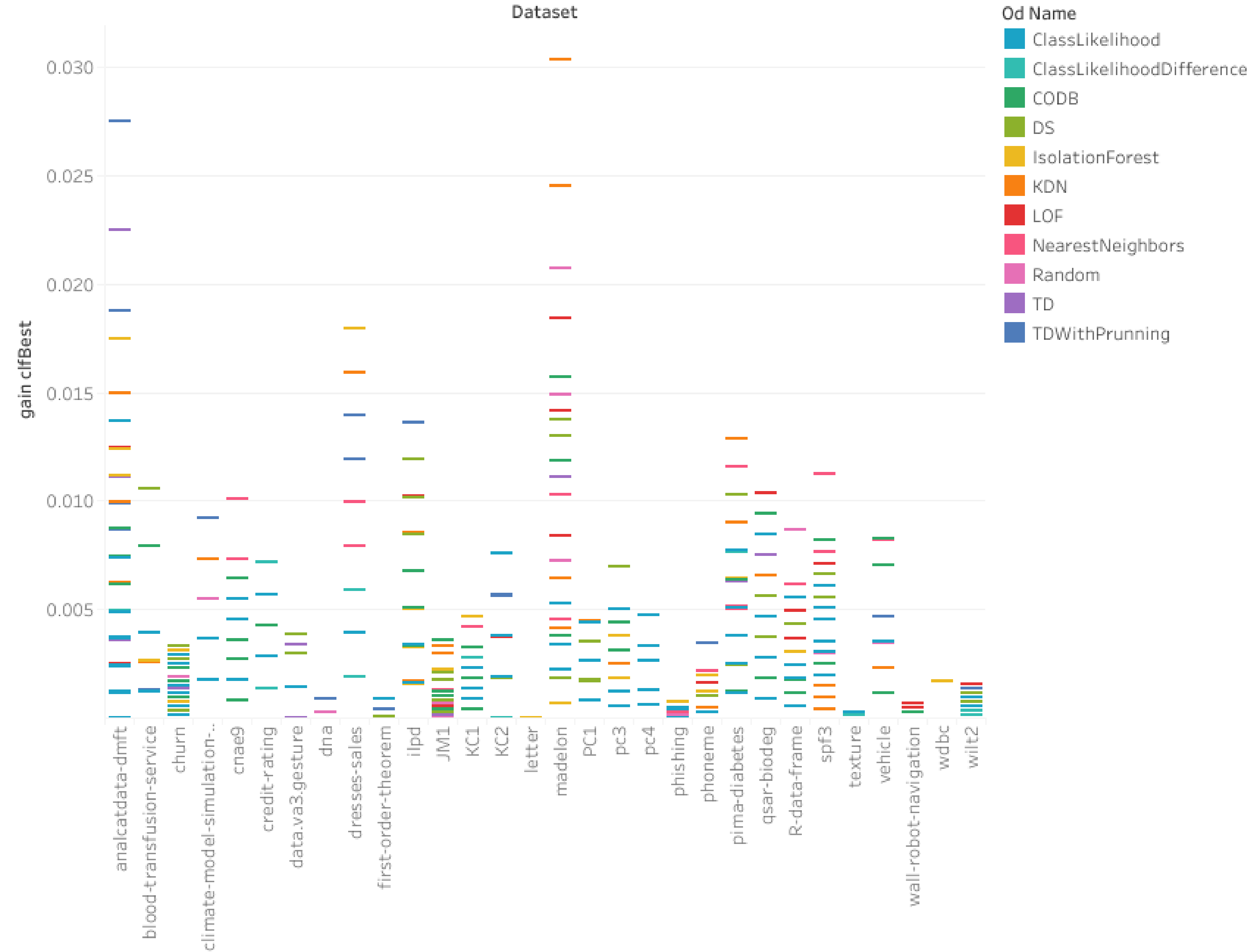
Sheet 3 (4)


Dataset
gain clfBest
0.030
0.025
0.020
0.015
0.010
0.005
analcatdata-dmft
blood-transfusion-service
churn
climate-model-simulation-..
cnae9
credit-rating
data.va3.gesture
dna
dresses-sales
first-order-theorem
ilpd
JM1
KC1
KC2
letter
madelon
PC1
pc3
pc4
phishing
phoneme
pima-diabetes
qsar-biodeg
R-data-frame
spf3
texture
vehicle
wall-robot-navigation
wdbc
wilt2
Clf
J48
JRip
LMT
Logistic
NaiveBayes
PART
RandomForest
SimpleLogistic
SMO

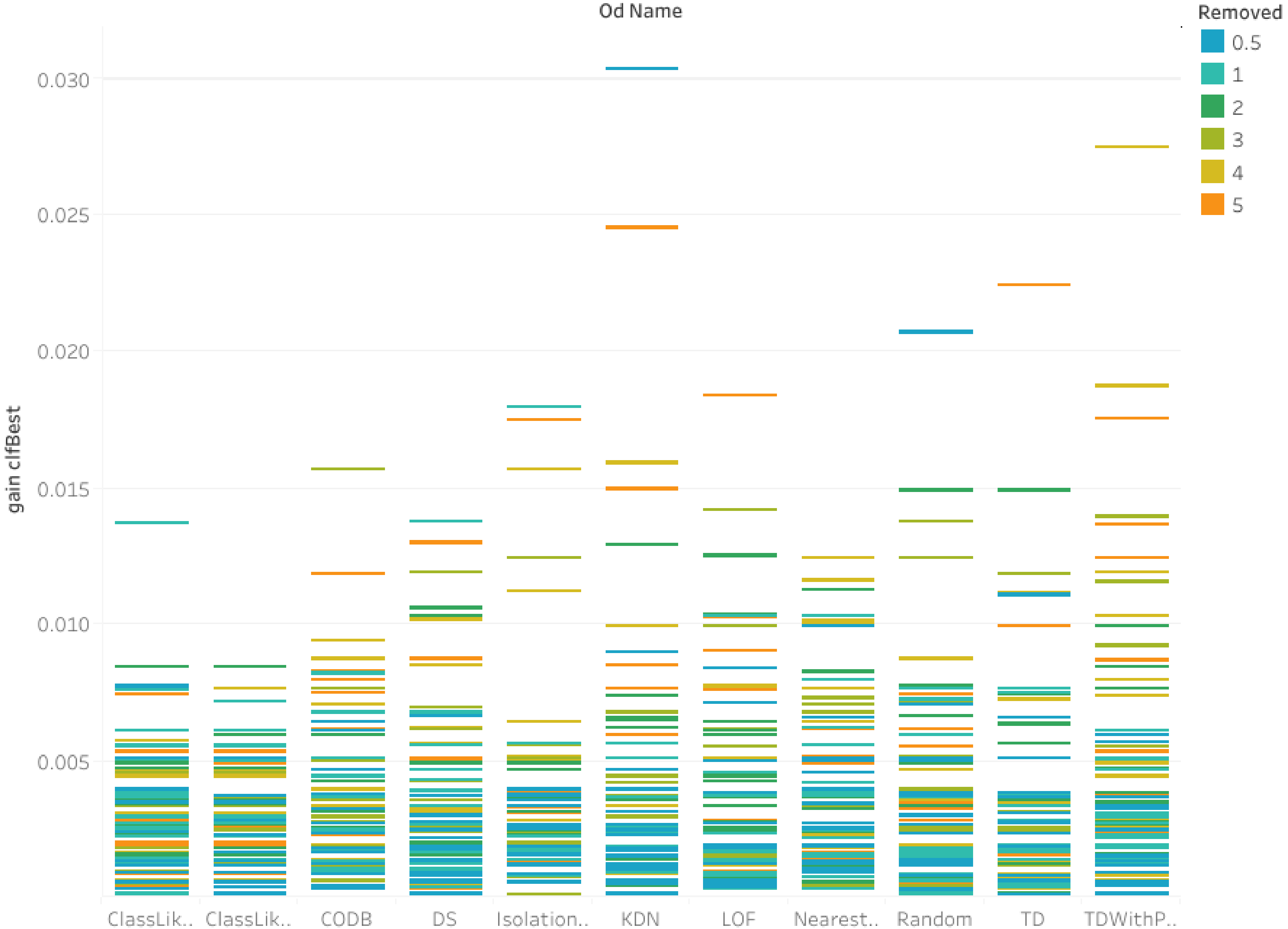
Sheet 3 (3)


Dataset
Clf Family
bayes
functions
rules
trees
gain clfBest
0.030
0.025
0.020
0.015
0.010
0.005
analcatdata-dmft
blood-transfusion-service
churn
climate-model-simulation-..
cnae9
credit-rating
data.va3.gesture
dna
dresses-sales
first-order-theorem
ilpd
JM1
KC1
KC2
letter
madelon
PC1
pc3
pc4
phishing
phoneme
pima-diabetes
qsar-biodeg
R-data-frame
spf3
texture
vehicle
wall-robot-navigation
wdbc
wilt2

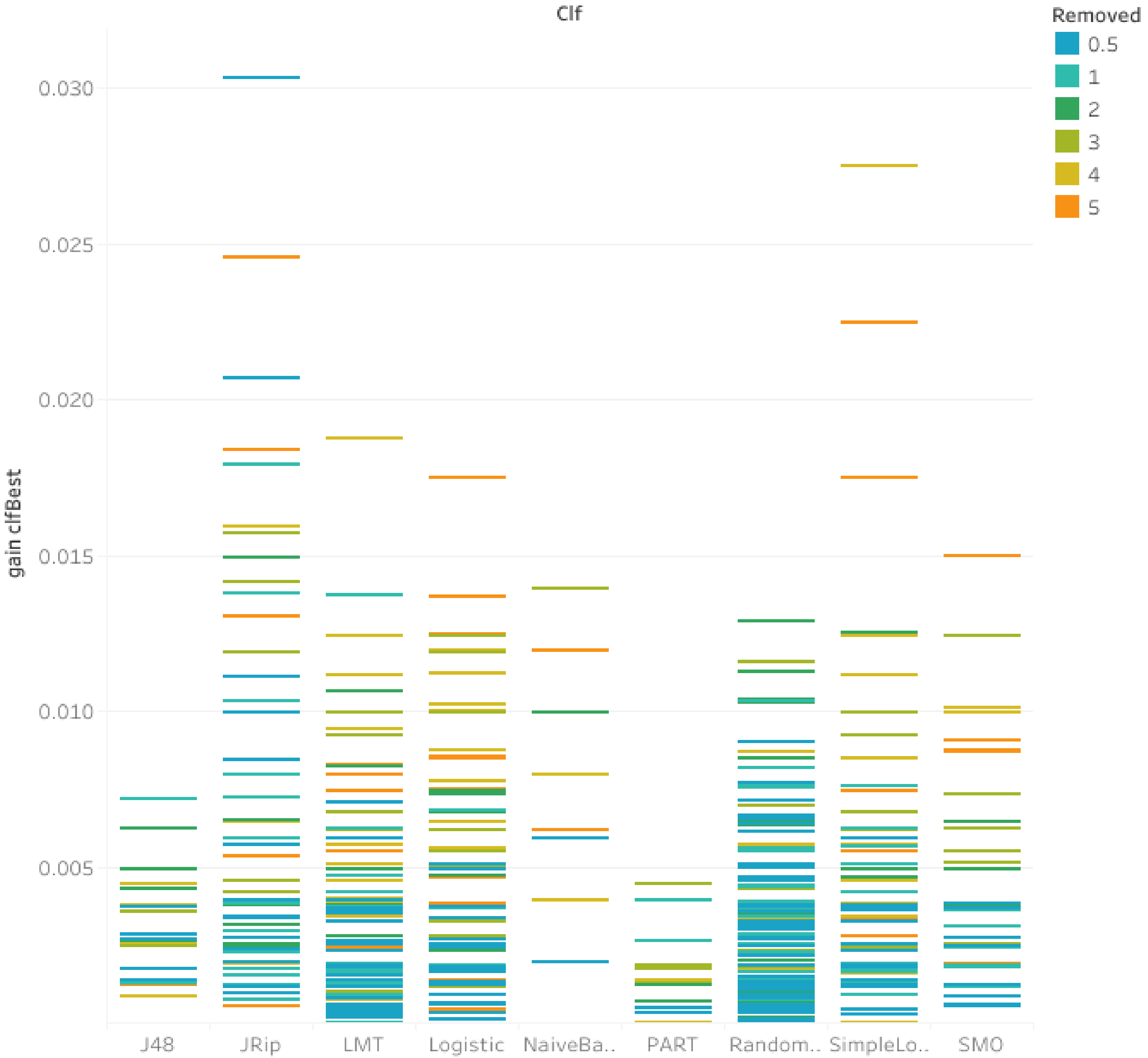
Sheet 3 (2)


