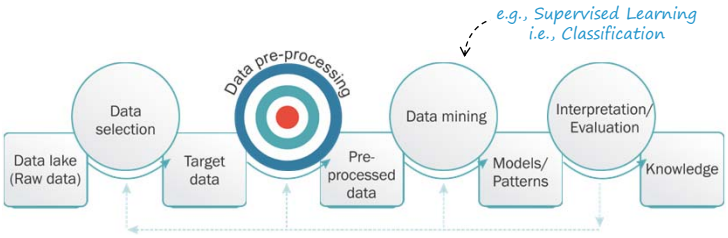
Learning the impact of data pre-processing for data analysis

Besim Bilalli | bbilalli@essi.upc.edu
Database Technologies and Information Management (DTIM)
Universitat Politècnica de Catalunya-BarcelonaTech (UPC)

Data analytics pipelines



- Data preparation is the act of manipulating raw data (which may come from disparate data sources) into a form that can readily and accurately be analyzed, e.g., for business purposes
- Data scientists spend considerable amount of time on data pre-processing (e.g., cleaning) before model training
  - It is often said that it consumes 80% of the analysis time

What does it include?

- Likely to include [1]:
  - Data discovery: identification of potentially relevant data sources (e.g., similar, join)
  - Data extraction: obtaining usable data form challenging and heterog. sources (e.g., deep web)
  - Data profiling: understanding basic properties of individual data sets (e.g., keys and relation.)
  - Format transformation: resolving inconsistencies in value representations
  - Source selection: choosing the data sets that are suitable for the problem at hand
  - Matching: identification of data sets that may contain the same type of information
  - Mapping: developing transformation programs that remove structural inconsistencies
  - Data repair: removal of constraint violations
  - Duplicate detection: identification of duplicate entries
  - Data integration: integration of data from various sources
  - …

Data pre-processing for ML [1]

Pre-processing affects the downstream ML application, but it is often unclear how

- It is typically performed for enabling the ML analysis, and not for improving it
  - So that, data conforms to the "syntax" of the models
  - Or, works that clean "semantic errors" too, but without taking into account how this impacts the subsequent workflow
- There are not many rigorous studies on how exactly pre-processing affects ML
  - **ML community** usually focuses on developing ML algorithms that are robust to some particular noise types of certain distributions
  - **Database (DB) community** has been mostly studying the problem of data cleaning alone without considering how data is consumed by downstream ML analytics (except some recent works like ActiveClean [SIGMOD '16], Learn2Clean [WWW '19], CleanML [ICDE '21], etc.)

Pre-processing impact on ML (1/3)

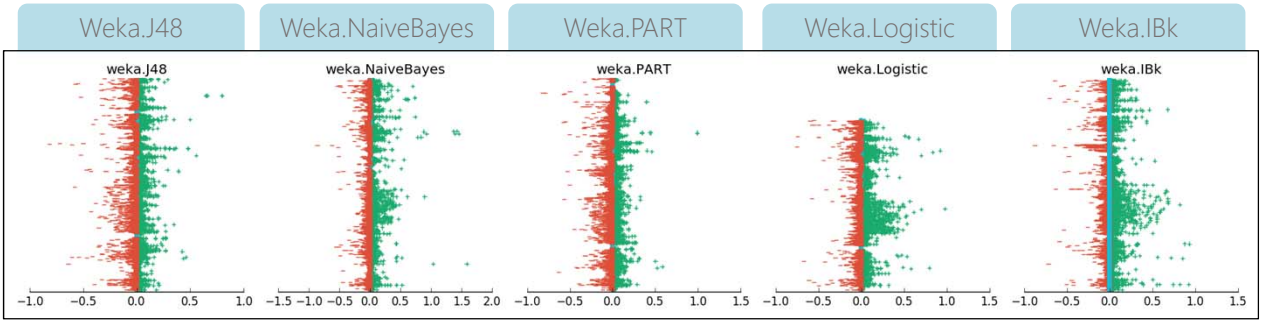
Pre-processing transformations

Operator	Technique	Attributes	Input Type	Output Type
Discretization	Supervised	Local	Continuous	Categorical
Discretization	Unsupervised	Local	Continuous	Categorical
Nominal to Binary	Supervised	Local	Categorical	Continuous
Nominal to Binary	Unsupervised	Local	Categorical	Continuous
Normalization	Unsupervised	Global	Continuous	Continuous
Standardization	Unsupervised	Global	Continuous	Continuous
Replace Missing Values	Unsupervised	Global	Continuous	Continuous
Replace Missing Values	Unsupervised	Global	Categorical	Categorical
PCA	Unsupervised	Global	Continuous	Continuous
...				

