Impact of filter feature selection on classification: an empirical study

Authors

Uchechukwu NJOKU, Alberto ABELLÓ

Besim BILALLI, Gianluca BONTEMPI (Université Libre de Bruxelles)

DOLAP 2022

29th March, 2022

Feature selection (FS)

The process of finding a subset of features in a dataset that are more relevant based on a **defined criteria**

Classification

- **Filter methods:** model independent

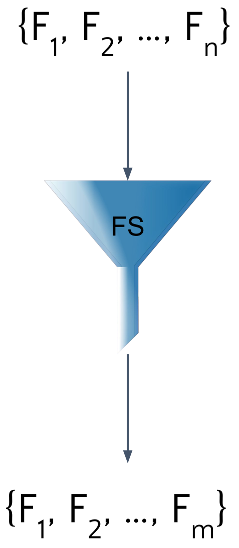
E.g: Gini, ReliefF, SPEC, CMIM, MRMR, JMI, RFS, CFS

- Wrapper methods: model dependent

E.g: Sequential forward(backward) selection

- Embedded methods: included in model

E.g: Tree based algorithms



$\{F_1, F_2, \ldots, F_n\}$

FS

$\{F_1, F_2, \ldots, F_m\}$

Why feature selection?

- Data understanding
- Data visualization
- Training time
- Storage

Contributions

Propose a systematic method to select representative datasets

Study and highlight the different effects of FS on the accuracy and training time of **binary and multiclass** classification models

Recommendations for choice of FS

Data selection

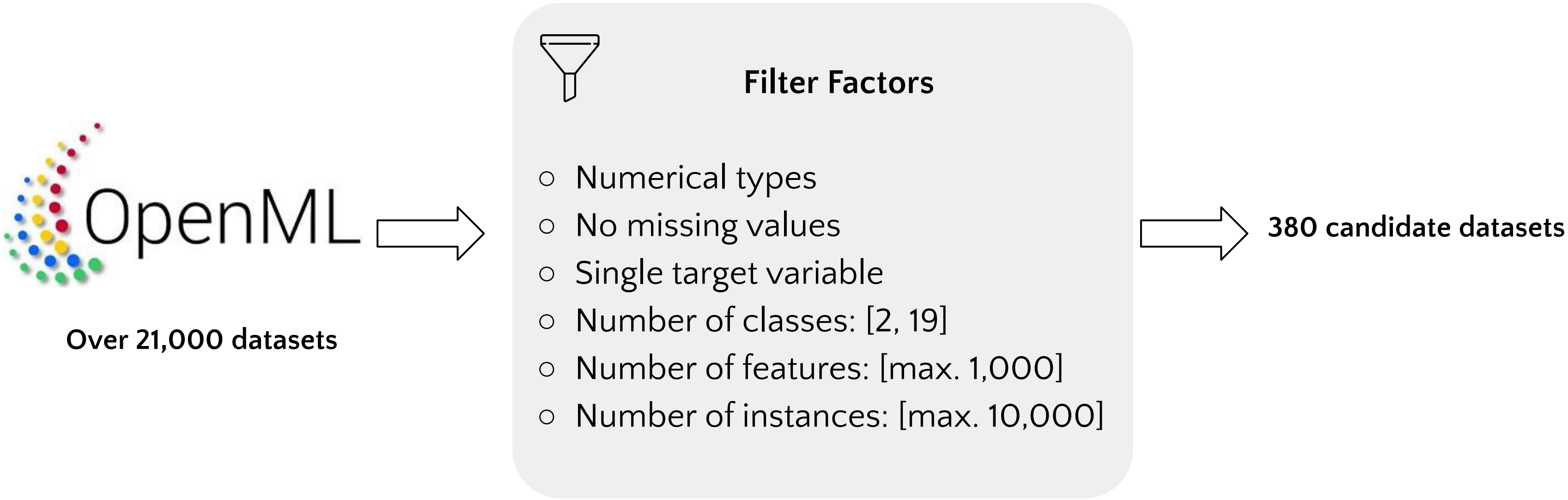
Why?

many more datasets than needed

Systematic approach steps:

1. Retrieve metadata: to gain insights into available data
2. Normalize metadata: for uniformity of values
3. Discretize metadata: to create clusters of similar datasets for selection