Dataset
gain clfBest
0.030
0.025
0.020
0.015
0.010
0.005
analcatdata-dmft
blood-transfusion-service
churn
climate-model-simulation-..
cnae9
credit-rating
data.va3.gesture
dna
dresses-sales
first-order-theorem
ilpd
JM1
KC1
KC2
letter
madelon
PC1
pc3
pc4
phishing
phoneme
pima-diabetes
qsar-biodeg
R-data-frame
spf3
texture
vehicle
wall-robot-navigation
wdbc
wilt2
Od Name
ClassLikelihood
ClassLikelihoodDifference
CODB
DS
IsolationForest
KDN
LOF
NearestNeighbors
Random
TD
TDWithPrunning

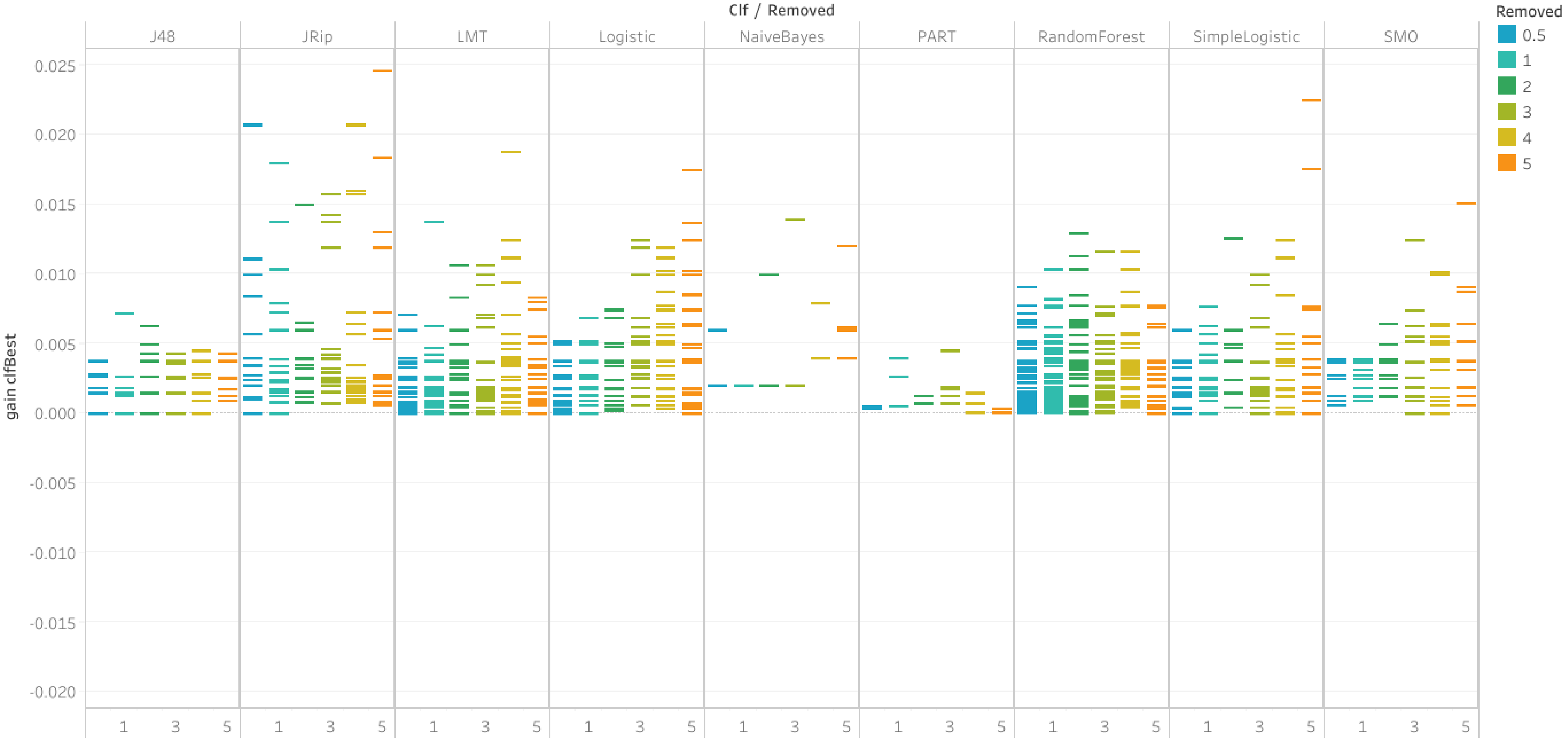
Sheet 3 (20)


Od Name
Removed
0.5
1
2
3
4
5
gain clfBest
0.030
0.025
0.020
0.015
0.010
0.005
ClassLik..
ClassLik..
CODB
DS
Isolation..
KDN
LOF
Nearest..
Random
TD
TDWithP..

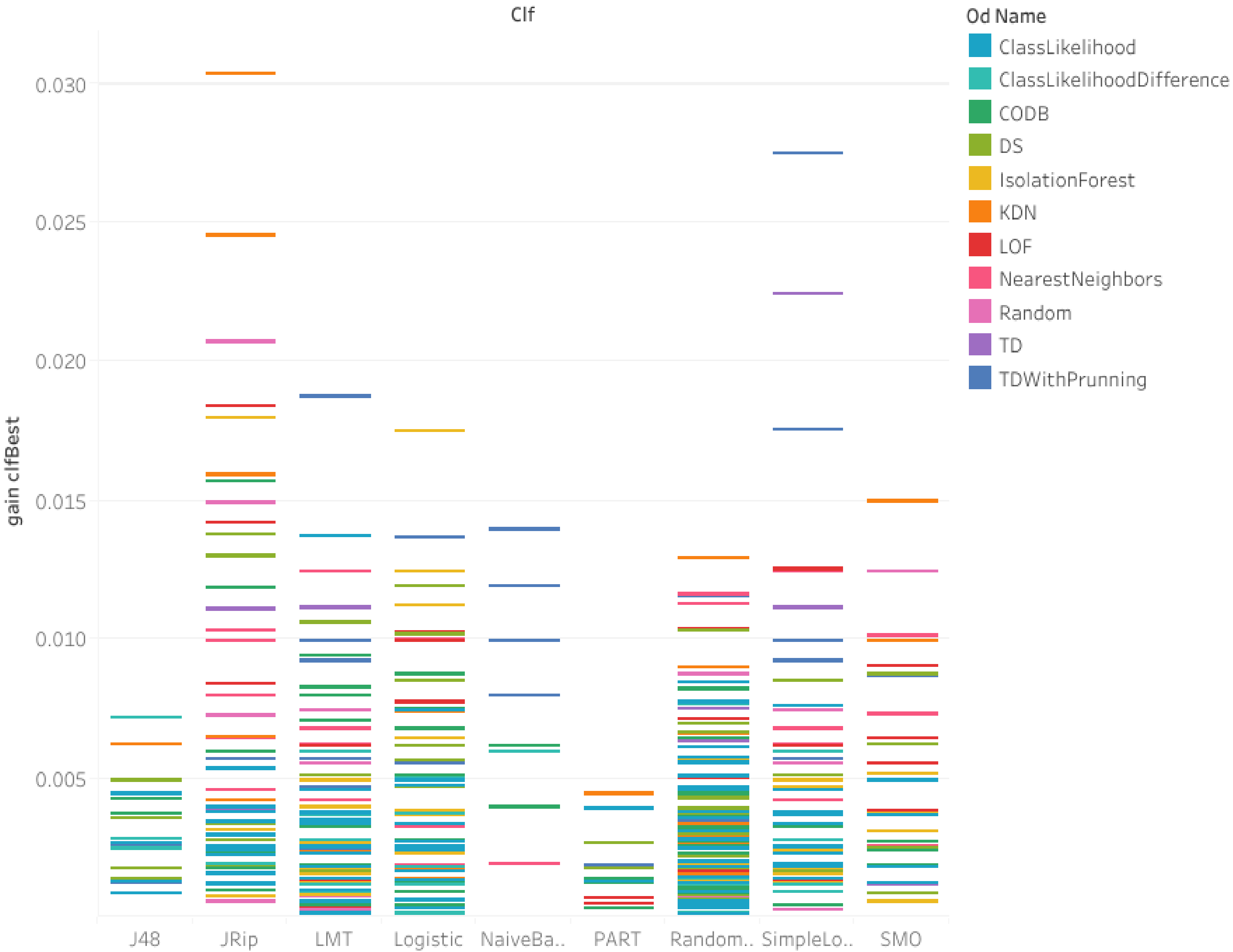
Sheet 3 (17)


Clf
gain clfBest
0.030
0.025
0.020
0.015
0.010
0.005
J48
JRip
LMT
Logistic
NaiveBa..
PART
Random..
SimpleLo..
SMO
Removed
0.5
1
2
3
4
5

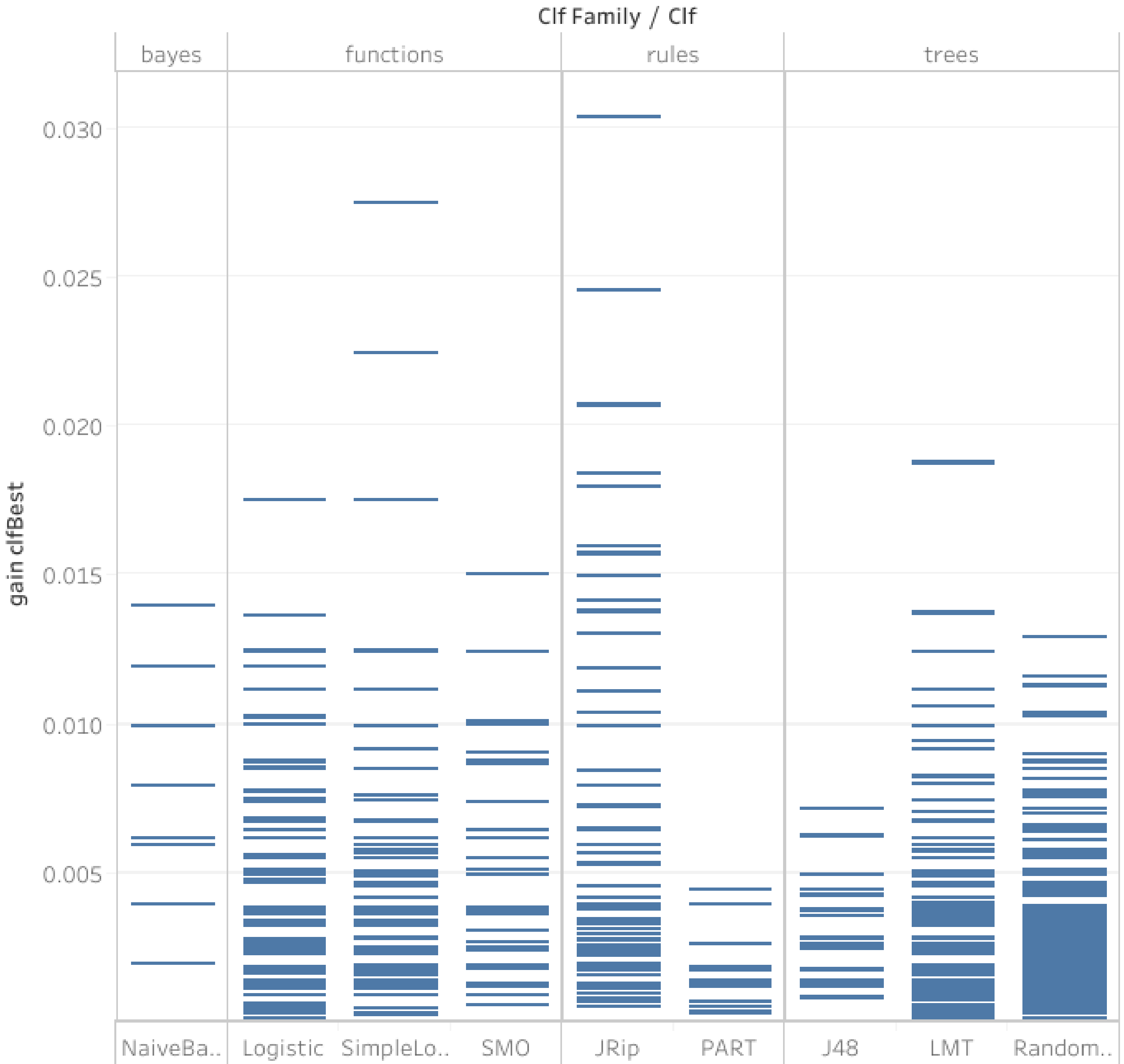
Sheet 3 (17)


Clf / Removed
J48
JRip
LMT
Logistic
NaiveBayes
PART
RandomForest
SimpleLogistic
SMO
Removed
0.5
1
2
3
4
5
gain clfBest
0.025
0.020
0.015
0.010
0.005
0.000
-0.005
-0.010
-0.015
-0.020
1
3
5