Transformations from Weka (filters) [default parametrizations]

Pre-processing impact on ML (2/3)

- Impact of different transformations on *classification accuracy* [1]

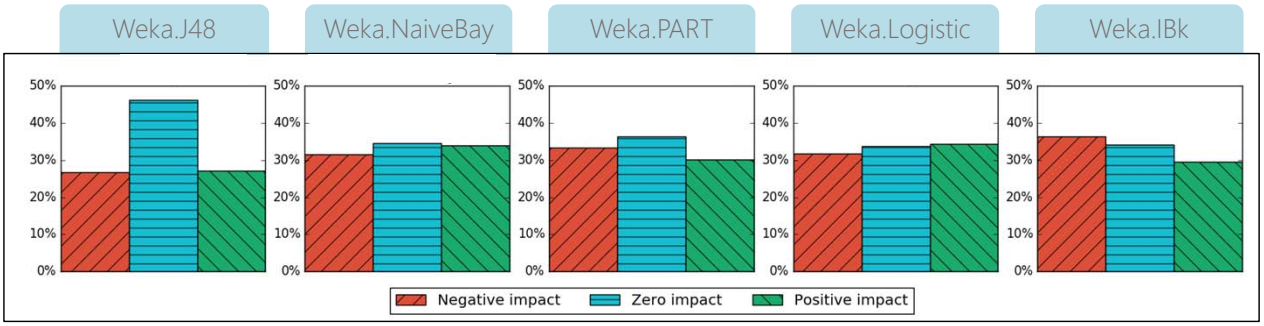


~500 datasets, ~25000 transformations in total

[1] Bilalli B., Abelló A. et al. Intelligent assistance for data pre-processing. CSI 2018

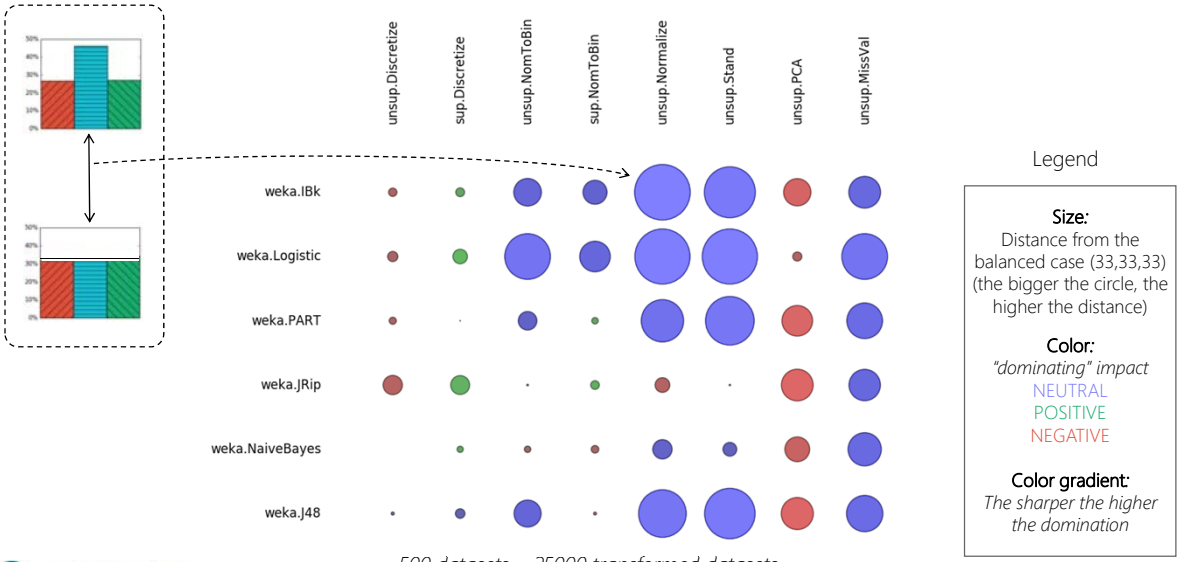
Pre-processing impact on ML (3/3)

