Universiteit Leiden

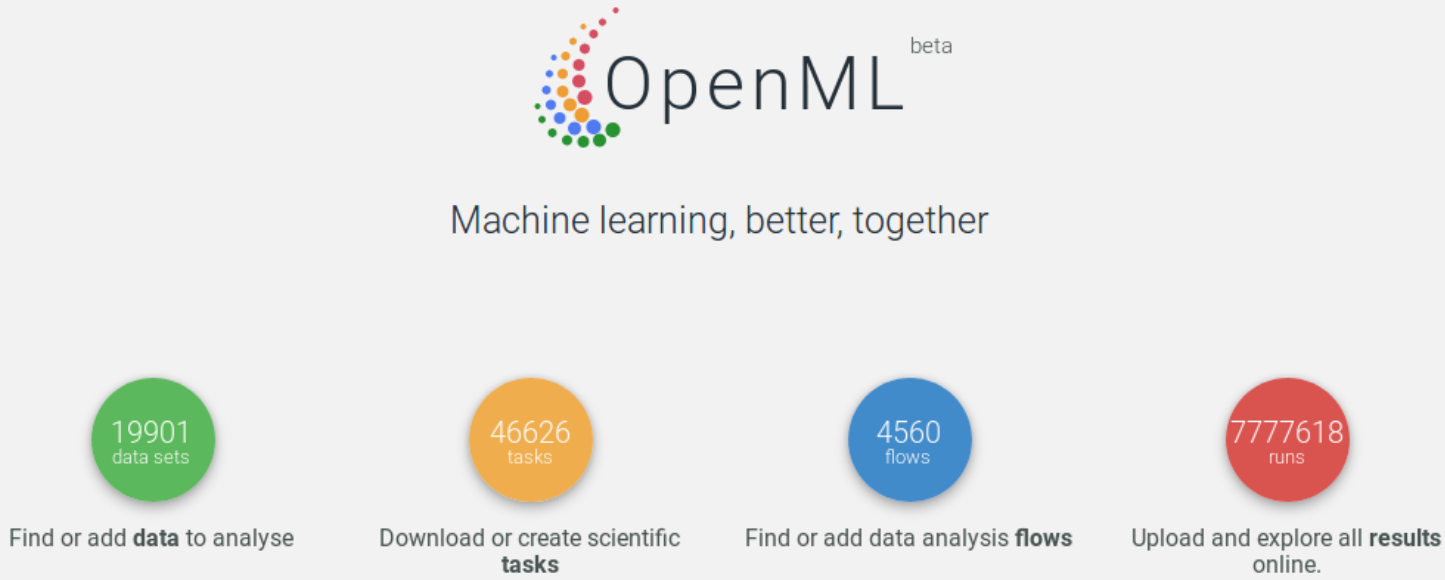
Metalearning: Advanced Topics

Pavel Brazdil, Jan N. van Rijn, Henry Gouk, Felix Mohr

Meta-Knowledge Transfer Communication in Different Systems

September 23rd, 2022

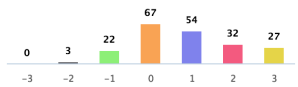
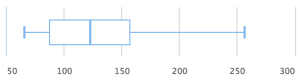
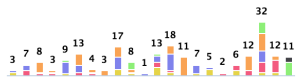
Openml.org


OpenML
beta
Machine learning, better, together
19901
data sets
Find or add **data** to analyse
46626
tasks
Download or create scientific **tasks**
4560
flows
Find or add data analysis **flows**
7777618
runs
Upload and explore all **results** online.

Datasets

- Data (ARFF) uploaded or referenced, versioned
- Analysed, characterized, organized on line
- Indexed based on name, meta-features, tags, etc.
- Support for other data formats (on request)

26 features

symboling **(target)**	nominal	6 unique values 0 missing	
normalized-losses	numeric	51 unique values 41 missing	
make	nominal	22 unique values 0 missing	

Show all 26 features

72 properties

DefaultAccuracy	**0.33**	The predictive accuracy obtained by simply predicting the ...
NumberOfClasses	**7**	The number of classes in the class attribute.
NumberOfFeatures	**26**	The number of features (attributes) in the dataset. Also kn...
NumberOfInstances	**205**	The number of instances (examples) in the database.
NumberOfMissingVal...	**59**	Counts the total number of missing values in the dataset.
NumberOfNumericFe...	**15**	The number of symbolic features in the dataset.
NumberOfSymbolicF...	**10**	The number of symbolic features in the dataset.
ClassCount	**7**	DataQuality extracted from Fantail Library

sklearn.datasets.fetch_openml

sklearn.datasets. **fetch_openml** (*name=None, version='active', data_id=None, data_home=None, target_column='default-target', cache=True, return_X_y=False*) [source]

Fetch dataset from openml by name or dataset id.

Datasets are uniquely identified by either an integer ID or by a combination of name and version (i.e. there might be multiple versions of the 'iris' dataset). Please give either name or data_id (not both). In case a name is given, a version can also be provided.

Read more in the User Guide.

Note: EXPERIMENTAL

The API is experimental in version 0.20 (particularly the return value structure), and might have small backward-incompatible changes in future releases.

Parameters:

name : *str or None*

String identifier of the dataset. Note that OpenML can have multiple datasets with the same name.

version : *integer or 'active', default='active'*

Version of the dataset. Can only be provided if also `name` is given. If 'active' the oldest version that's still active is used. Since there may be more than one active version of a dataset, and those versions may fundamentally be different from one another, setting an exact version is highly recommended.