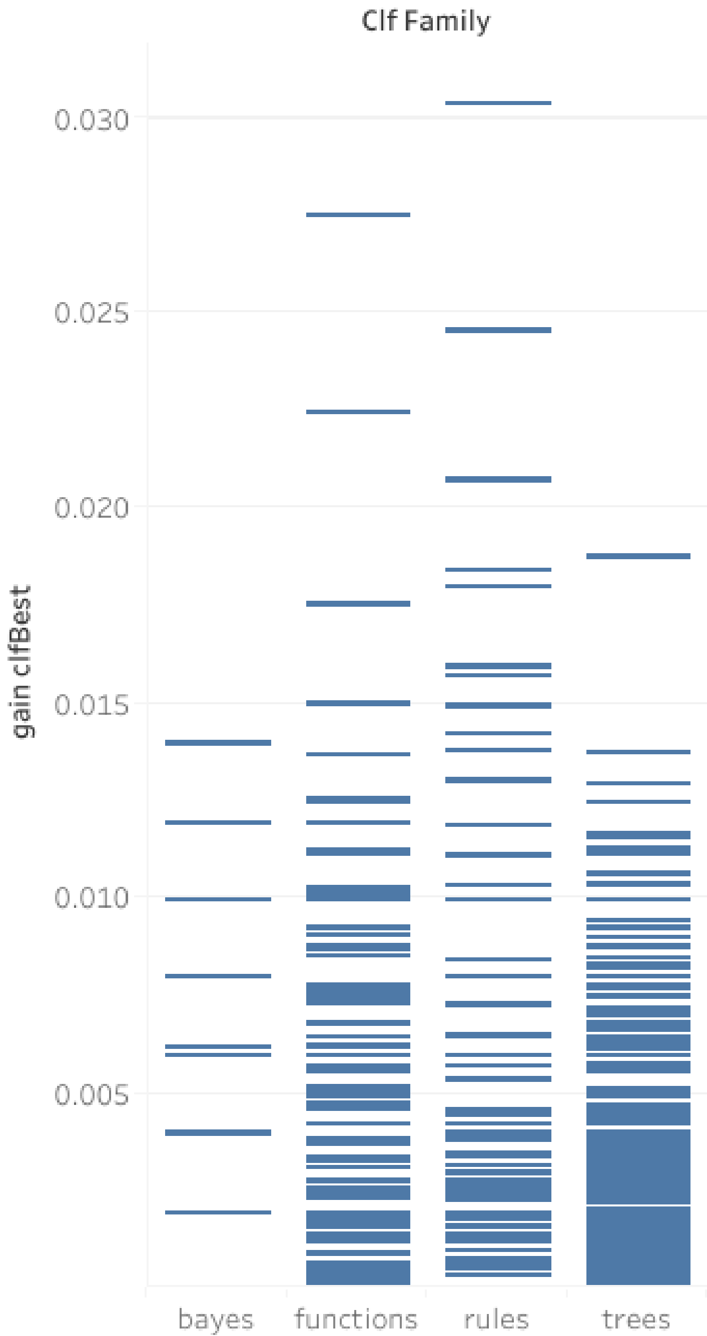
Sheet 3 (16)


Clf
Od Name
ClassLikelihood
ClassLikelihoodDifference
CODB
DS
IsolationForest
KDN
LOF
NearestNeighbors
Random
TD
TDWithPrunning
gain clfBest
0.030
0.025
0.020
0.015
0.010
0.005
J48
JRip
LMT
Logistic
NaiveBa..
PART
Random..
SimpleLo..
SMO

Sheet 3 (15)

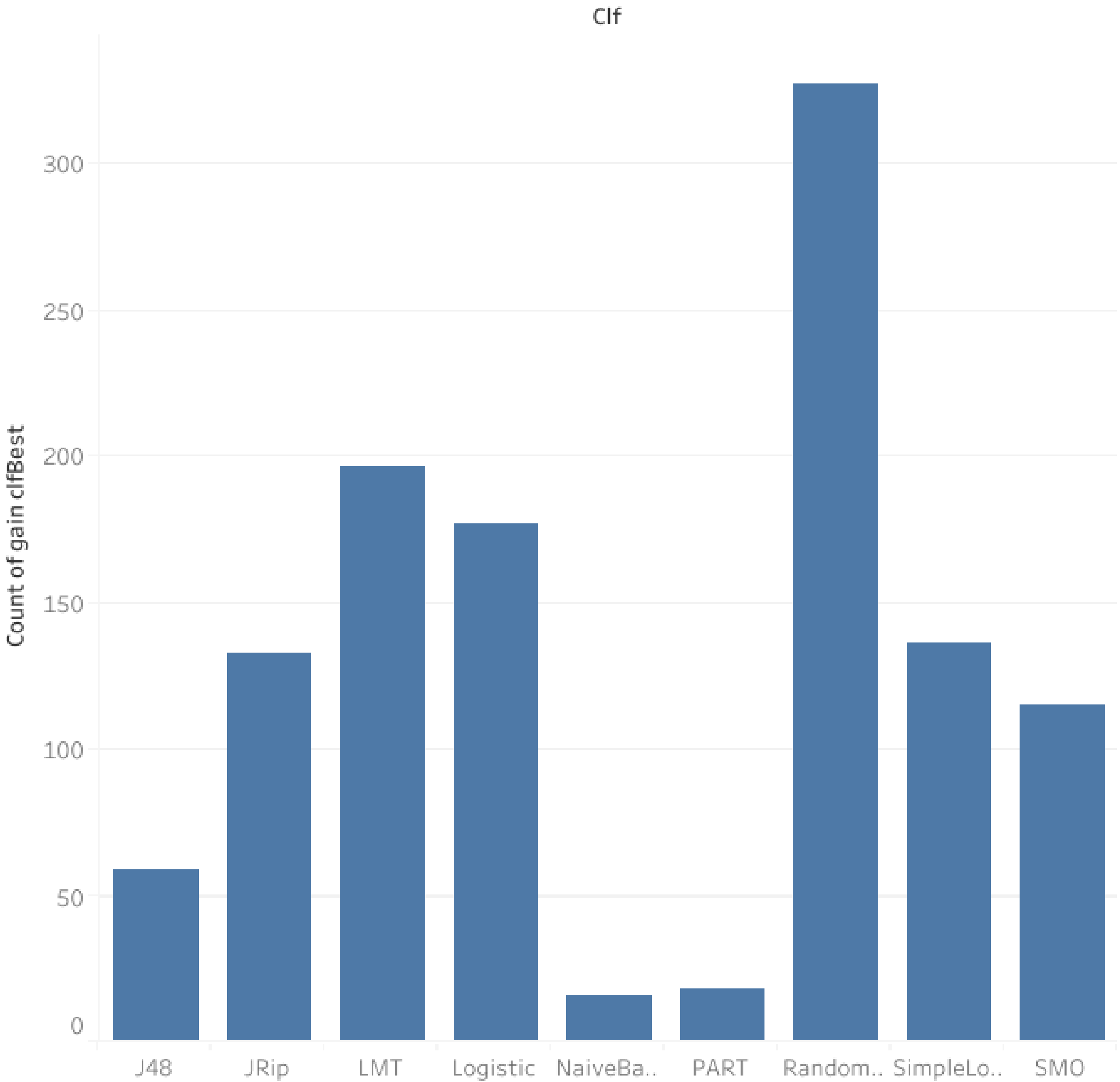


Sheet 3 (11)