- Distributions of impacts on *classification accuracy* [1]



~500 datasets, ~25000 transformations in total

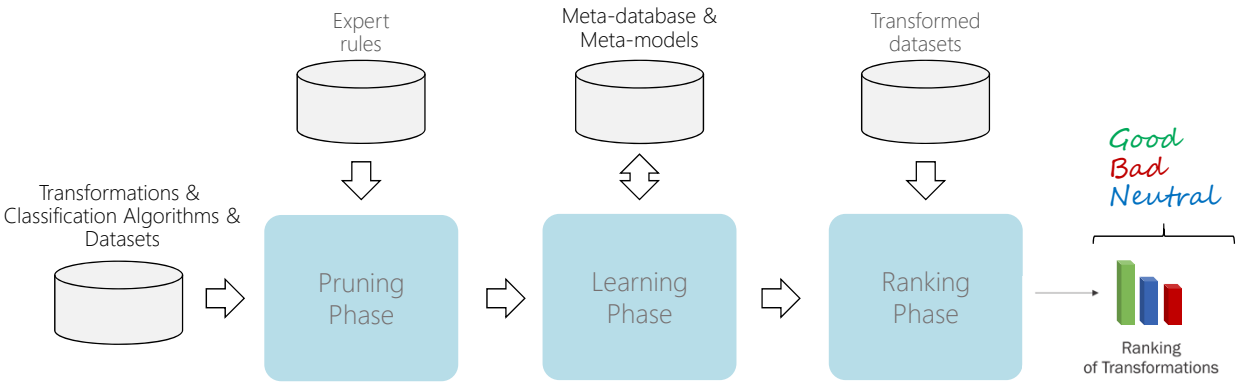
Impact *per transformation* type *per algorithm*



~500 datasets, ~25000 transformed datasets

(Meta) Learning the impact of transformations

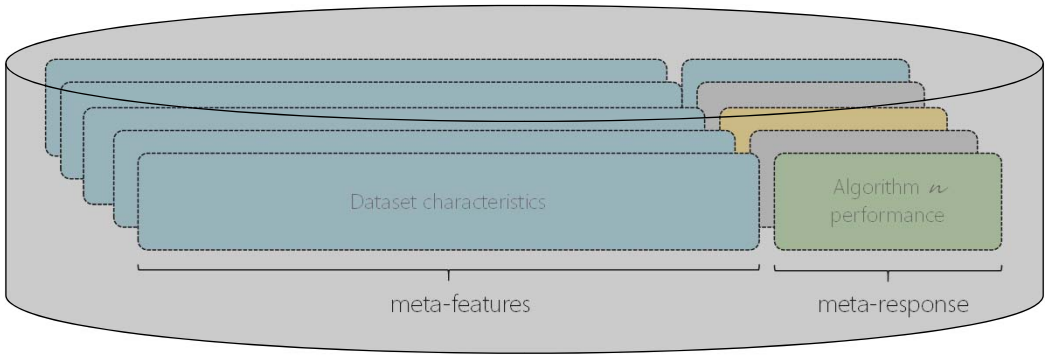
PRESISTANT Architecture [1]



(Meta) Learning the impact of transformations

Meta-database

Target metadata (meta-database)