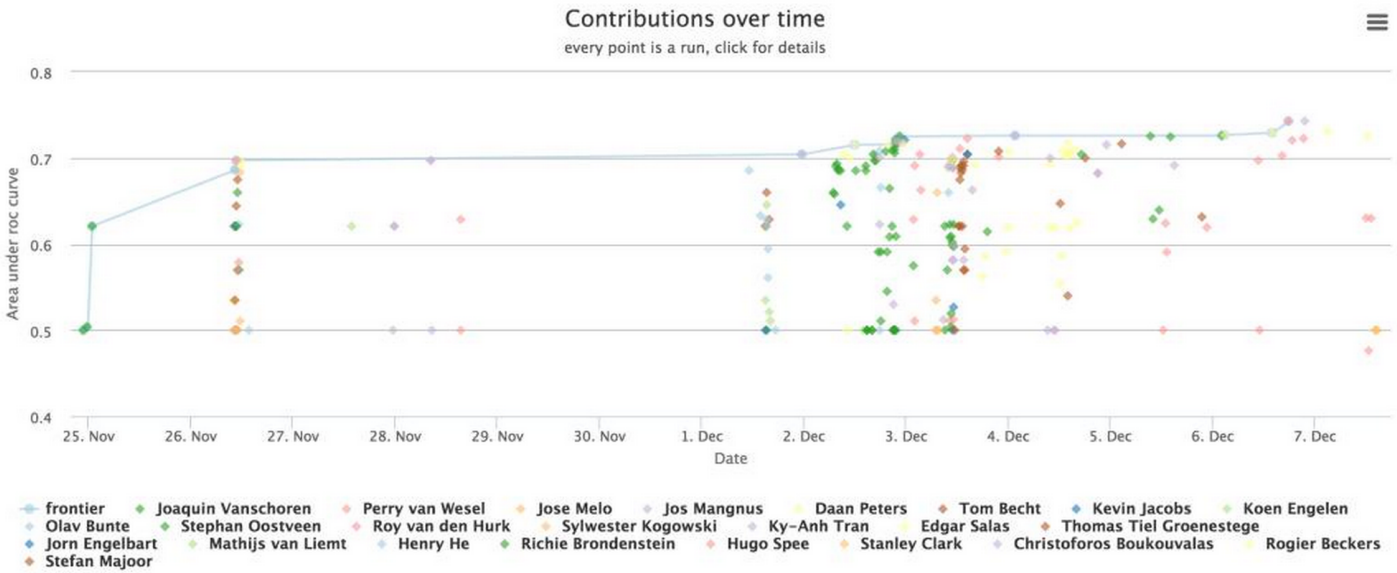
data_id : *int or None*

OpenML ID of the dataset. The most specific way of retrieving a dataset. If data_id is not given, name (and potential version) are used to obtain a dataset.

data_home : *string or None, default None*

Tasks

- Data alone does not define an experiment
- Tasks contain: data, target attribute, goals, procedures
- Readable by tools, automates experimentation
- Real time 'leaderboard' and overview



Tasks

Various Task Types:

- Supervised Classification
- Supervised Regression
- Learning Curve Classification (time intensive)
- Data Stream Classification (on line learning)
- Survival Analysis
- Clustering (Work in Progress)
- Machine Learning Challenge
- Easily extendable to more ..

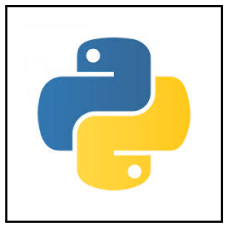
Flows (algorithms)

- Run locally, auto-registered by tools
- Integrations + APIs (REST, Java, R, Python, . . .)




Flows (algorithms)

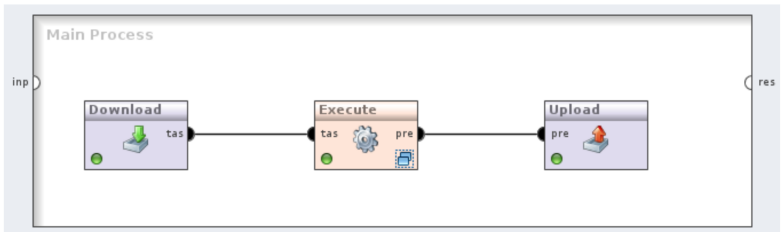
- Run locally, auto-registered by tools
- Integrations + APIs (REST, Java, R, Python, ...)



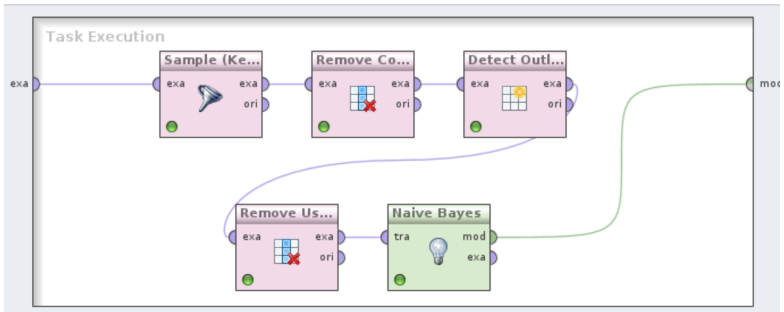

```
1 import openml
2 from sklearn import ensemble
3
4 task = openml.tasks.get_task(3954)
5 clf = ensemble.RandomForestClassifier()
6 run = openml.runs.run_model_on_task(task, clf)
7 result = run.publish()
```

Flows (algorithms)

- Run locally, auto-registered by tools
- Integrations + APIs (REST, Java, R, Python, . . .)



(a) Main Workflow



★ Runs

- Flow uploads predictions
- Predictions are evaluated on OpenML
- Reproducible, linked to data, flows and researcher
- Contains:
  - predictions
  - hyperparameter settings
  - model information
  - evaluation measures

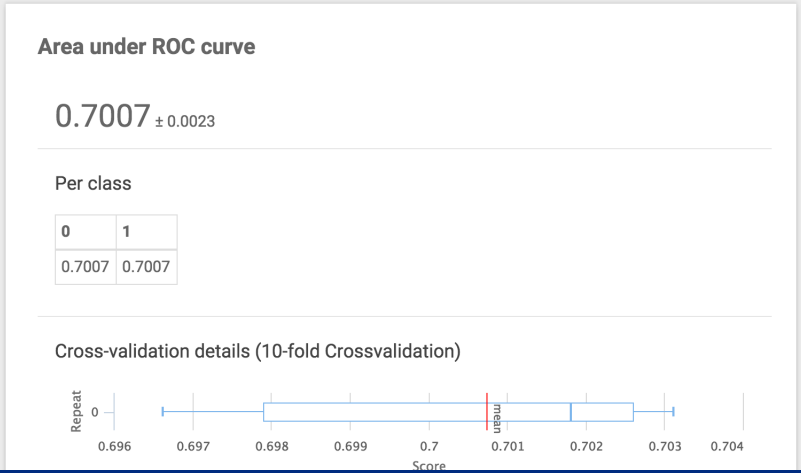
Result files

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

Model readable model
A human-readable description of the model that was built.