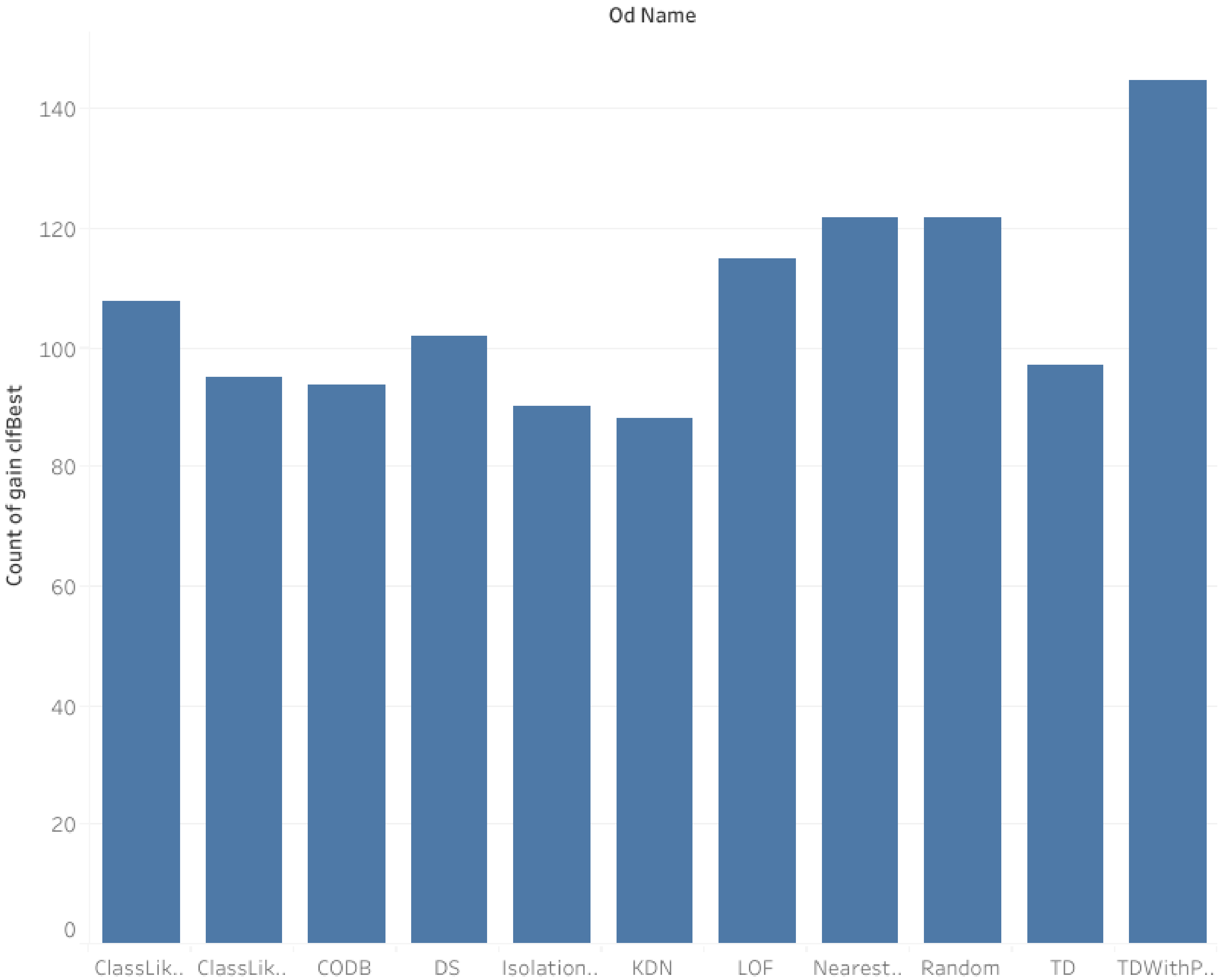


Count of gain_clfBest

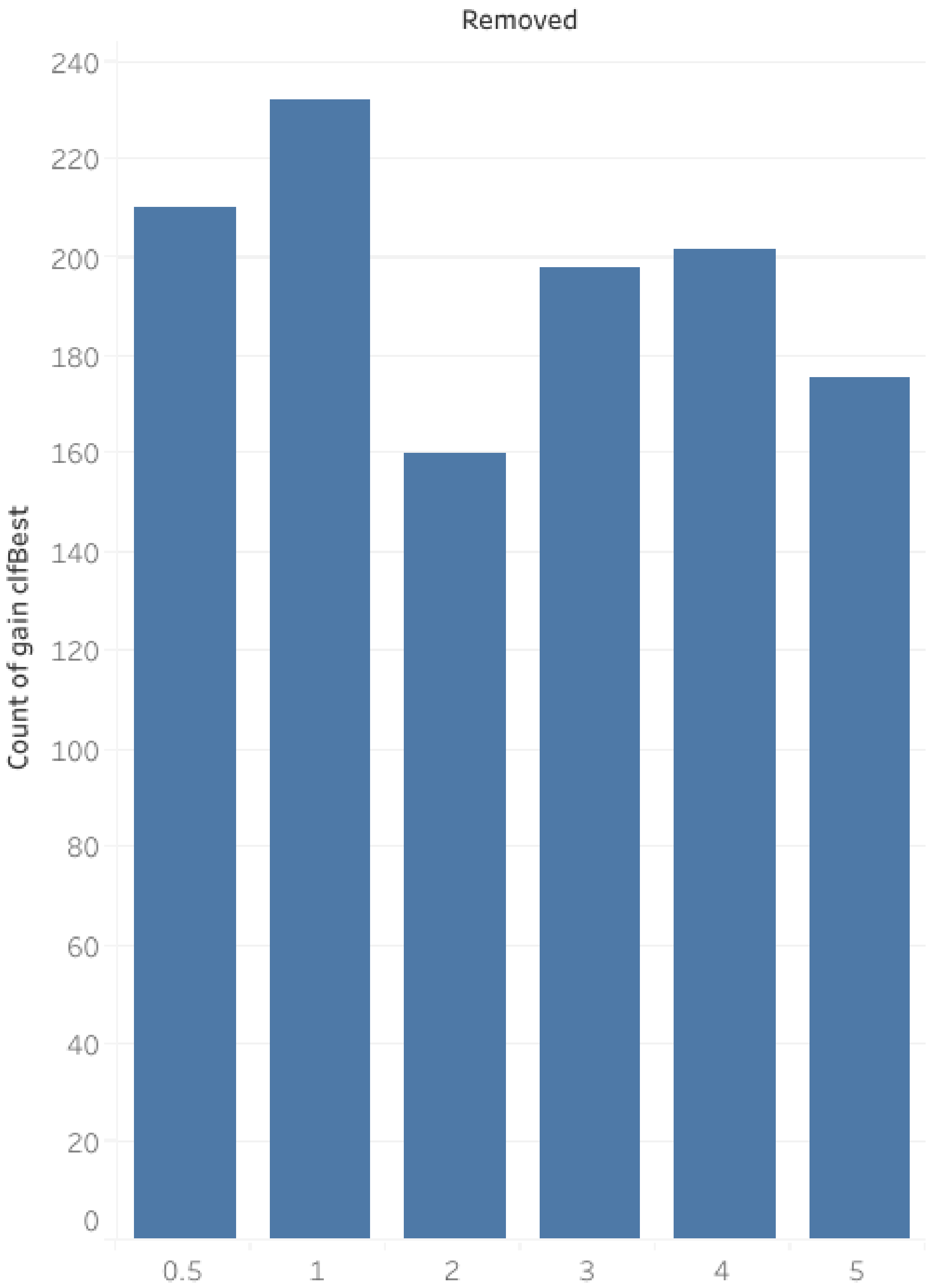
Sheet 3 (27)


Clf
Count of gain clfBest
0
50
100
150
200
250
300
J48
JRip
LMT
Logistic
NaiveBa..
PART
Random..
SimpleLo..
SMO

Sheet 3 (26)

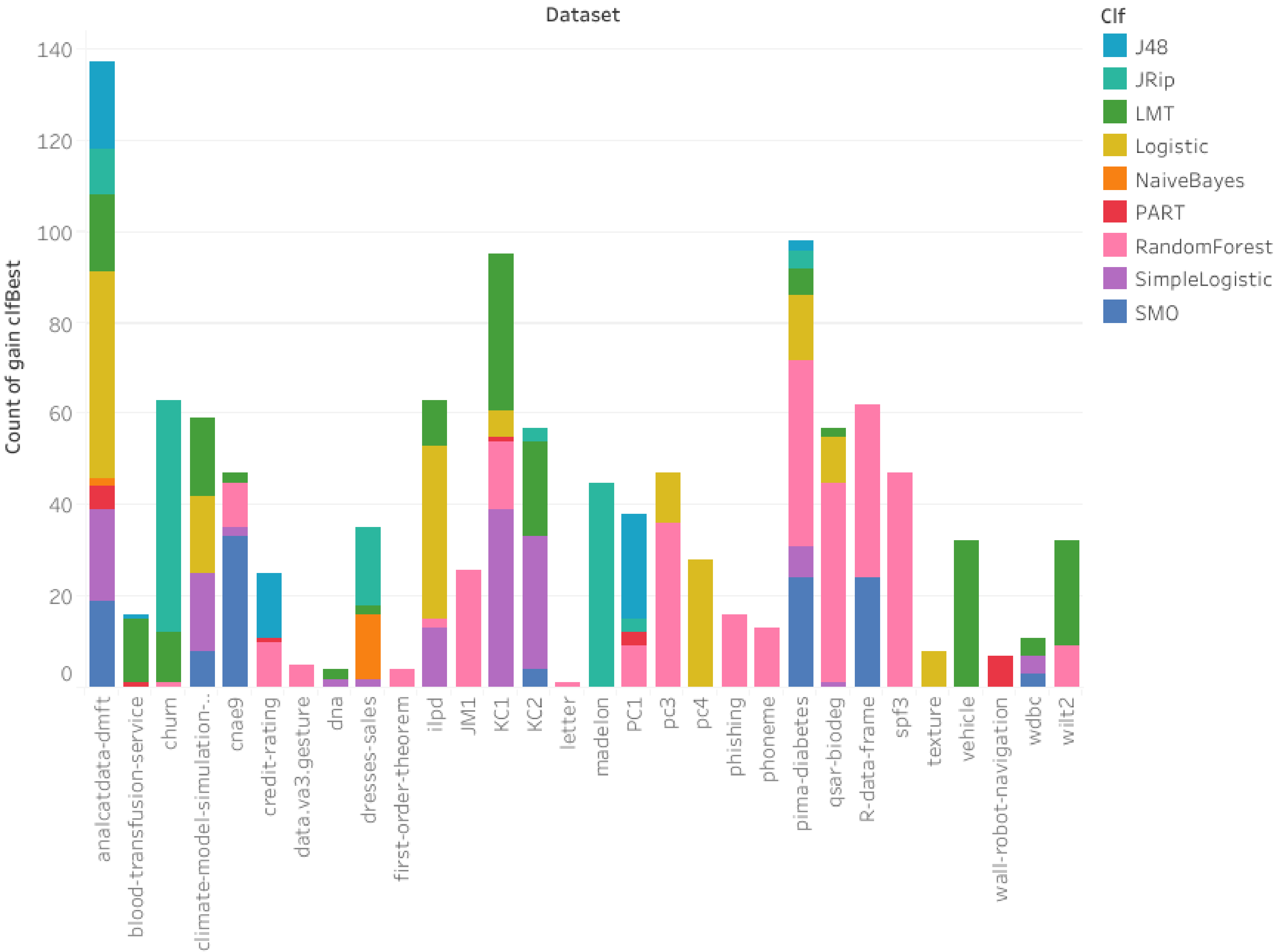

Od Name
Count of gain clfBest
0
20
40
60
80
100
120
140
ClassLik..
ClassLik..
CODB
DS
Isolation..
KDN
LOF
Nearest..
Random
TD
TDWithP..

Sheet 3 (25)

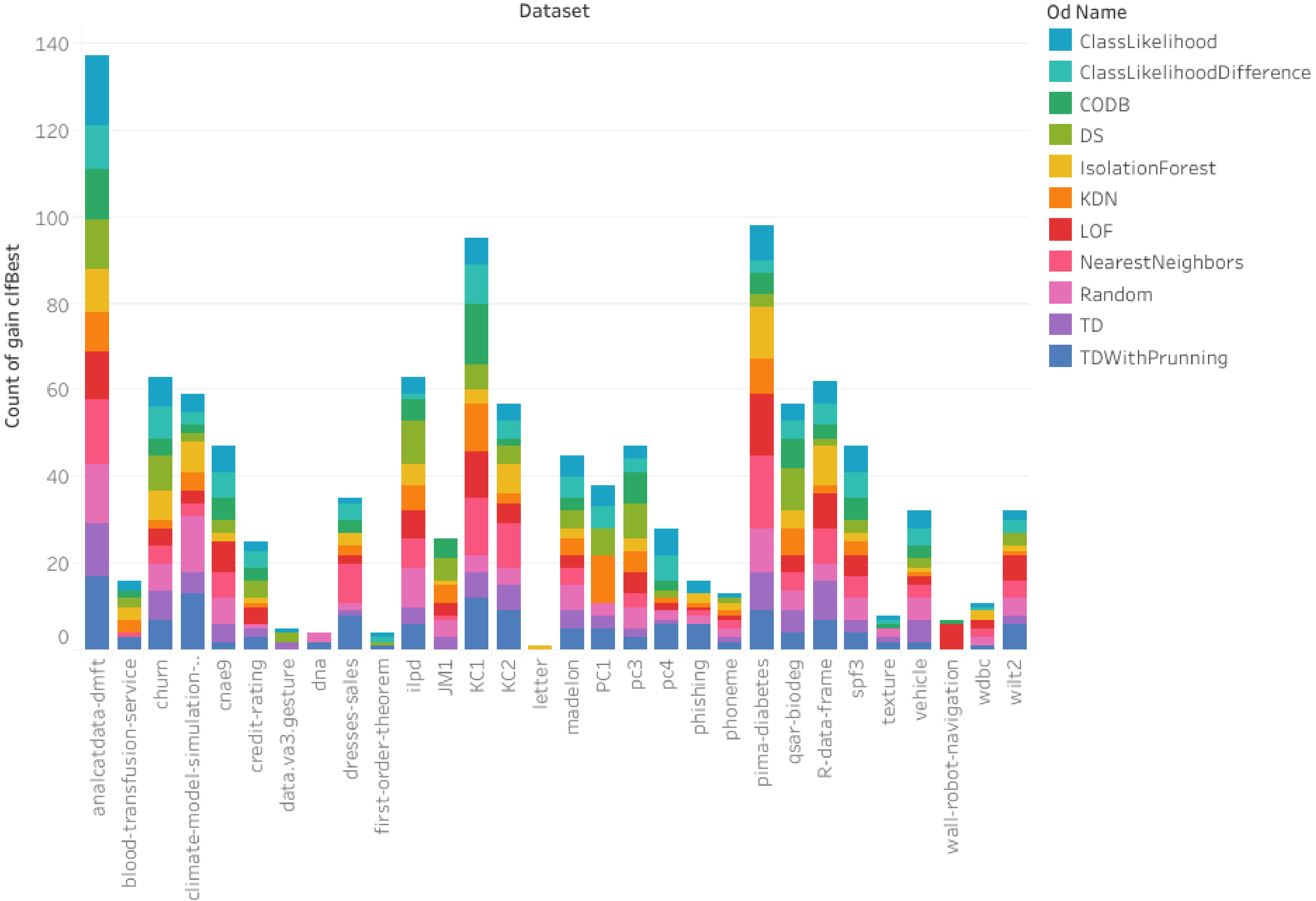


Removed
Count of gain clfBest
240
220
200
180
160
140
120
100
80
60
40
20
0
0.5
1
2
3
4
5

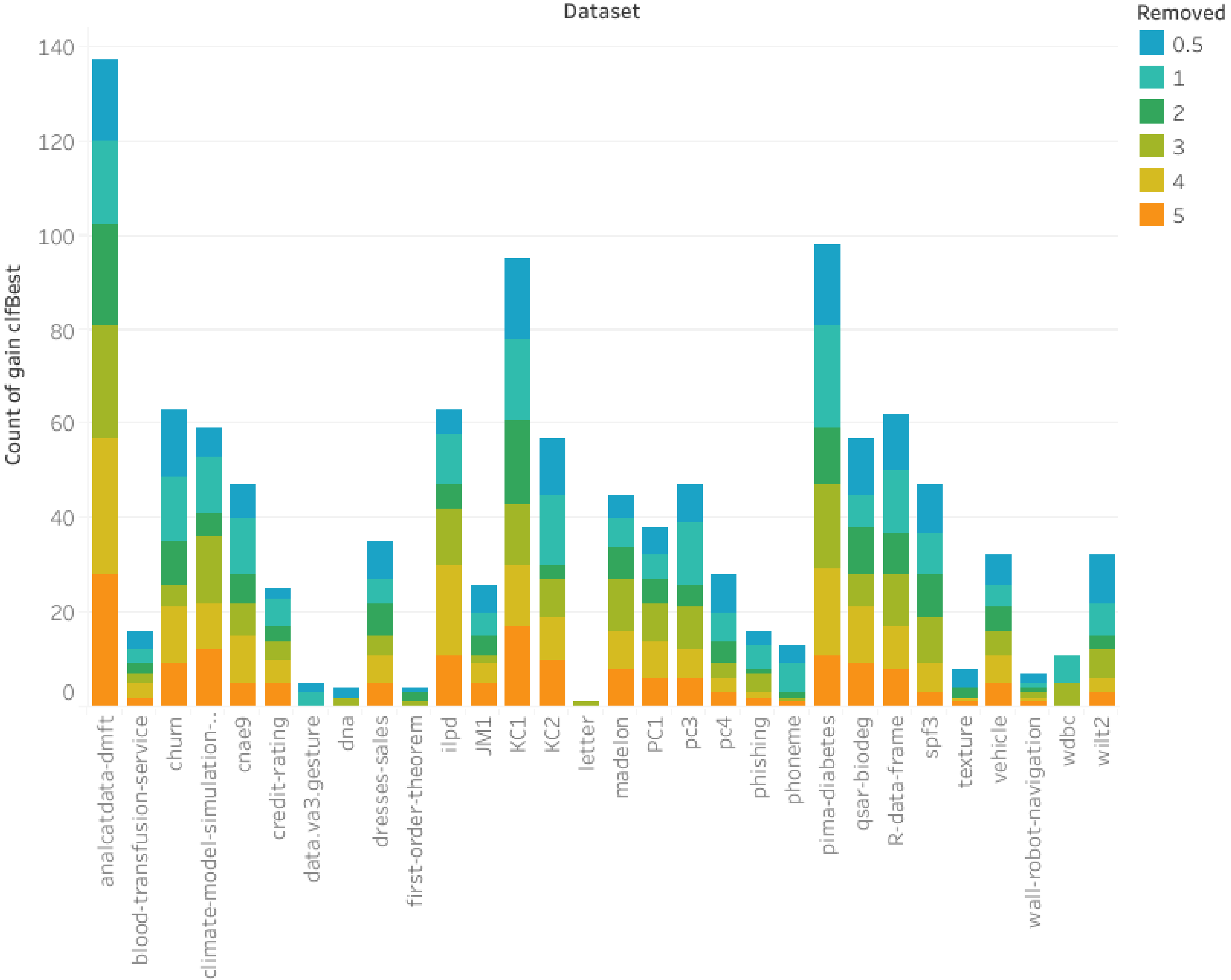
Sheet 3 (24)