Data selection: OpenML Use Case



OpenML

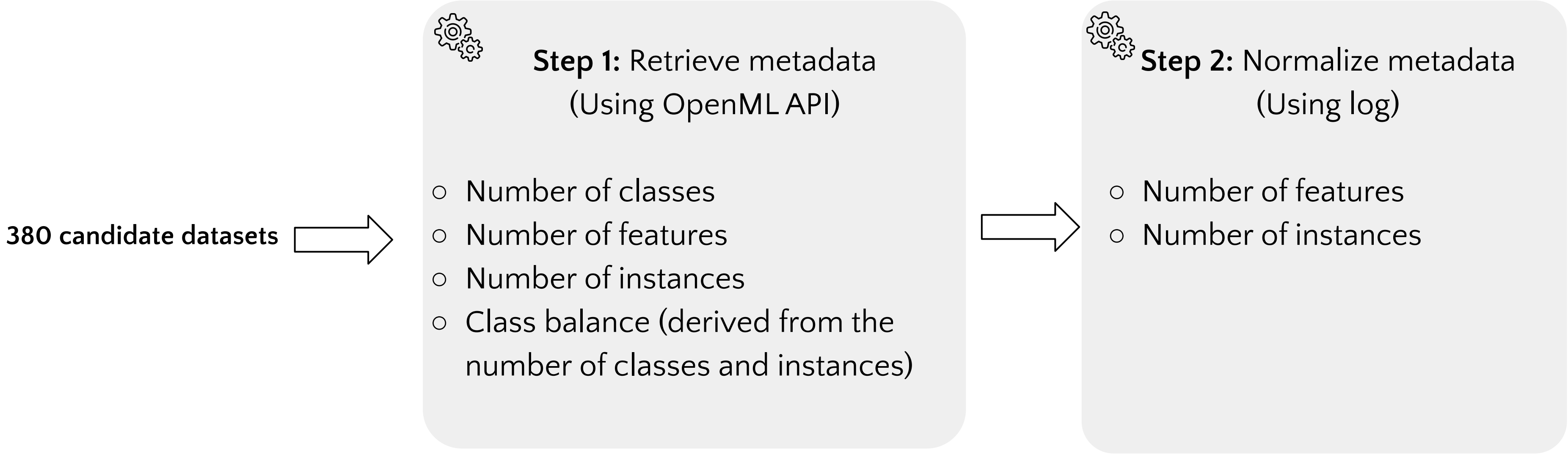
Over 21,000 datasets

Filter Factors

- Numerical types
- No missing values
- Single target variable
- Number of classes: [2, 19]
- Number of features: [max. 1,000]
- Number of instances: [max. 10,000]

380 candidate datasets

Data selection: OpenML Use Case



380 candidate datasets

Step 1: Retrieve metadata
(Using OpenML API)

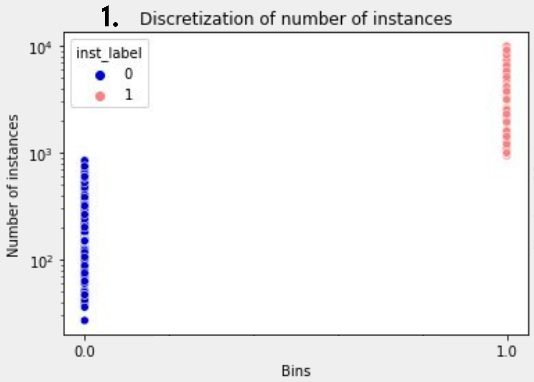
○ Number of classes
○ Number of features
○ Number of instances
○ Class balance (derived from the number of classes and instances)

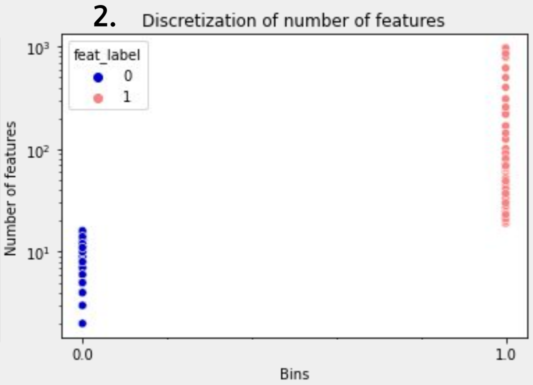
Step 2: Normalize metadata
(Using log)

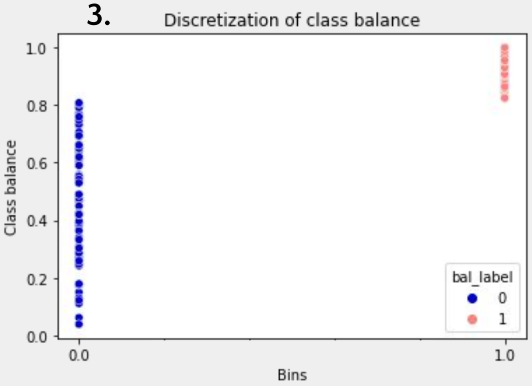
○ Number of features
○ Number of instances

Data selection: OpenML Use Case

Step 3: Discretize metadata (using agglomerative clustering)







4. Number of classes discretized into 0:binary (2 classes) and 1:multiclass (>2 classes)

- All datasets fall under one of 16 (2^4) possible factor combinations
- Two representatives chosen randomly