Model serialized model
A serialized description of the model that can be read by the tool that generated it.

Predictions arff
ARFF file with instance level predictions generated by the model



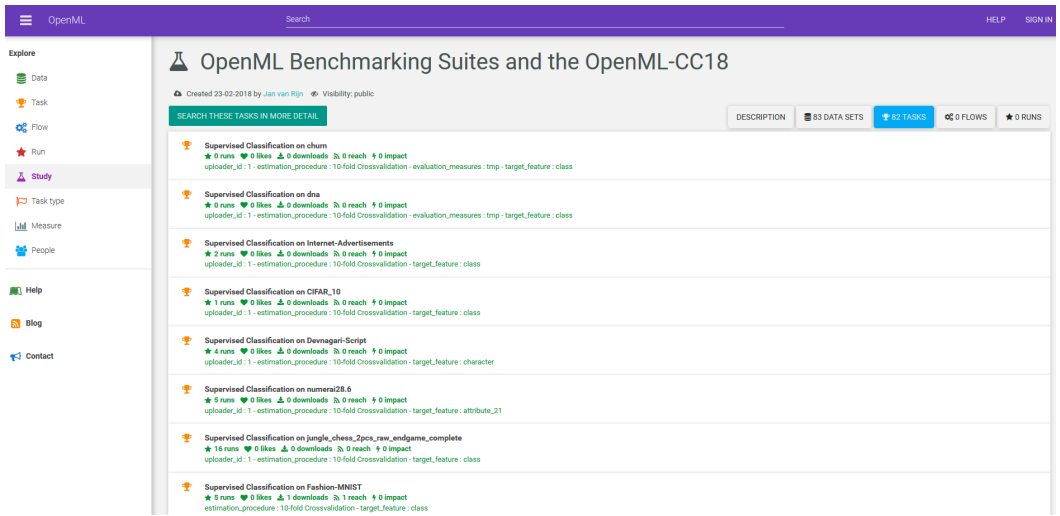
★ Runs

Answer basic questions about performance of algorithms to study . . .

- Computation is done client side
- The results (serialized model, predictions) are uploaded to OpenML
- OpenML evaluates the results and calculates default evaluation measures
- Users can calculate custom evaluation measures and share them on OpenML

Studies

- Bundles Data, Flows, Tasks and Runs
- Link attached to publication
- (Work in Progress): attach a notebook



Intermediate Summary

OpenML ...

- offers an eco-system to make Machine Learning experiments re-useable to other scientists
- is integrated in various programming languages and ML toolboxes, including Scikit-learn, Weka, ...
- has an active community; community meetings take place twice a year
- has more than 20,000 unique monthly users (and growing)

Projects

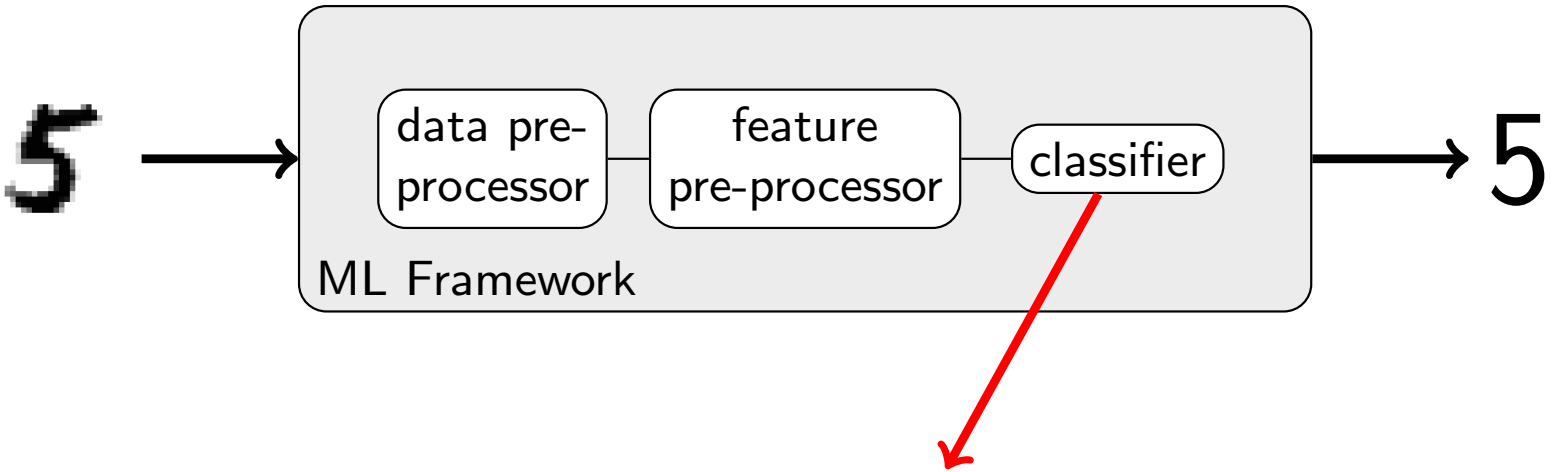
Several projects done with OpenML:

- OpenML Benchmark Suites [Bischl et al., 2021] - not today
- Hyperparameter Importance [van Rijn and Hutter, 2018]
- Myth Busting Urban Legends [Post et al., 2016, Strang et al., 2018] - if time allows
- (Learning good defaults) [Pfisterer et al., 2021] - not today