Dataset
Count of gain clfBest
140
120
100
80
60
40
20
0
analcatdata-dmft
blood-transfusion-service
churn
climate-model-simulation-..
cnae9
credit-rating
data.va3.gesture
dna
dresses-sales
first-order-theorem
ilpd
JM1
KC1
KC2
letter
madelon
PC1
pc3
pc4
phishing
phoneme
pima-diabetes
qsar-biodeg
R-data-frame
spf3
texture
vehicle
wall-robot-navigation
wdbc
wilt2
Clf
J48
JRip
LMT
Logistic
NaiveBayes
PART
RandomForest
SimpleLogistic
SMO

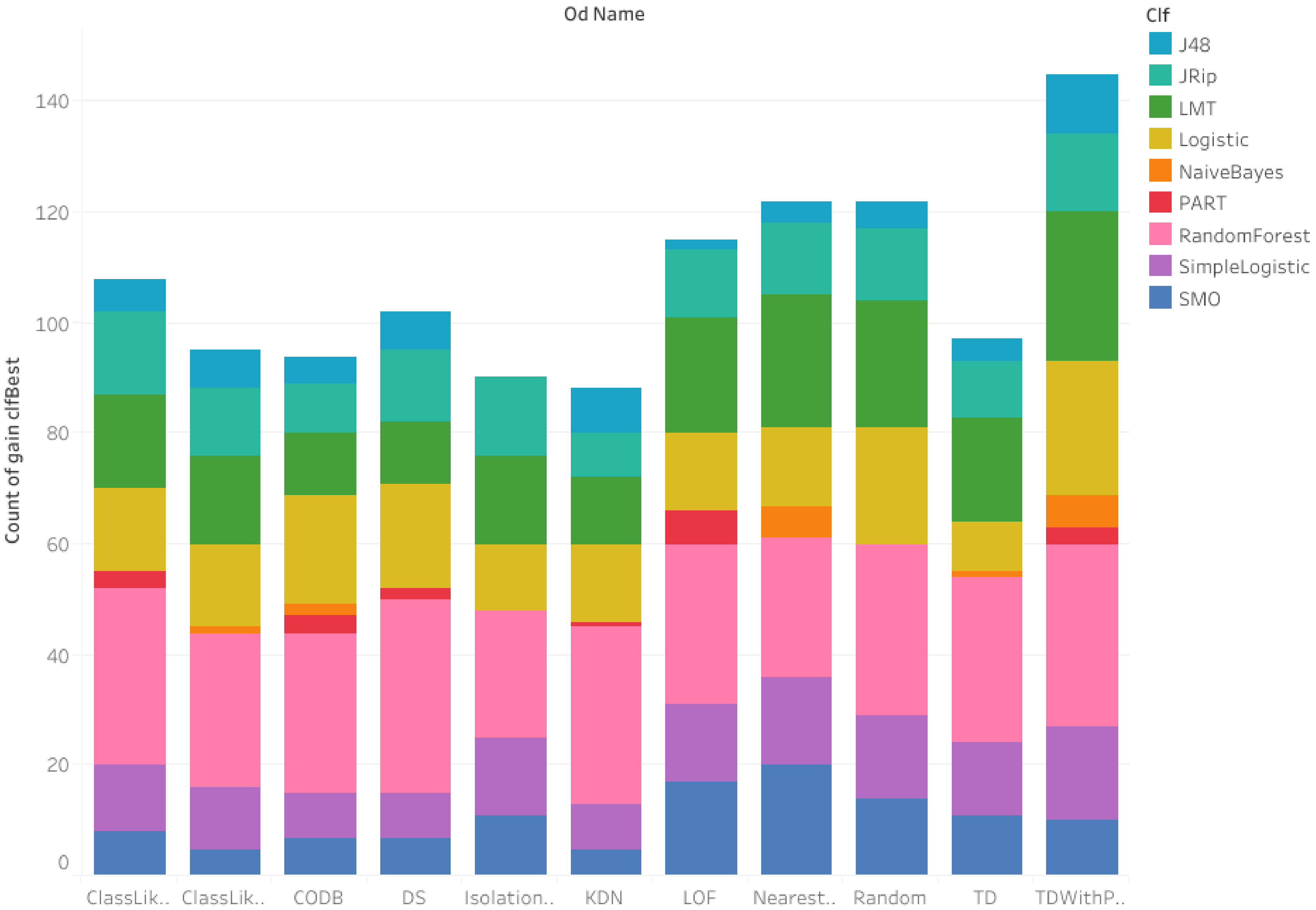
Sheet 3 (23)


Dataset
Count of gain clfBest
140
120
100
80
60
40
20
0
analcatdata-dmft
blood-transfusion-service
churn
climate-model-simulation-..
cnae9
credit-rating
data.va3.gesture
dna
dresses-sales
first-order-theorem
ilpd
JM1
KC1
KC2
letter
madelon
PC1
pc3
pc4
phishing
phoneme
pima-diabetes
qsar-biodeg
R-data-frame
spf3
texture
vehicle
wall-robot-navigation
wdbc
wilt2
Od Name
ClassLikelihood
ClassLikelihoodDifference
CODB
DS
IsolationForest
KDN
LOF
NearestNeighbors
Random
TD
TDWithPrunning

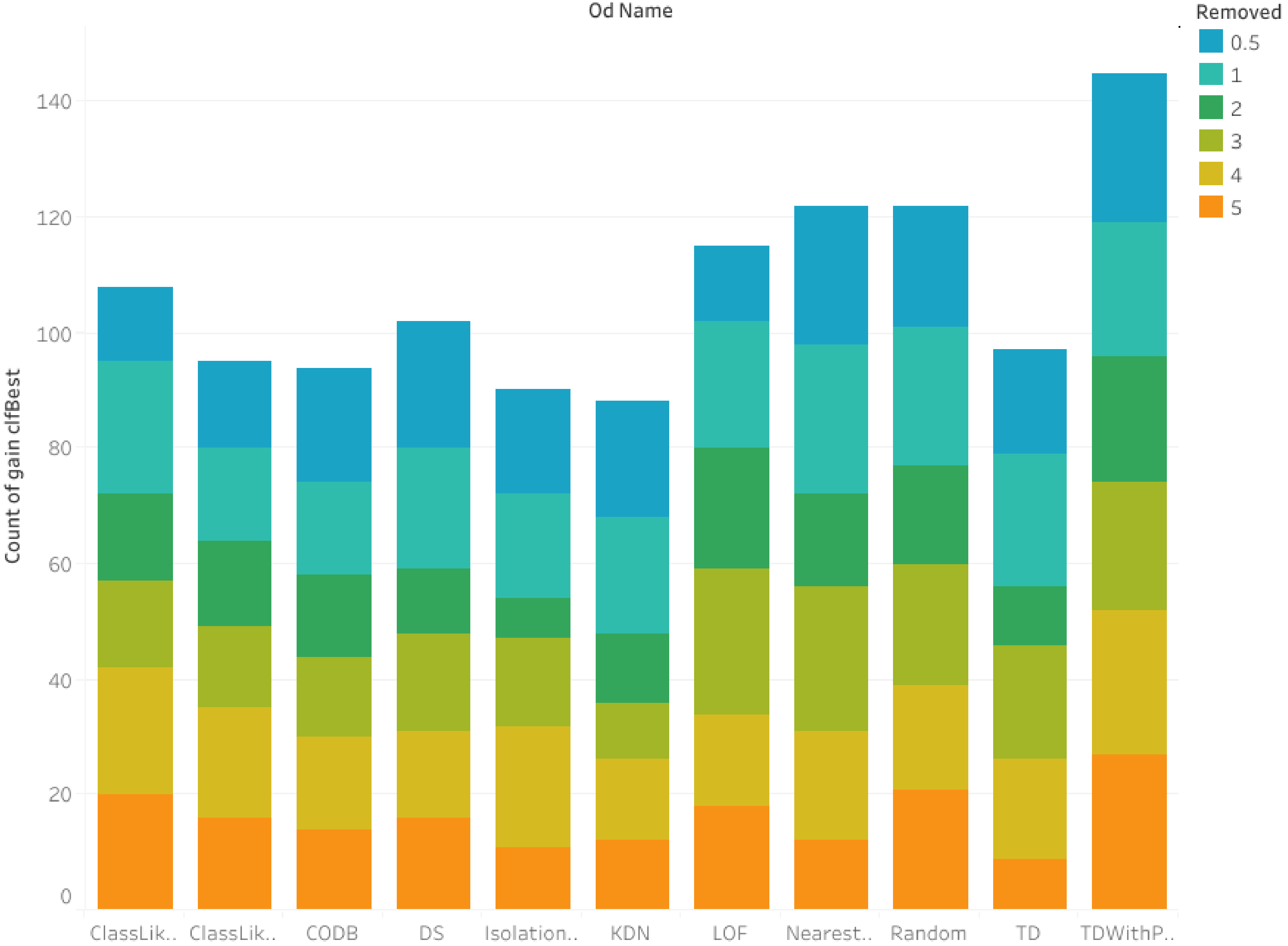
Sheet 3 (22)


Dataset
Removed
0.5
1
2
3
4
5
Count of gain clfBest
140
120
100
80
60
40
20
0
analcatdata-dmft
blood-transfusion-service
churn
climate-model-simulation-..
cnae9
credit-rating
data.va3.gesture
dna
dresses-sales
first-order-theorem
ilpd
JM1
KC1
KC2
letter
madelon
PC1
pc3
pc4
phishing
phoneme
pima-diabetes
qsar-biodeg
R-data-frame
spf3
texture
vehicle
wall-robot-navigation
wdbc
wilt2