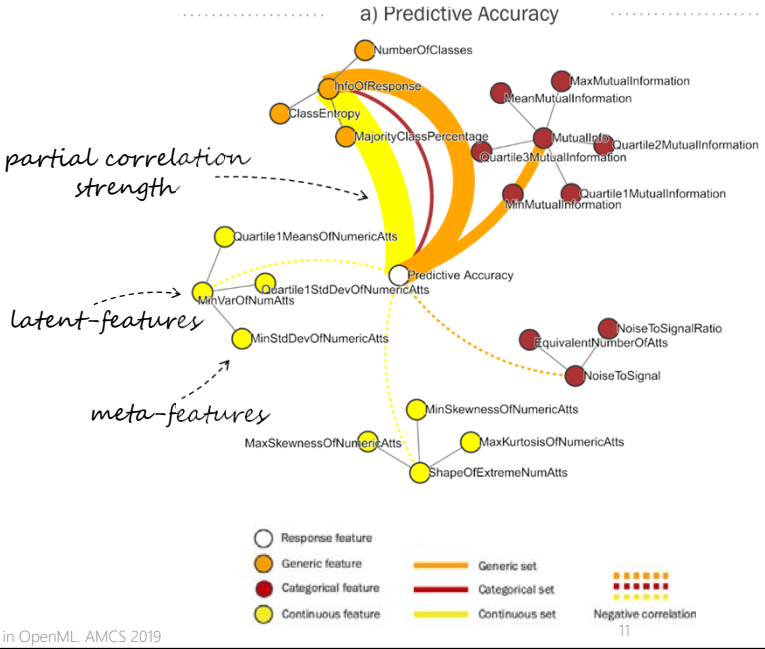


1. FEATURE EXTRACTION

Latent-Feature	Meta-feature	Corr.
Information of Categorical Attributes	MeanAttributeEntropy	0.853
	MinAttributeEntropy	0.925
	Q1AttributeEntropy	0.920
	Q2AttributeEntropy	0.878
Shape of Numeric Attributes	MeanKurtosisOfNumericAtts	0.933
	MeanSkewnessOfNumericAtts	0.874
	MinKurtosisOfNumericAtts	0.927
	Q1KurtosisOfNumericAtts	0.971
	Q1SkewnessOfNumericAtts	0.755
	Q2KurtosisOfNumericAtts	0.962
	Q2SkewnessOfNumericAtts	0.944
	Q3KurtosisOfNumericAtts	0.969
	Q3SkewnessOfNumericAtts	0.927
Variability of Numeric Attributes	MeanMeansOfNumericAtts	0.987
	MeanStdDevOfNumericAtts	-0.994
	MinMeansOfNumericAtts	0.999
	MaxStdDevOfNumericAtts	-0.999
Min. Variability of Num. Attributes	MinStdDevOfNumericAtts	0.821
	Q1MeansOfNumericAtts	0.719
	Q1StdDevOfNumericAtts	0.887
Modality of Categorical Attributes	MaxNominalAttDistinctValues	0.899
	MeanNominalAttDistinctValues	0.874
	StdvNominalAttDistinctValues	0.942
Number of Instances	NumberOfInstances	0.978
	MajorityClassSize	0.935
	MinorityClassSize	0.843

14 L.F.

2. FEATURE SELECTION



[1] Bilalli B., Abelló A. et al. On the predictive power of meta-features in OpenML. AMCS 2019

11

[1] Bilalli B., Abelló A. et al. PRESISTANT: Learning based assistant for data pre-processing. DKE 2019

PRESISTANT *results (1)*

How much do we *gain* from ~~top-K~~ ?

G_{π,T_d}

A permutation/ordering π of the gain values G_{Td} on the entire list of transformations in a given dataset d, results in the ordered list of gains
[we are interested in our *recommended* permutation, *best*, and *worst*]

- Since, the lower the position, the less valuable it is for the user; *use Discounted Cumulative Gain.*
- DCG progressively reduces the gain as the rank decreases.

$$DCG_{G_{\pi,T_d}} = \sum_{i=1}^{N} \frac{G_{\pi,T_d}[i]}{log_2(i+1)}$$

- How close we are to the best ranking?

$$nDCG_d = \frac{DCG_{G_{rec,T_d}} - DCG_{G_{worst,T_d}}}{DCG_{G_{best,T_d}} - DCG_{G_{worst,T_d}}}$$

Algorithm	$\overline{nDCG}$		#Datasets considered[a]
	All trans.	Top-1	
J48	0.73	0.79	503
Naive Bayes	0.78	0.84	504
PART	0.73	0.79	503
Logistic	0.64	0.67	447
IBk	0.77	0.84	500