Universiteit Leiden

Hyperparameter Importance

Hyperparameter Optimization



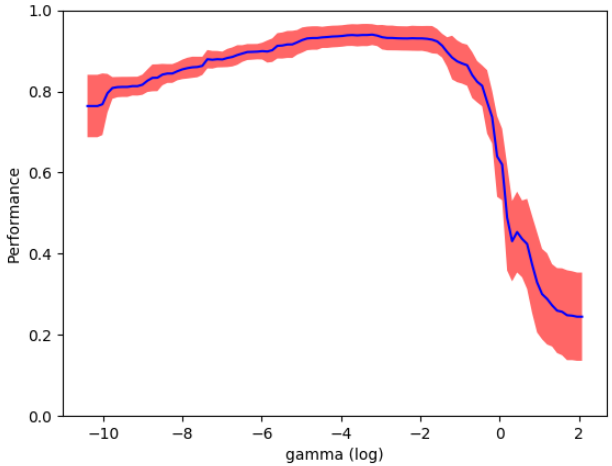
classifier	$\#\lambda$
Adaboost	4
Bernoulli Naive Bayes	2
Decision Tree	4
Extra Trees	5
Gradient Boosting	6
k-NN	3
LDA	4
...	

Hyperparameter Importance

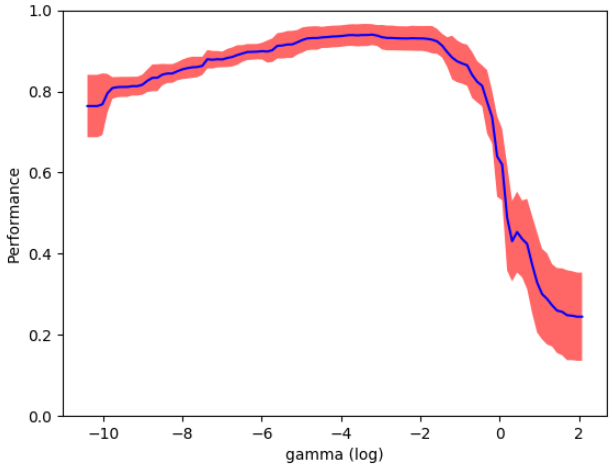
Experimental Setting by van Rijn and Hutter [2018]:

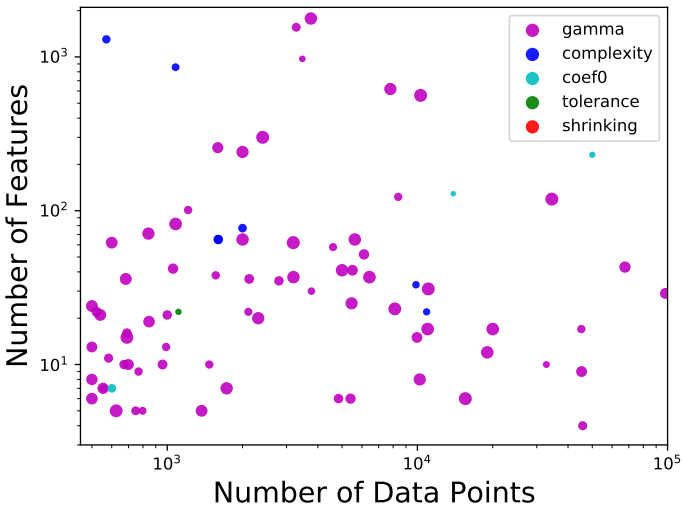
- All datasets from the OpenML-100 [Bischl et al., 2021]
- Four classifiers from scikit-learn (Random Forest, Adaboost and SVM with various kernels)
- Include results with at least 200 runs (try to generate if not enough)
- Functional ANOVA [Sobol, 1993, Hutter et al., 2014]
- Limitation: No conditional hyperparameters