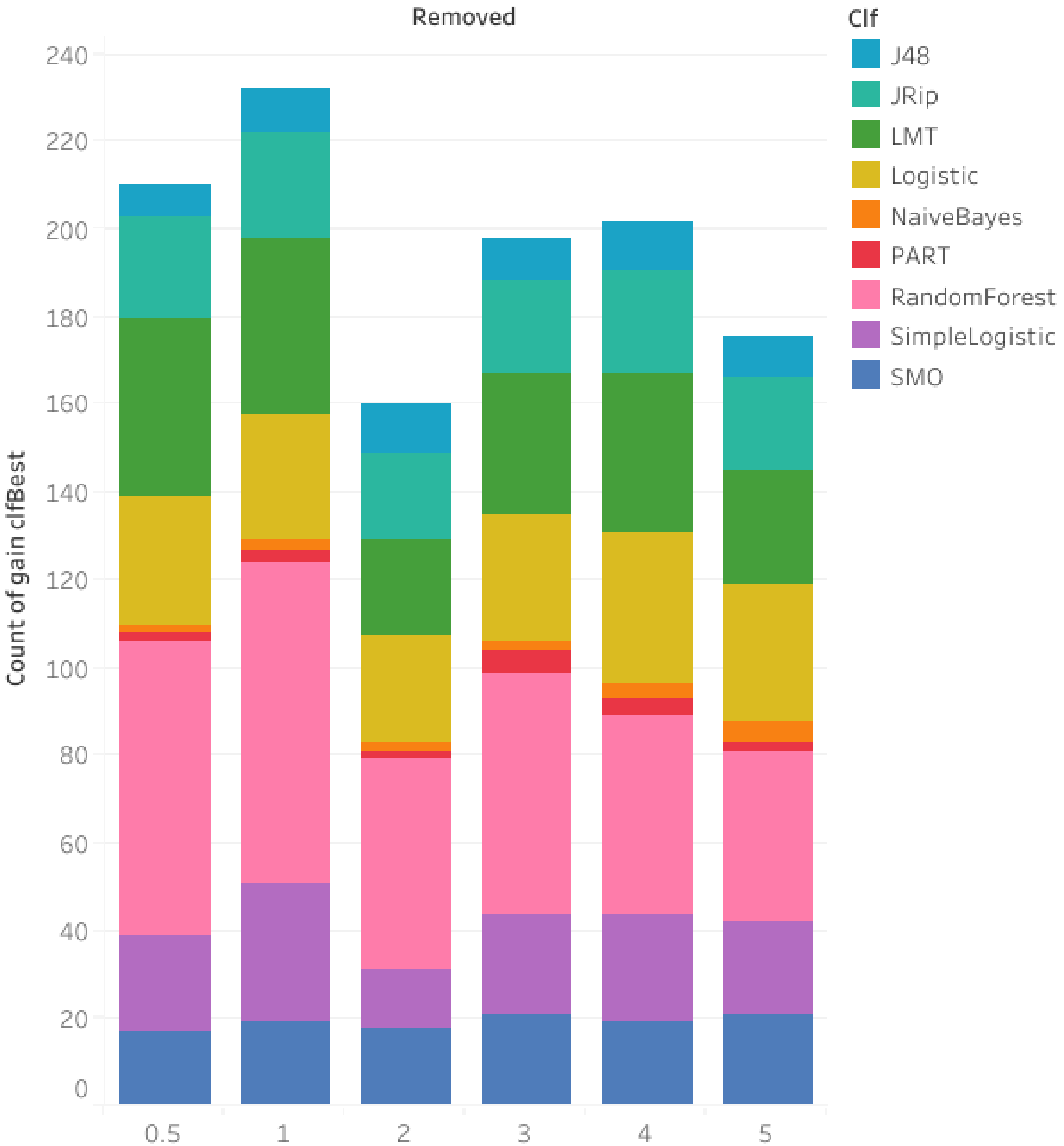
Sheet 3 (29)


Od Name
Count of gain clfBest
0
20
40
60
80
100
120
140
ClassLik..
ClassLik..
CODB
DS
Isolation..
KDN
LOF
Nearest..
Random
TD
TDWithP..
Clf
J48
JRip
LMT
Logistic
NaiveBayes
PART
RandomForest
SimpleLogistic
SMO

Sheet 3 (30)


Od Name
Removed
0.5
1
2
3
4
5
Count of gain clfBest
0
20
40
60
80
100
120
140
ClassLik..
ClassLik..
CODB
DS
Isolation..
KDN
LOF
Nearest..
Random
TD
TDWithP..

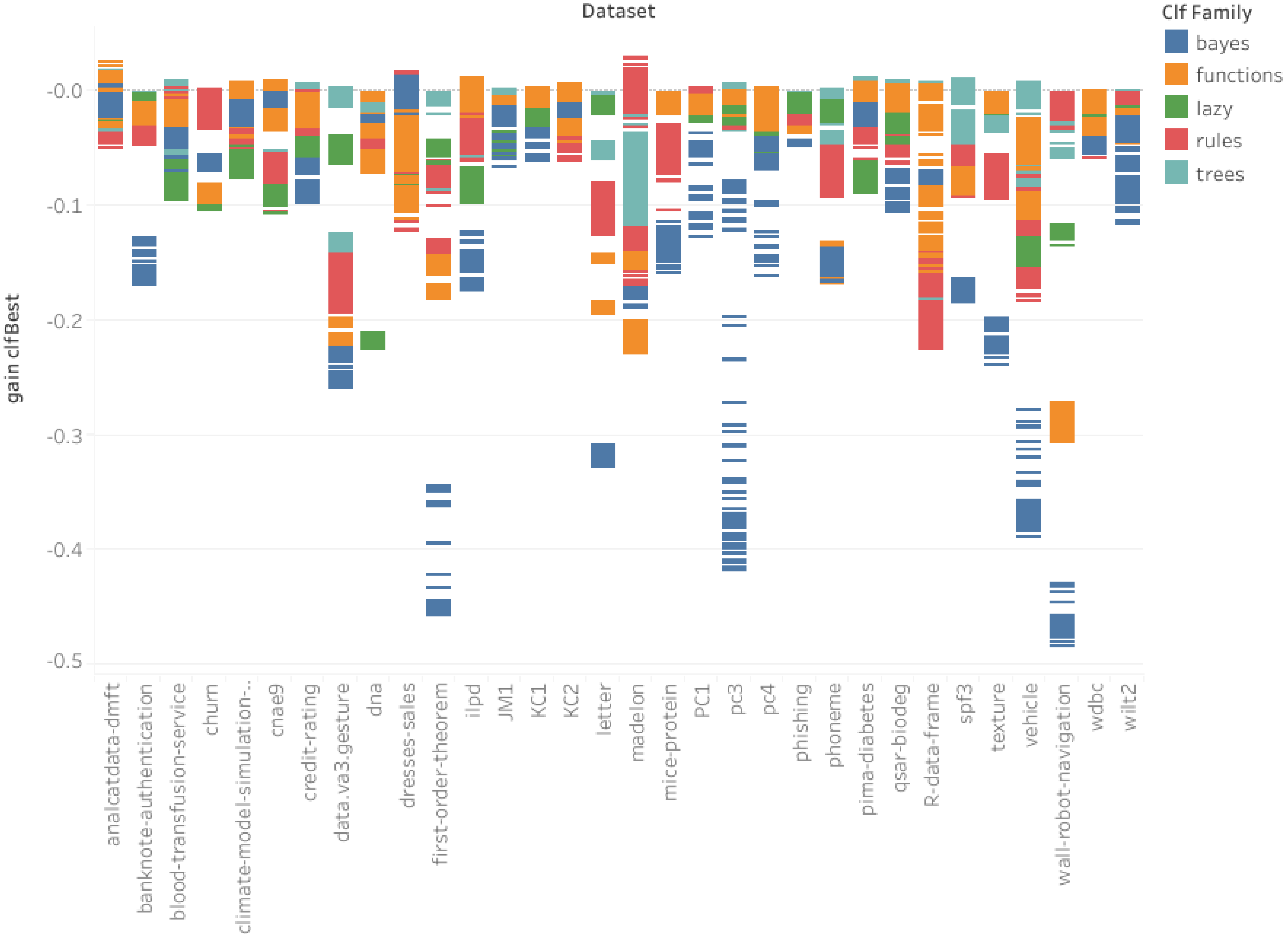
Sheet 3 (28)



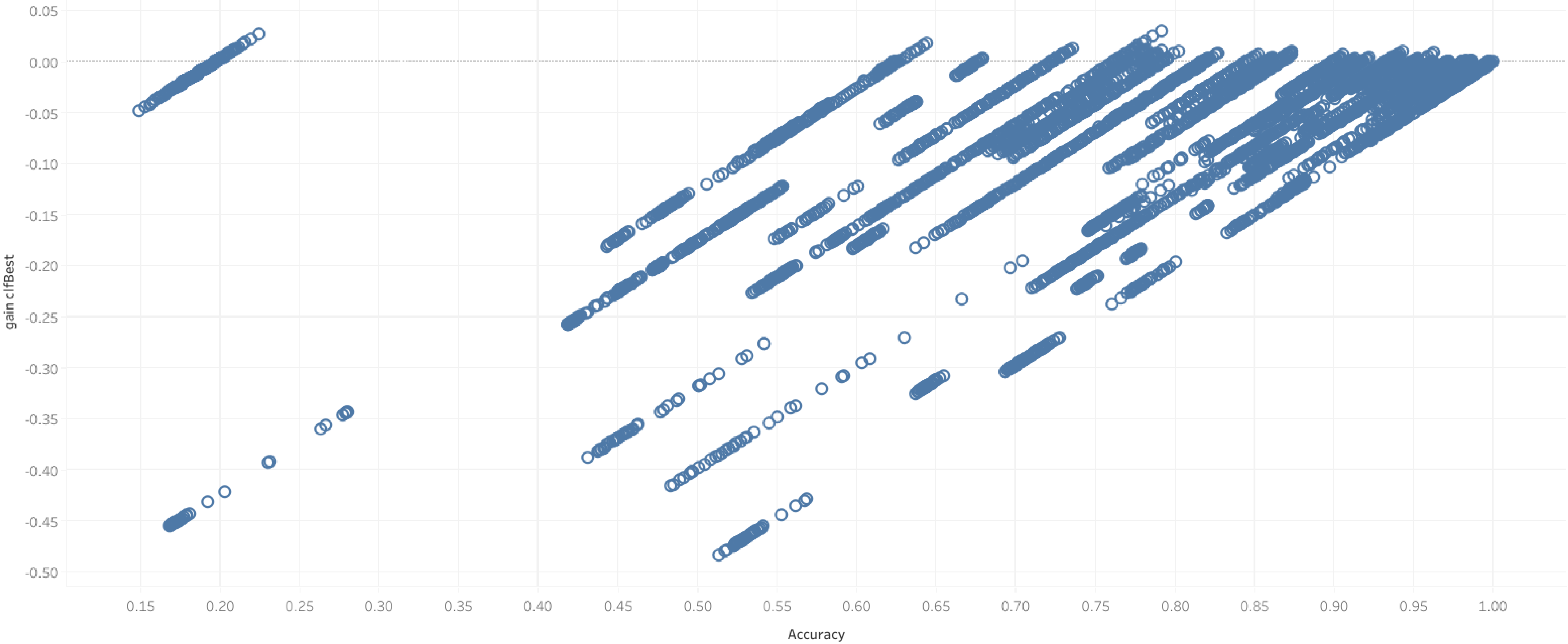
Removed
Count of gain clfBest
0
20
40
60
80
100
120
140
160
180
200
220
240
0.5
1
2
3
4
5
Clf
J48
JRip
LMT
Logistic
NaiveBayes
PART
RandomForest
SimpleLogistic
SMO

Other graphs

Sheet 3 (3)

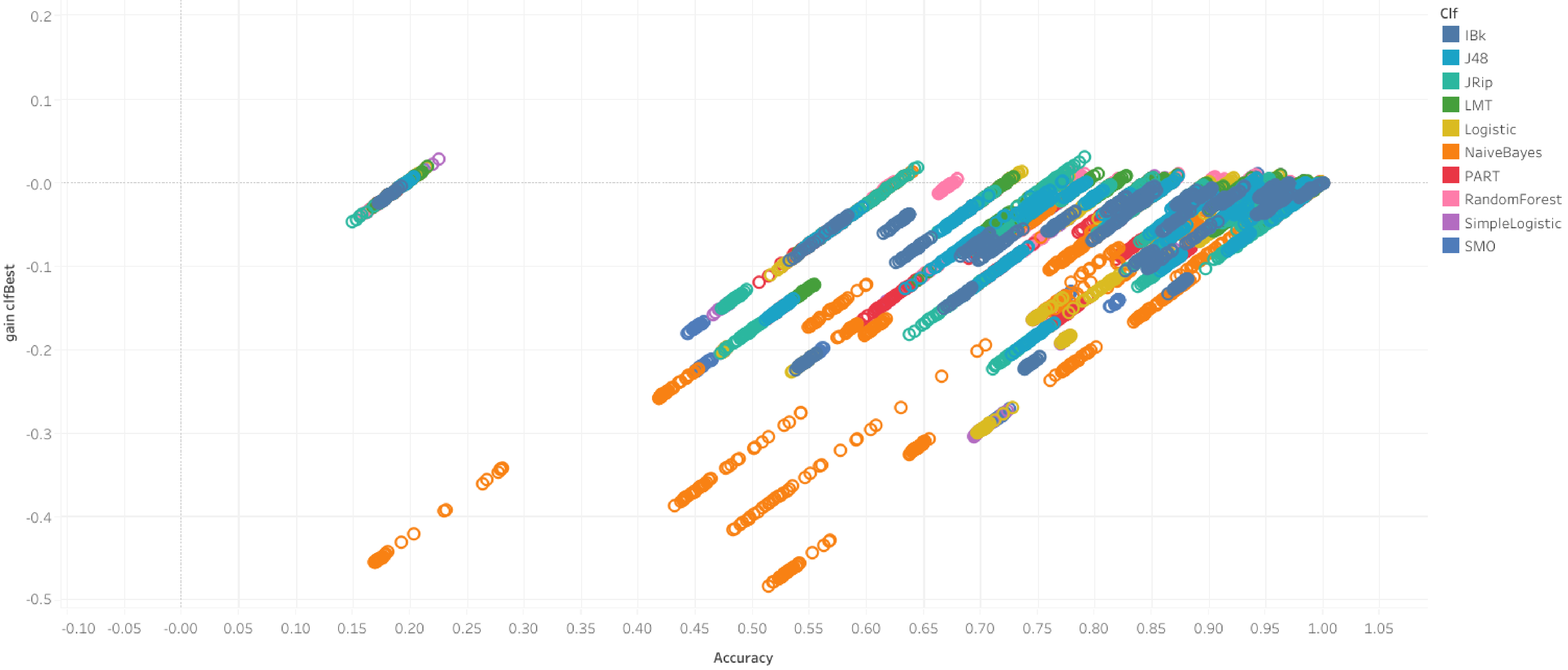

Dataset
Clf Family
bayes
functions
lazy
rules
trees
gain clfBest
-0.0
-0.1
-0.2
-0.3
-0.4
-0.5
analcatdata-dmft
banknote-authentication
blood-transfusion-service
churn
climate-model-simulation-..
cnae9
credit-rating
data.va3.gesture
dna
dresses-sales
first-order-theorem
ilpd
JM1
KC1
KC2
letter
madelon
mice-protein
PC1
pc3
pc4
phishing
phoneme
pima-diabetes
qsar-biodeg
R-data-frame
spf3
texture
vehicle
wall-robot-navigation
wdbc
wilt2