[a]Number of datasets with at least 1 relevant (non-neutral) transformation

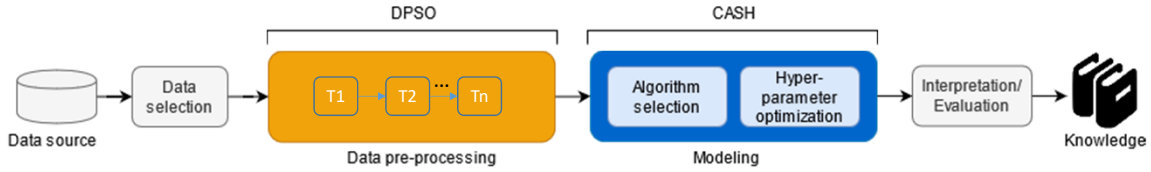
So far, only single transformations were applied to a dataset.

What about pre-processing pipelines (e.g., chain of transformations applied at once)?

13

Data pre-processing pipelines

Prototypes and pipelines



- Data pipeline prototypes (logical pipelines) are defined as fixed, ordered sequences of kinds of pre-processing transformations, where each kind of transformation can be instantiated by a specific set of *operators*, into executable pipelines (physical pip.)

List of transformations:

- *E* - Encoding;
- *N* - Normalization;
- *D* - Discretization;
- *I* - Imputation;
- *R* - Rebalancing;
- *F* - Feature Engineering.

Exhaustive set of data pipeline prototypes:

ID	Pipeline prototype	ID	Pipeline prototype
1	$I \rightarrow E \rightarrow N \rightarrow D \rightarrow F \rightarrow R$	13	$I \rightarrow E \rightarrow F \rightarrow N \rightarrow D \rightarrow R$
2	$I \rightarrow E \rightarrow N \rightarrow D \rightarrow R \rightarrow F$	14	$I \rightarrow E \rightarrow F \rightarrow N \rightarrow R \rightarrow D$
3	$I \rightarrow E \rightarrow N \rightarrow F \rightarrow D \rightarrow R$	15	$I \rightarrow E \rightarrow F \rightarrow D \rightarrow N \rightarrow R$
4	$I \rightarrow E \rightarrow N \rightarrow F \rightarrow R \rightarrow D$	16	$I \rightarrow E \rightarrow F \rightarrow D \rightarrow R \rightarrow N$
5	$I \rightarrow E \rightarrow N \rightarrow R \rightarrow D \rightarrow F$	17	$I \rightarrow E \rightarrow F \rightarrow R \rightarrow N \rightarrow D$
6	$I \rightarrow E \rightarrow N \rightarrow R \rightarrow F \rightarrow D$	18	$I \rightarrow E \rightarrow F \rightarrow R \rightarrow D \rightarrow N$
7	$I \rightarrow E \rightarrow D \rightarrow N \rightarrow F \rightarrow R$	19	$I \rightarrow E \rightarrow R \rightarrow N \rightarrow D \rightarrow F$
8	$I \rightarrow E \rightarrow D \rightarrow N \rightarrow R \rightarrow F$	20	$I \rightarrow E \rightarrow R \rightarrow N \rightarrow F \rightarrow D$
9	$I \rightarrow E \rightarrow D \rightarrow F \rightarrow N \rightarrow R$	21	$I \rightarrow E \rightarrow R \rightarrow D \rightarrow N \rightarrow F$
10	$I \rightarrow E \rightarrow D \rightarrow F \rightarrow R \rightarrow N$	23	$I \rightarrow E \rightarrow R \rightarrow D \rightarrow F \rightarrow N$
11	$I \rightarrow E \rightarrow D \rightarrow R \rightarrow N \rightarrow F$	23	$I \rightarrow E \rightarrow R \rightarrow F \rightarrow N \rightarrow D$
12	$I \rightarrow E \rightarrow D \rightarrow R \rightarrow F \rightarrow N$	24	$I \rightarrow E \rightarrow R \rightarrow F \rightarrow D \rightarrow N$