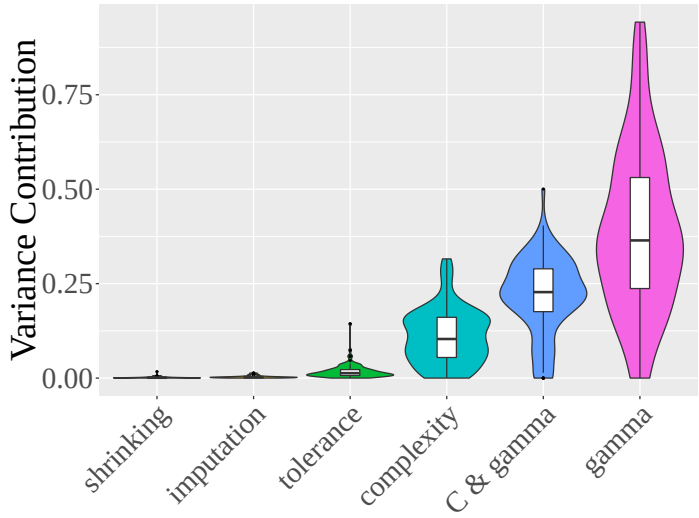
Hyperparameter Importance



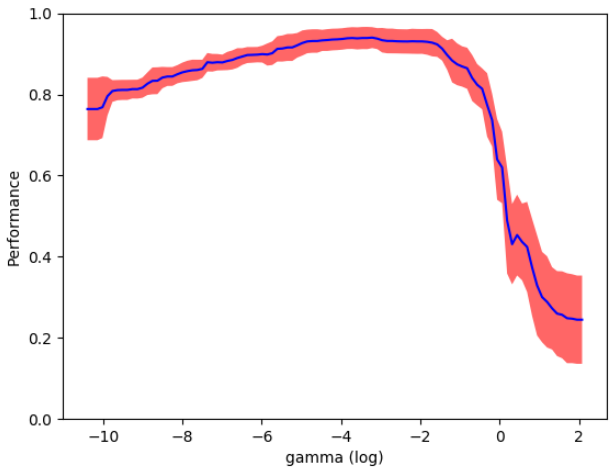
Hyperparameter Importance

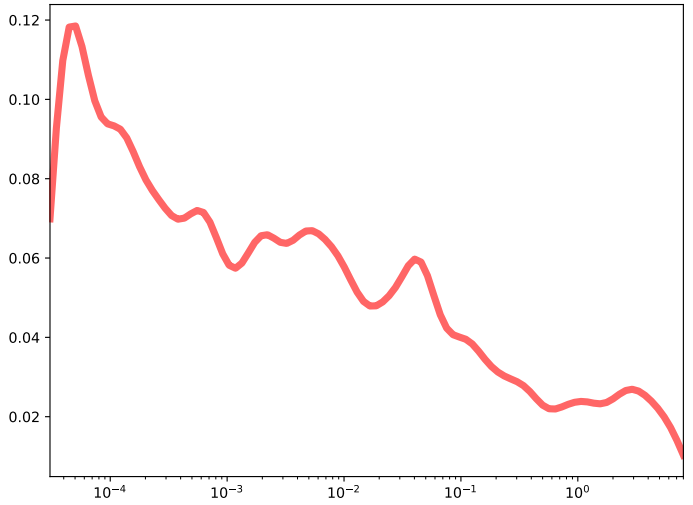


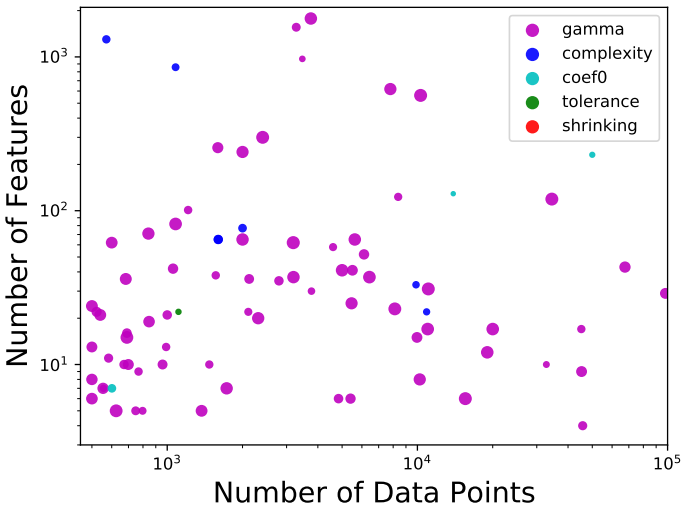


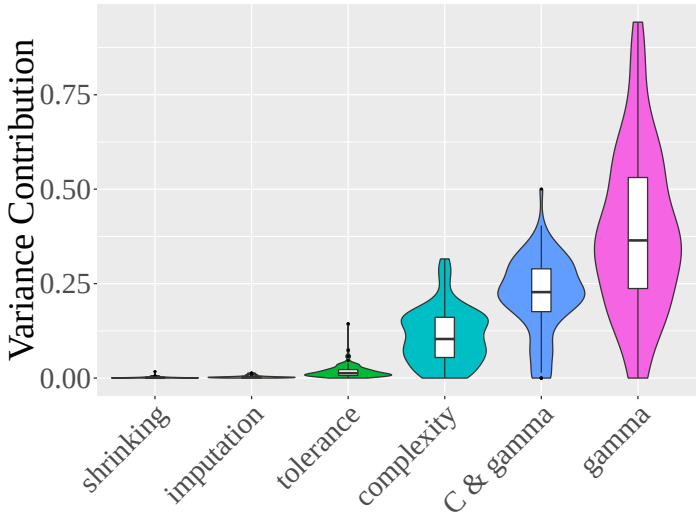


Hyperparameter Importance

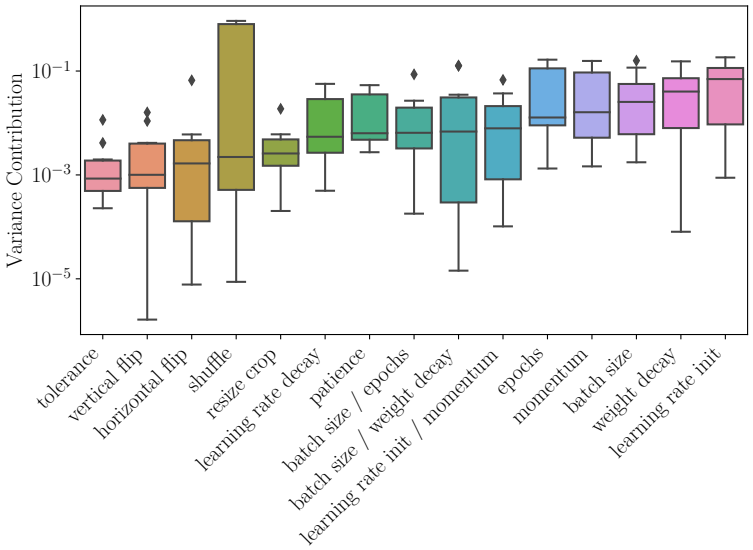




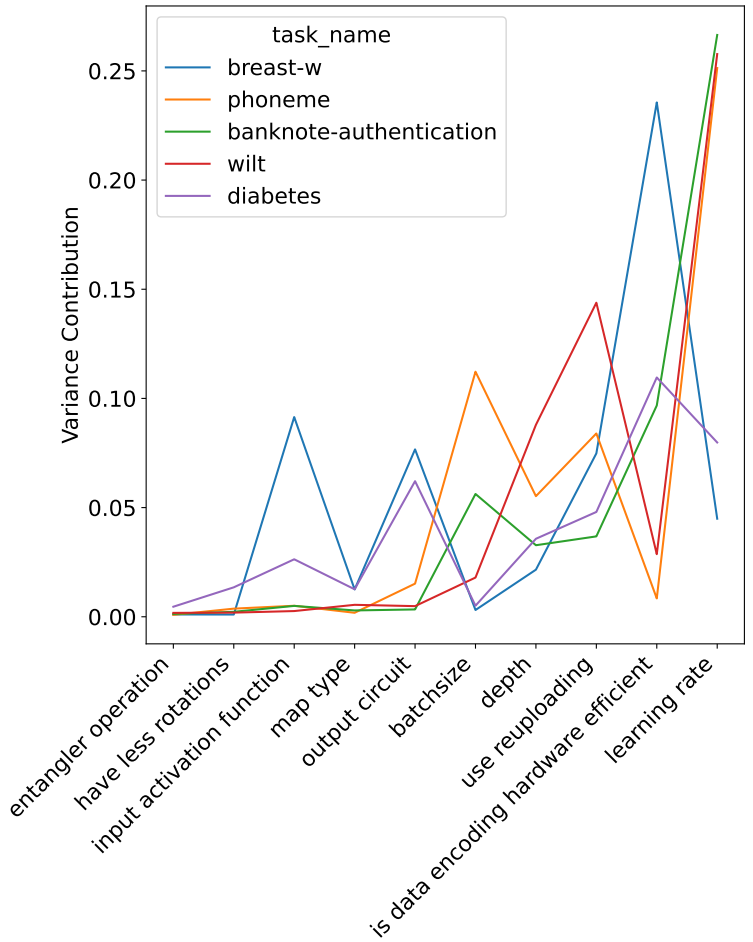




Hyperparameter Importance



[Sharma et al., 2019]

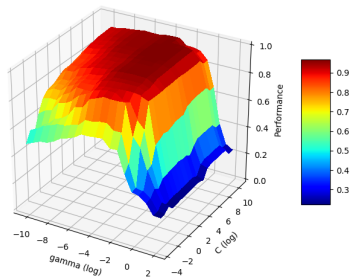


[Moussa et al., 2022]

Hyperparameter Importance

Initial conclusions from a study by van Rijn and Hutter [2018]:

- Functional ANOVA is a consistent tool for Hyperparameter Importance Analysis
- Obtained expected results (gamma and complexity) and new insights (Random Forest, Adaboost)
- Imputation of Missing Values
- Inferring priors leads to statistically significant better Hyperparameter Optimization results (Random Search and Hyperband)