Sheet 3 (10)



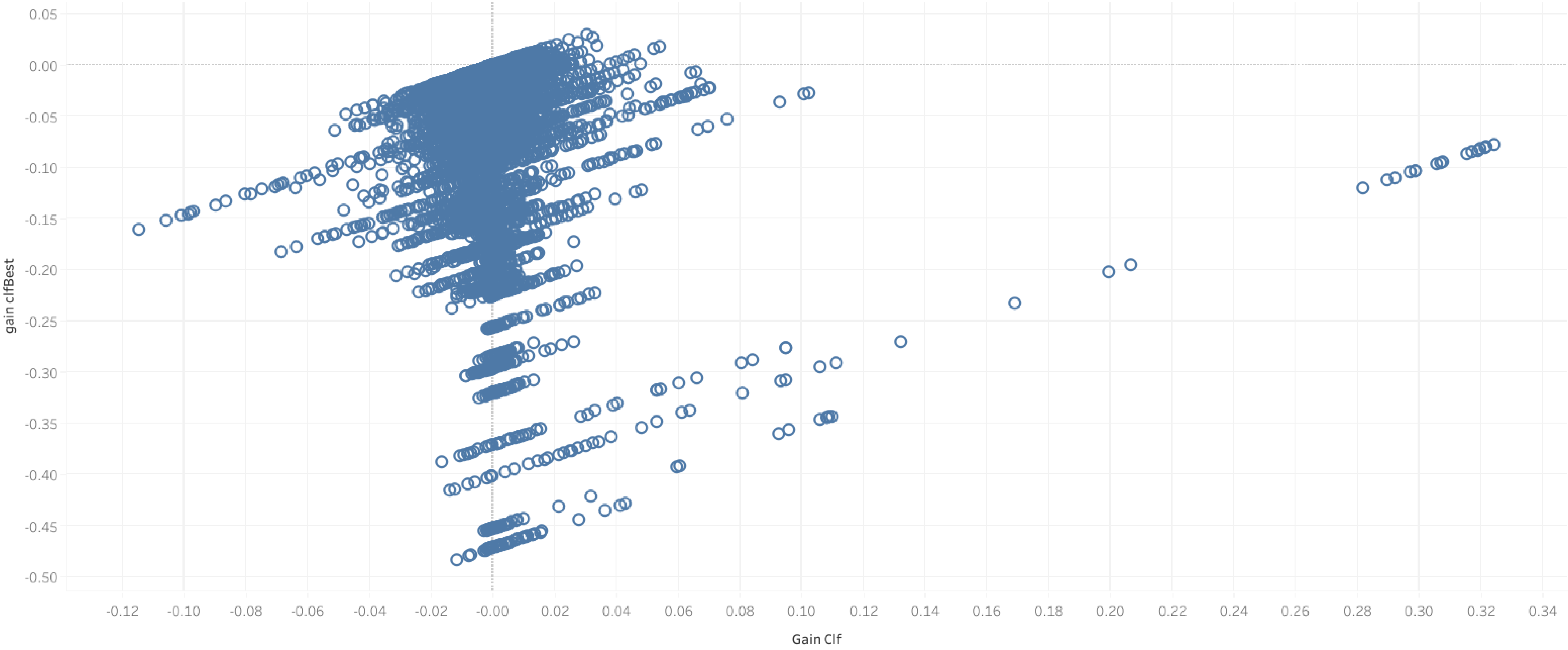
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
-0.35
-0.40
-0.45
-0.50
gain clfBest
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Accuracy

Sheet 3 (18)



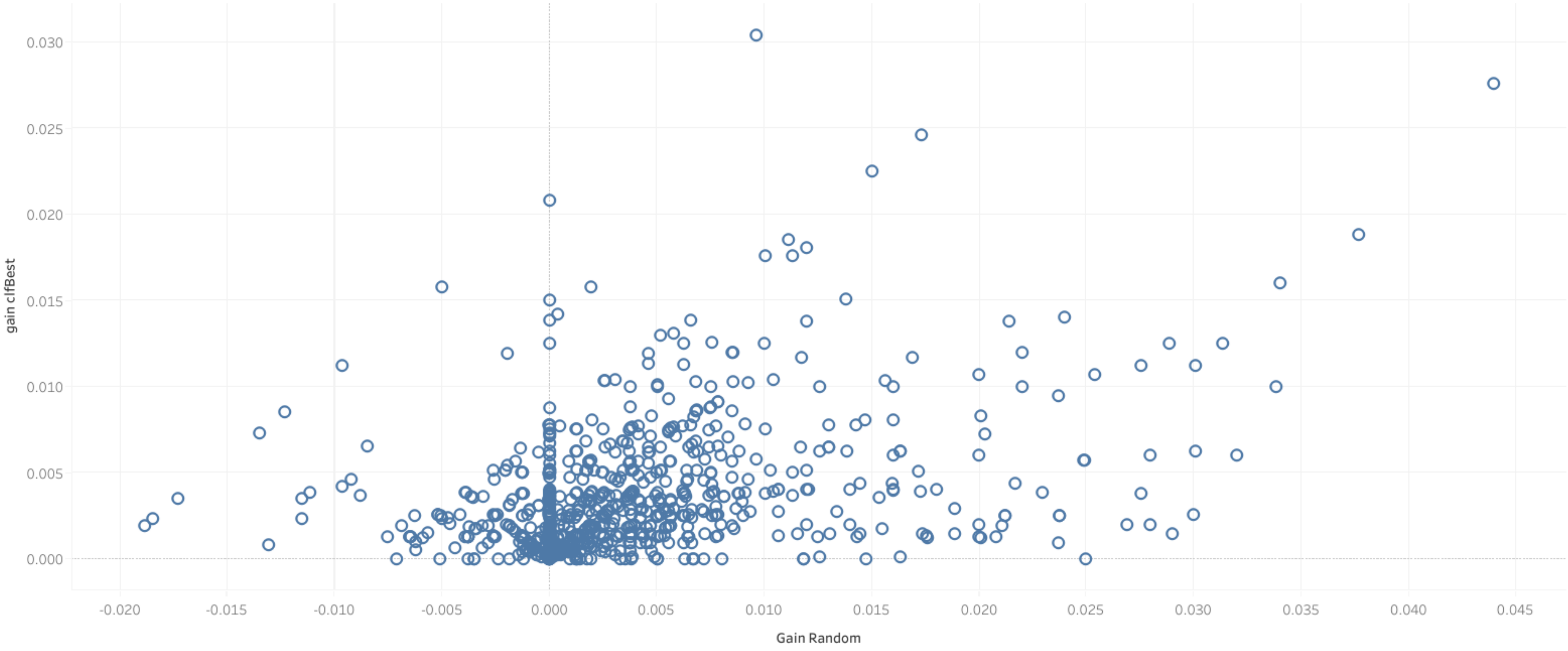
gain clfBest
0.2
0.1
-0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.10
-0.05
-0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
1.05
Accuracy
Clf
IBk
J48
JRip
LMT
Logistic
NaiveBayes
PART
RandomForest
SimpleLogistic
SMO

Sheet 3 (9)

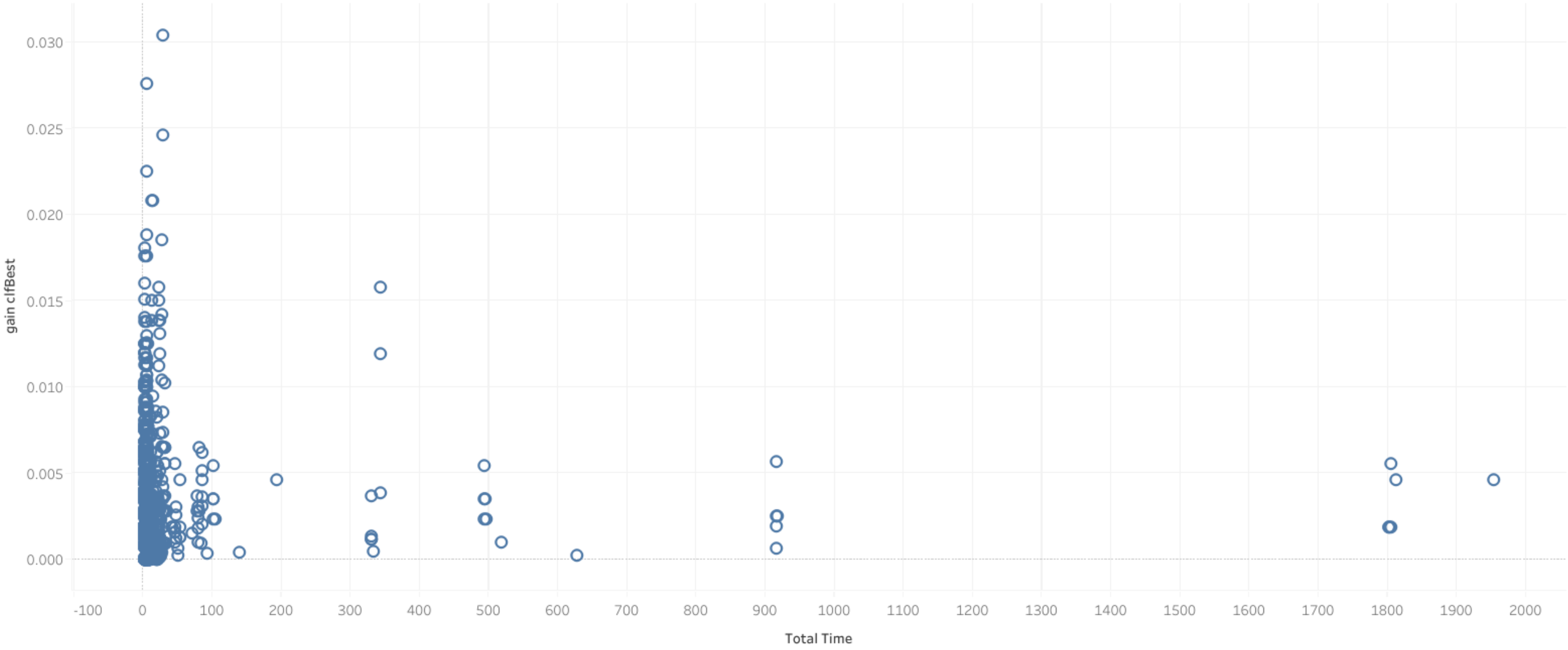


gain clfBest
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
-0.35
-0.40
-0.45
-0.50
-0.12
-0.10
-0.08
-0.06
-0.04
-0.02
-0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
0.24
0.26
0.28
0.30
0.32
0.34
Gain Clf