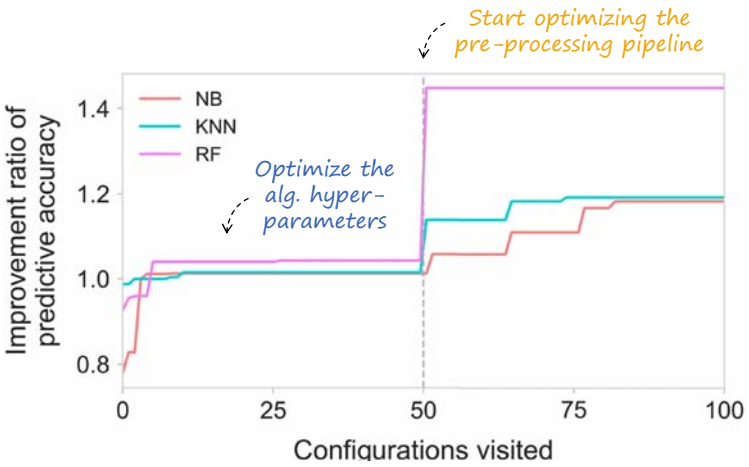
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

DTIM www.essi.upc.edu/dtim

14

Optimizing pre-processing pipelines

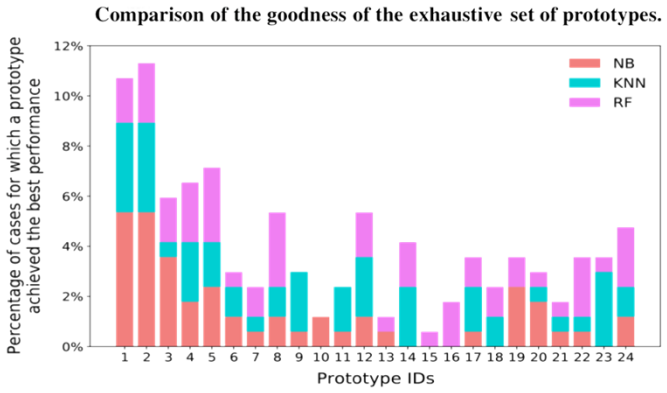
- AutoML experiment: `bank-marketing` dataset [1]



Every pipeline prototype is good for some dataset!

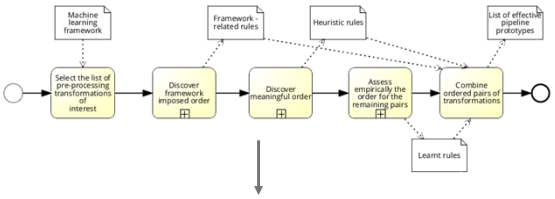
Exhaustive set of data pipeline prototypes:

ID	Pipeline prototype	ID	Pipeline prototype
1	$I \rightarrow E \rightarrow N \rightarrow D \rightarrow F \rightarrow R$	13	$I \rightarrow E \rightarrow F \rightarrow N \rightarrow D \rightarrow R$
2	$I \rightarrow E \rightarrow N \rightarrow D \rightarrow R \rightarrow F$	14	$I \rightarrow E \rightarrow F \rightarrow N \rightarrow R \rightarrow D$
3	$I \rightarrow E \rightarrow N \rightarrow F \rightarrow D \rightarrow R$	15	$I \rightarrow E \rightarrow F \rightarrow D \rightarrow N \rightarrow R$
4	$I \rightarrow E \rightarrow N \rightarrow F \rightarrow R \rightarrow D$	16	$I \rightarrow E \rightarrow F \rightarrow D \rightarrow R \rightarrow N$
5	$I \rightarrow E \rightarrow N \rightarrow R \rightarrow D \rightarrow F$	17	$I \rightarrow E \rightarrow F \rightarrow R \rightarrow N \rightarrow D$
6	$I \rightarrow E \rightarrow N \rightarrow R \rightarrow F \rightarrow D$	18	$I \rightarrow E \rightarrow F \rightarrow R \rightarrow D \rightarrow N$
7	$I \rightarrow E \rightarrow D \rightarrow N \rightarrow F \rightarrow R$	19	$I \rightarrow E \rightarrow R \rightarrow N \rightarrow D \rightarrow F$
8	$I \rightarrow E \rightarrow D \rightarrow N \rightarrow R \rightarrow F$	20	$I \rightarrow E \rightarrow R \rightarrow N \rightarrow F \rightarrow D$
9	$I \rightarrow E \rightarrow D \rightarrow F \rightarrow N \rightarrow R$	21	$I \rightarrow E \rightarrow R \rightarrow D \rightarrow N \rightarrow F$
10	$I \rightarrow E \rightarrow D \rightarrow F \rightarrow R \rightarrow N$	23	$I \rightarrow E \rightarrow R \rightarrow D \rightarrow F \rightarrow N$
11	$I \rightarrow E \rightarrow D \rightarrow R \rightarrow N \rightarrow F$	23	$I \rightarrow E \rightarrow R \rightarrow F \rightarrow N \rightarrow D$
12	$I \rightarrow E \rightarrow D \rightarrow R \rightarrow F \rightarrow N$	24	$I \rightarrow E \rightarrow R \rightarrow F \rightarrow D \rightarrow N$