- Video: `https://www.youtube.com/watch?v=mS4vL7_rSWQ`




BUSTING THE
MYTHS

Myth Busting for Data Mining

Universiteit Leiden

- Papers are generally build upon claims that are not well grounded, e.g.,
  - "We performed data transformation X because it is common practise."
  - "We set hyperparameter Y to value Z because the authors recommended these values."
- We can empirically analyze the validity of these claims on the meta-data from OpenML

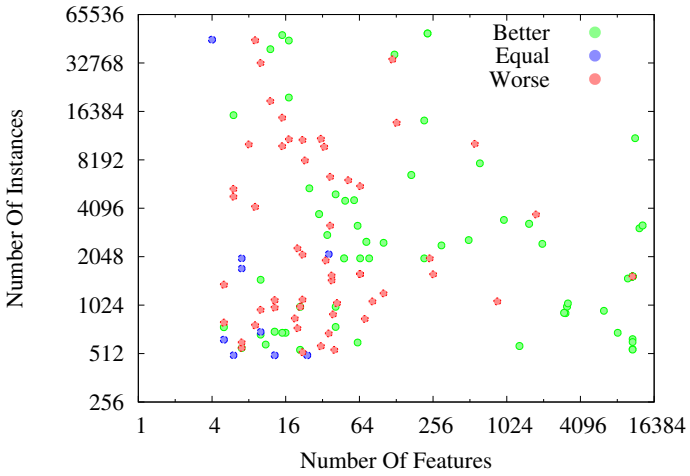
Effect of Feature Selection

Experimental Setting by Post et al. [2016]:

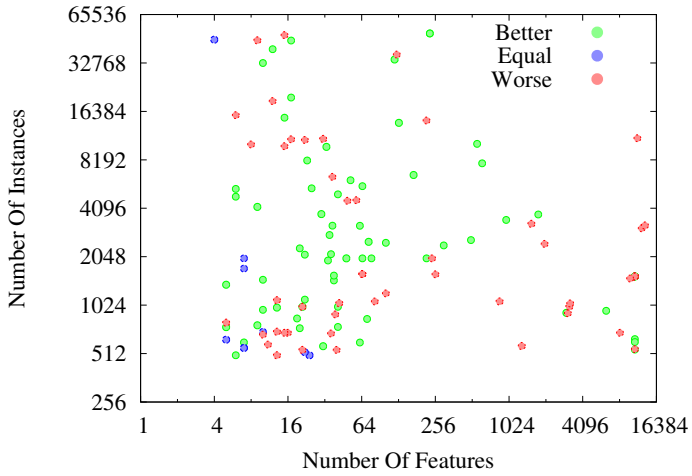
- 400 binary classification datasets from OpenML
- 12 algorithms from Weka
- Correlation-based Feature Subset Selection
- We added runs that not existed on OpenML
- Recorded Area Under the ROC curve
- Limitation: Hyperparameter Optimization