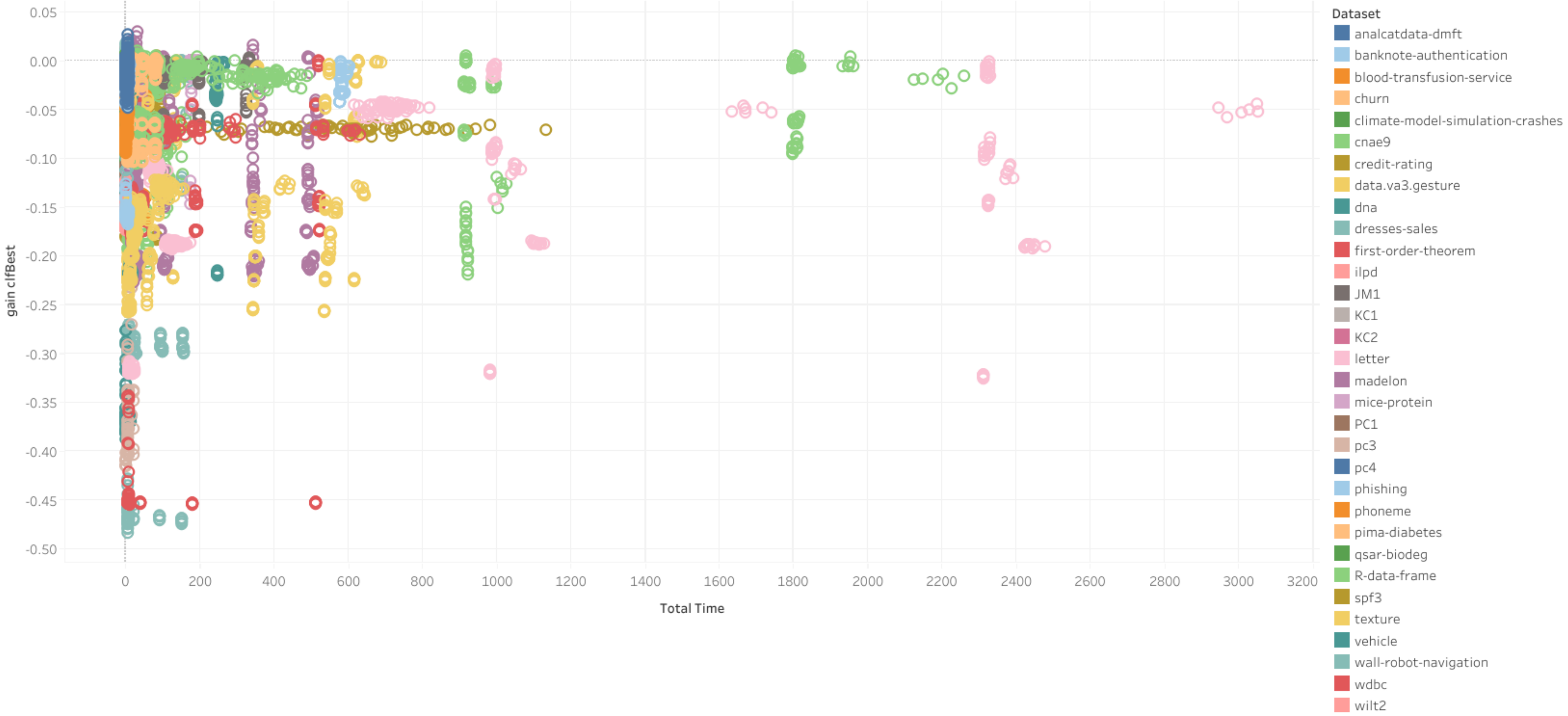
Sheet 3 (8)


0.030
0.025
0.020
0.015
0.010
0.005
0.000
gain clfBest
-0.020
-0.015
-0.010
-0.005
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
0.045
Gain Random

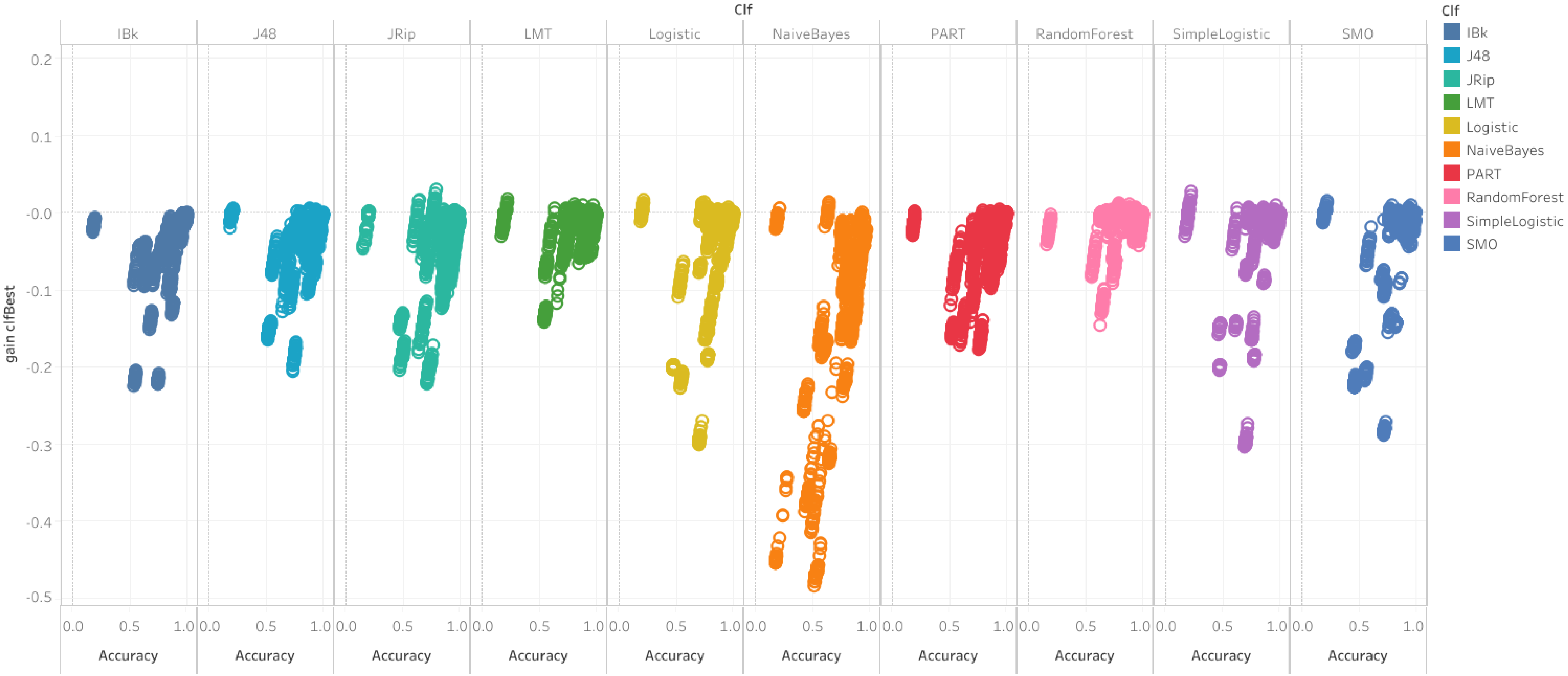
Sheet 3 (5)


gain clfBest
0.030
0.025
0.020
0.015
0.010
0.005
0.000
-100
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
Total Time

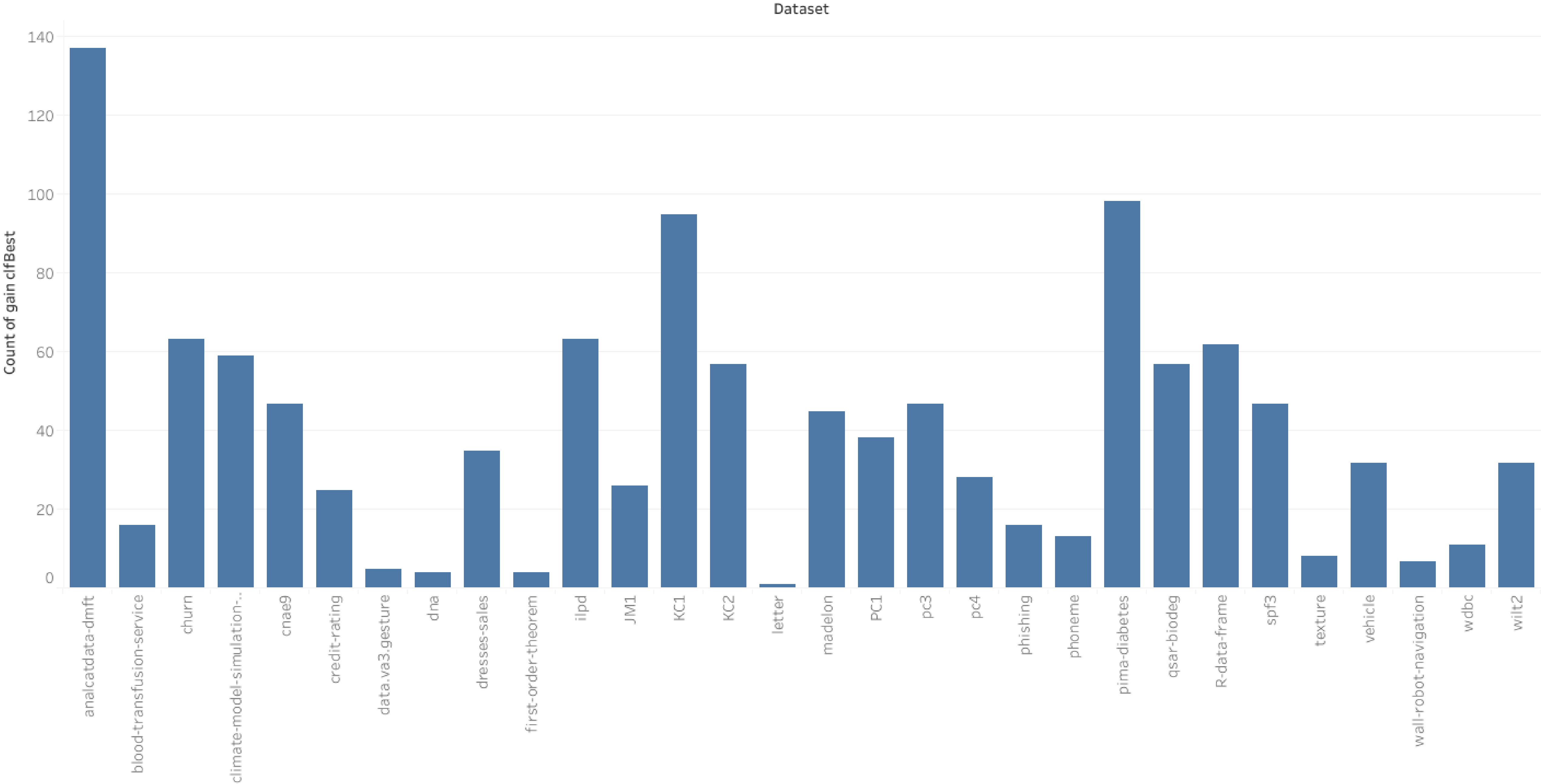
Sheet 3 (5)


gain clfBest
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
-0.35
-0.40
-0.45
-0.50
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
Total Time
Dataset
analcatdata-dmft
banknote-authentication
blood-transfusion-service
churn
climate-model-simulation-crashes
cnae9
credit-rating
data.va3.gesture
dna
dresses-sales
first-order-theorem
ilpd
JM1
KC1
KC2
letter
madelon
mice-protein
PC1
pc3
pc4
phishing
phoneme
pima-diabetes
qsar-biodeg
R-data-frame
spf3
texture
vehicle
wall-robot-navigation
wdbc
wilt2

Sheet 3 (19)


Clf
IBk
J48
JRip
LMT
Logistic
NaiveBayes
PART
RandomForest
SimpleLogistic
SMO
gain clfBest
0.2
0.1
-0.0
-0.1
-0.2
-0.3
-0.4
-0.5
0.0
0.5
1.0
Accuracy
Clf
IBk
J48
JRip
LMT
Logistic
NaiveBayes
PART
RandomForest
SimpleLogistic
SMO

Sheet 3 (21)


Dataset
Count of gain clfBest
140
120
100
80
60
40
20
0
analcatdata-dmft
blood-transfusion-service
churn
climate-model-simulation-..
cnae9
credit-rating
data.va3.gesture
dna
dresses-sales
first-order-theorem
ilpd
JM1
KC1
KC2
letter
madelon
PC1
pc3
pc4
phishing
phoneme
pima-diabetes
qsar-biodeg
R-data-frame
spf3
texture
vehicle
wall-robot-navigation
wdbc
wilt2

- https://docs.openml.org/
- https://www.openml.org/
- https://www.openml.org/search?type=data
- https://www.openml.org/d/50
- https://www.openml.org/search?type=task
- https://www.openml.org/t/145804
- https://www.openml.org/search?type=flow
- https://www.openml.org/f/7791
- https://www.openml.org/search?type=run
- https://www.openml.org/r/23504
- https://docs.openml.org/APIs/
- https://en.wikipedia.org/wiki/Cohen's_kappa
- https://en.wikipedia.org/wiki/Receiver_operating_characteristic#:~:text=A%20receiver%20operating%20characteristic%20curve,why%20it%20is%20so%20named.
- Besim BILALLI, Alberto ABELL´O, Tom´as ALUJA-BANET. On the predictive power of meta-features in OpenML. International Journal of Applied Mathematics and Computer Science. 2017, Vol 27, Iss 4, 2083-8492.