Our approach

1. Generate effective pipeline prototypes (study transformations in pairs):
   - Framework + Heuristic + Learnt rules
2. Optimize the effective set
3. Select the best one

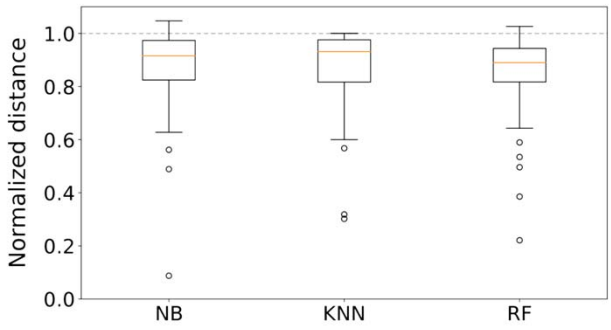


Effective set of data pipeline prototypes:

ID	Pipeline prototype
1	$I \rightarrow E \rightarrow N \rightarrow R \rightarrow F$
2	$I \rightarrow E \rightarrow N \rightarrow F \rightarrow R$
3	$I \rightarrow E \rightarrow R \rightarrow D \rightarrow F$
4	$I \rightarrow E \rightarrow D \rightarrow R \rightarrow F$
5	$I \rightarrow E \rightarrow D \rightarrow F \rightarrow R$

Optimizing pre-processing pipelines

Effective vs exhaustive pipelines



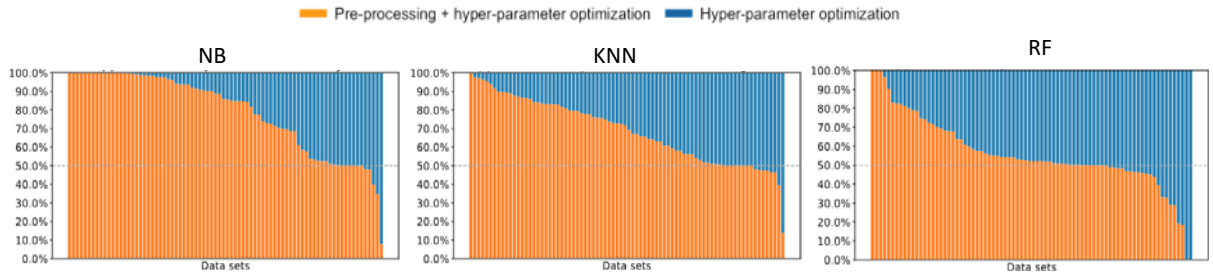
- With 24 times less time budget, our proposed pipeline prototypes were able to obtain results that were as good as 90% in the median of the 'optimal' ones found through an exhaustive search

Optimizing pre-processing pipelines

Pre-processing vs hyper-parameter optimization

- AutoML experiment: `80 datasets from OpenML`

What is more beneficial? Assigning the whole optimization time to hyper-parameter optimization or splitting it with pre-processing?



Conclusions and future work

- The results indicate that pre-processing can boost the performance of the ML algorithm; It must be considered as an integral part of the data analytics optimization process.

i. study the characteristics of the datasets that do not react well to data pre-processing (find out why)

ii. study more datasets and algorithms to further validate the approach

iii. perform meta-learning to recommend the most useful prototype, or at least study the relationship between characteristics of the datasets and the effective pipeline prototypes

Besim Bilalli
bbilalli@essi.upc.edu
www.essi.upc.edu/dtim/people

Questions?