Universiteit Leiden

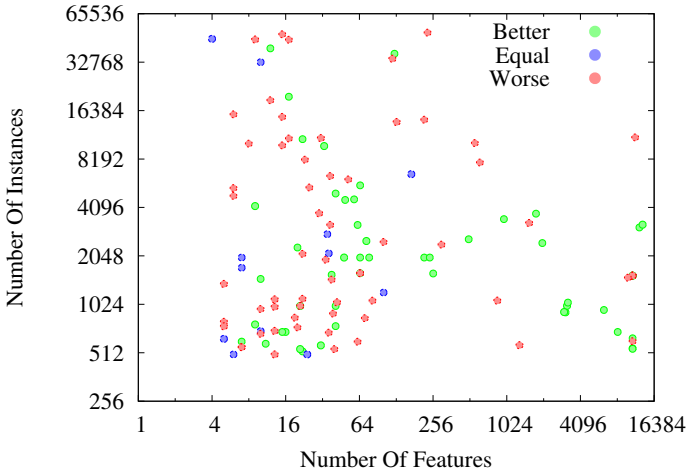
Effect of Feature Selection



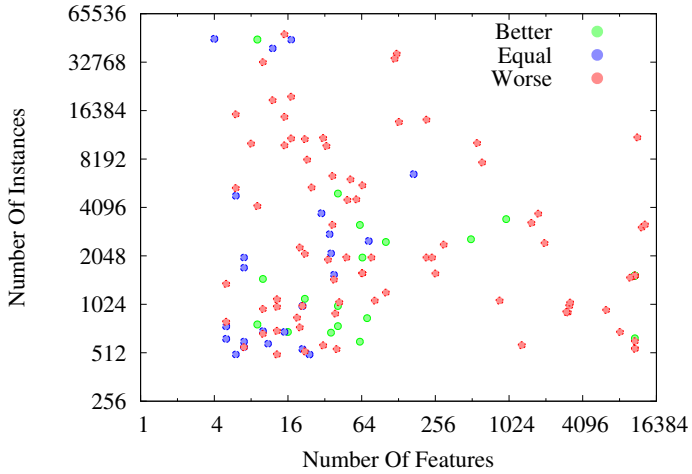
k-NN



Naive Bayes



J48



SMO

Universiteit Leiden

Linear vs. Non-linear

Reoccuring topic in literature, see [Strang et al., 2018] for several examples

	Linear	Non-linear
ease of tuning		
interpretability		
fit risk		
performance		

Universiteit Leiden

Linear vs. Non-linear

Reoccuring topic in literature, see [Strang et al., 2018] for several examples

	Linear	Non-linear
ease of tuning	+	-
interpretability		
fit risk		
performance		

Universiteit Leiden

Linear vs. Non-linear

Reoccuring topic in literature, see [Strang et al., 2018] for several examples

	Linear	Non-linear
ease of tuning	+	-
interpretability	+	-
fit risk		
performance		

Universiteit Leiden

Linear vs. Non-linear

Reoccuring topic in literature, see [Strang et al., 2018] for several examples

	Linear	Non-linear
ease of tuning	+	-
interpretability	+	-
fit risk	underfit	overfit
performance		

Universiteit Leiden

Linear vs. Non-linear

Reoccuring topic in literature, see [Strang et al., 2018] for several examples

	Linear	Non-linear
ease of tuning	+	-
interpretability	+	-
fit risk	underfit	overfit
performance	-	+

Universiteit Leiden

Linear vs. Non-linear

Reoccuring topic in literature, see [Strang et al., 2018] for several examples

	Linear	Non-linear
ease of tuning	+	-
interpretability	+	-
fit risk	underfit	overfit
performance	-	+
Tree	Decision Stump	Decision Tree

Universiteit Leiden

Linear vs. Non-linear

Reoccuring topic in literature, see [Strang et al., 2018] for several examples

	Linear	Non-linear
ease of tuning	+	-
interpretability	+	-
fit risk	underfit	overfit
performance	-	+
Tree	Decision Stump	Decision Tree
SVM	Linear Kernel	Gaussian Kernel

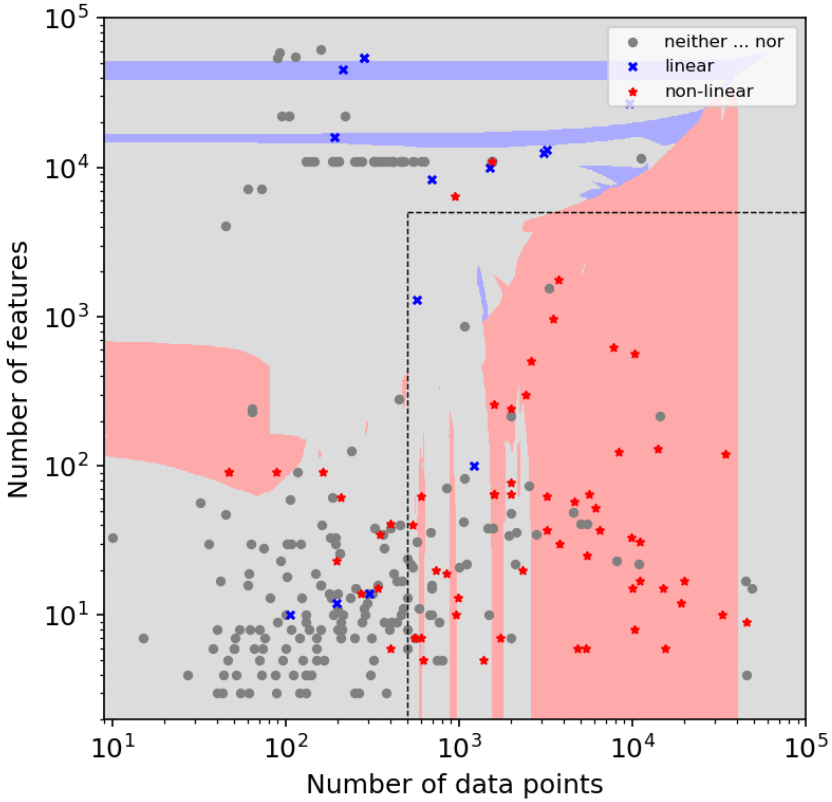
Universiteit Leiden

Linear vs. Non-linear

Reoccuring topic in literature, see [Strang et al., 2018] for several examples

	Linear	Non-linear
ease of tuning	+	-
interpretability	+	-
fit risk	underfit	overfit
performance	-	+
Tree	Decision Stump	Decision Tree
SVM	Linear Kernel	Gaussian Kernel
Neural Network	Perceptron	MLP

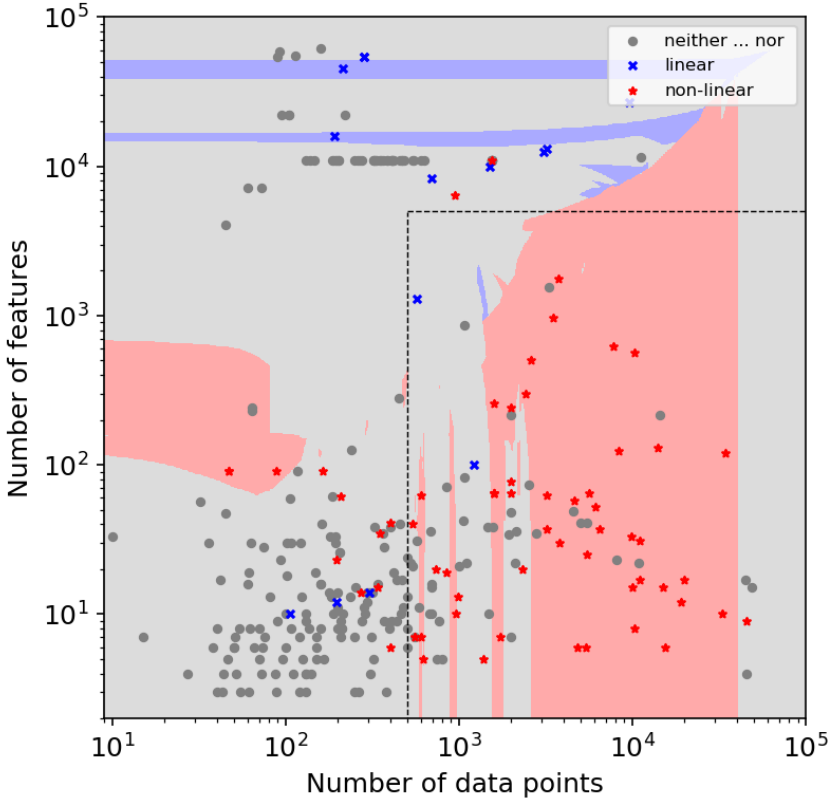
Linear vs. Non-Linear



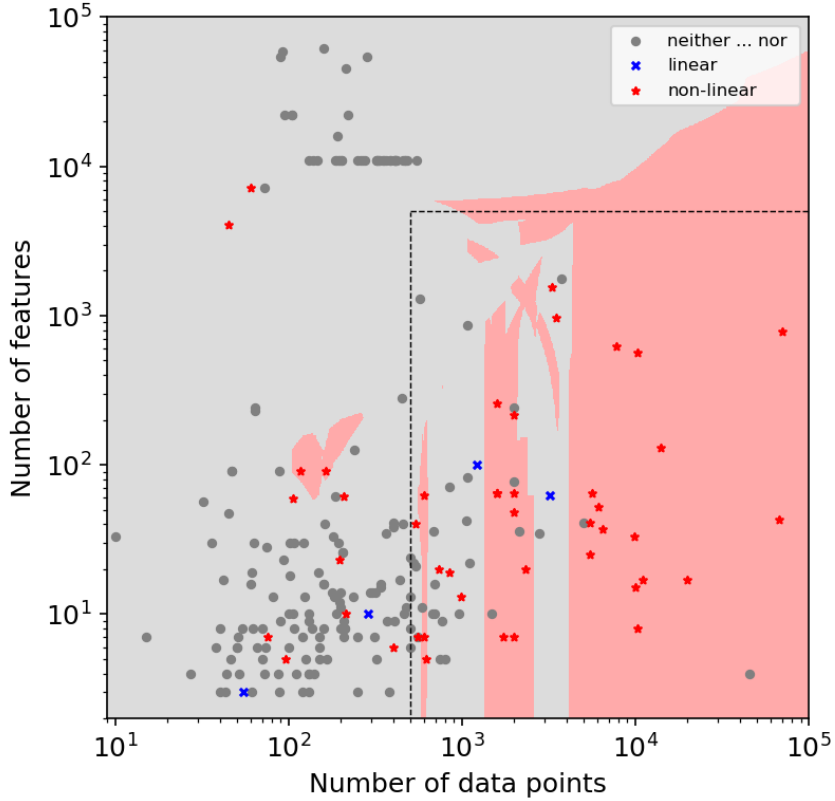
SVM

Universiteit Leiden

Linear vs. Non-Linear



SVM



Neural Networks

Universiteit Leiden

Effect of Feature Selection

Conclusions:

- Most results as expected:
  - Feature selection is often beneficial for the classifiers for which we expect it to be: k-NN and Naive Bayes
  - Non-linear classifiers exclusively better than linear classifier
- Whether or not to use feature selection can be learned (see paper)
- Low amount of datasets on which feature selection significantly effects performance potentially indicates data bias
- Realization: Limitation of OpenML100

Universiteit Leiden

References

B. Bischl, G. Casalicchio, M. Feurer, P. Gijsbers, F. Hutter, M. Lang, R. G. Mantovani, J. N. van Rijn, and J. Vanschoren. OpenML benchmarking suites. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, NIPS'21, 2021.

F. Hutter, H. H. Hoos, and K. Leyton-Brown. An efficient approach for assessing hyperparameter importance. In *Proc. of ICML 2014*, pages 754–762, 2014.

C. Moussa, J. N. van Rijn, T. Bäck, and V. Dunjko. Hyperparameter importance of quantum neural networks across small datasets. In *Discovery Science*. Springer International Publishing, 2022.

F. Pfisterer, J. N. van Rijn, P. Probst, A. C. Müller, and B. Bischl. Learning multiple defaults for machine learning algorithms, 2021.

M. J. Post, P. van der Putten, and J. N. van Rijn. Does Feature Selection Improve Classification? A Large Scale Experiment in OpenML. In *Advances in Intelligent Data Analysis XV*, pages 158–170. Springer, 2016.

A. Sharma, J. N. van Rijn, F. Hutter, and A. Müller. Hyperparameter importance for image classification by residual neural networks. In P. Kralj Novak, T. Šmuc, and S. Džeroski, editors, *Discovery Science*, pages 112–126. Springer International Publishing, 2019.

I. M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1(4):407–414, 1993.

B. Strang, P. van der Putten, J. N. van Rijn, and F. Hutter. Don't Rule Out Simple Models Prematurely: a Large Scale Benchmark Comparing Linear and Non-linear Classifiers in OpenML. In *Advances in Intelligent Data Analysis XVII*. Springer, 2018.

J. N. van Rijn and F. Hutter. Hyperparameter importance across datasets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2367–2376. ACM, 2018.