Datasets: OpenML Use Case

- **Number of features:**

0: [2, 16]	205 datasets
1: [19, 971]	175 datasets

- **Number of Instances**

0: [27, 846]	246 datasets
1: [937, 9.989]	134 datasets

- **Classes**

0 – Binary	262 datasets
1 – Multiclass	118 datasets

- **Class balance**

0: [0,0385, 0,8083]	84 datasets
1: [0,8232, 1,0]	296 datasets

CFIB

0000: 4

0001: 98

0010: 6

0011: 24

0100: 23

0101: 55

0110: 21

0111: 31

CFIB

1000: 4

1001: 46

1010: 18

1011: 5

1100: 5

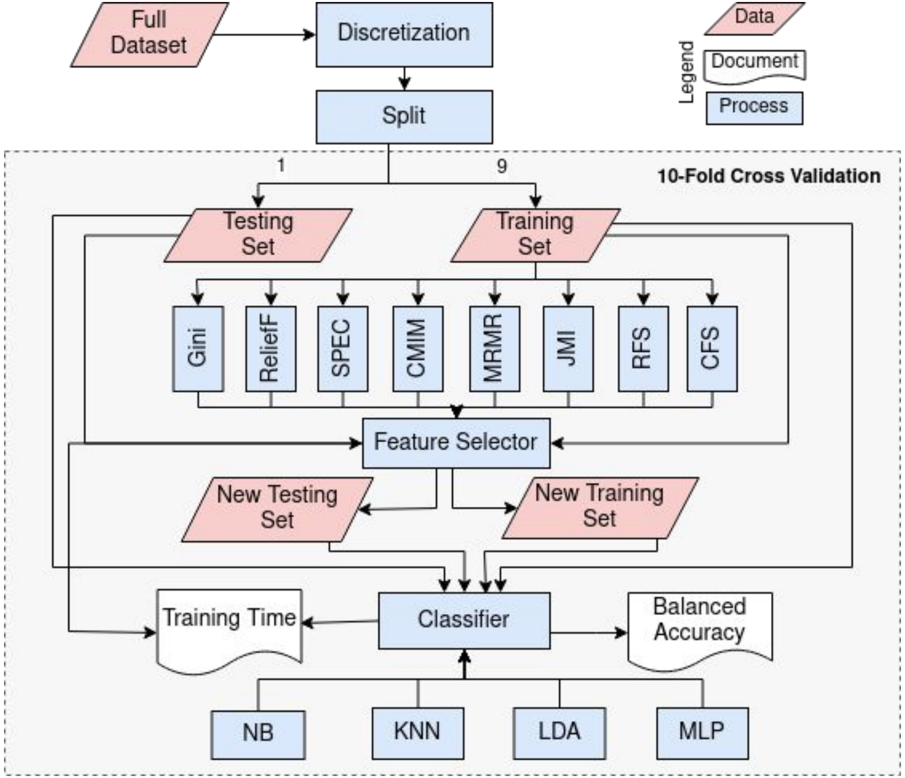
1101: 11

1110: 3

1111: 26

- **All datasets fall under one of 16 (2^4) possible factor combinations**
- **Two representatives chosen randomly**

Methodology



Work Load

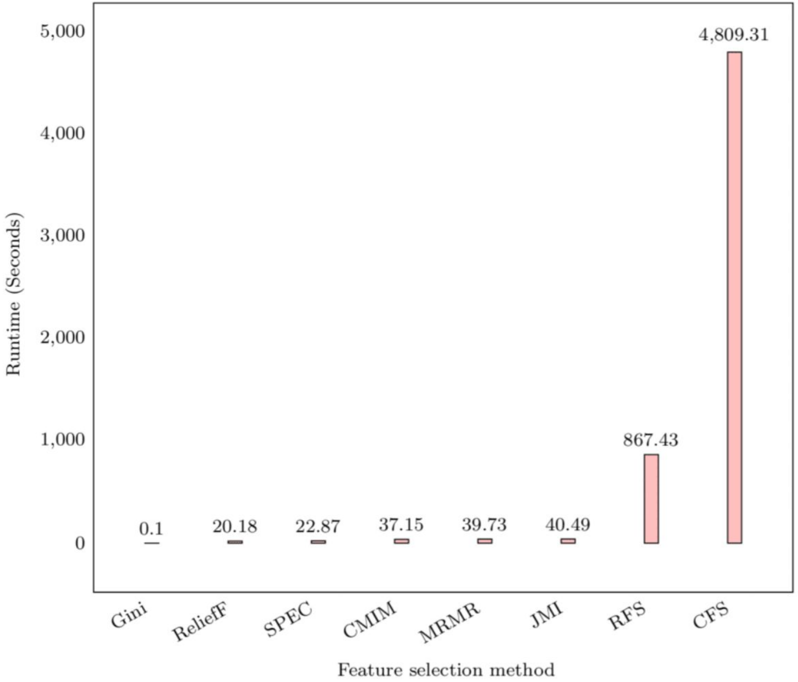
- 32 datasets
- 8 feature selection methods
- 4 classification algorithms
- 5 feature subset sizes [#features$^{(0.5,0.6,0.7,0.8,0.9)}$]

Metrics

- Feature selection runtime
- Classifier accuracy change
- Classifier runtime change

Results

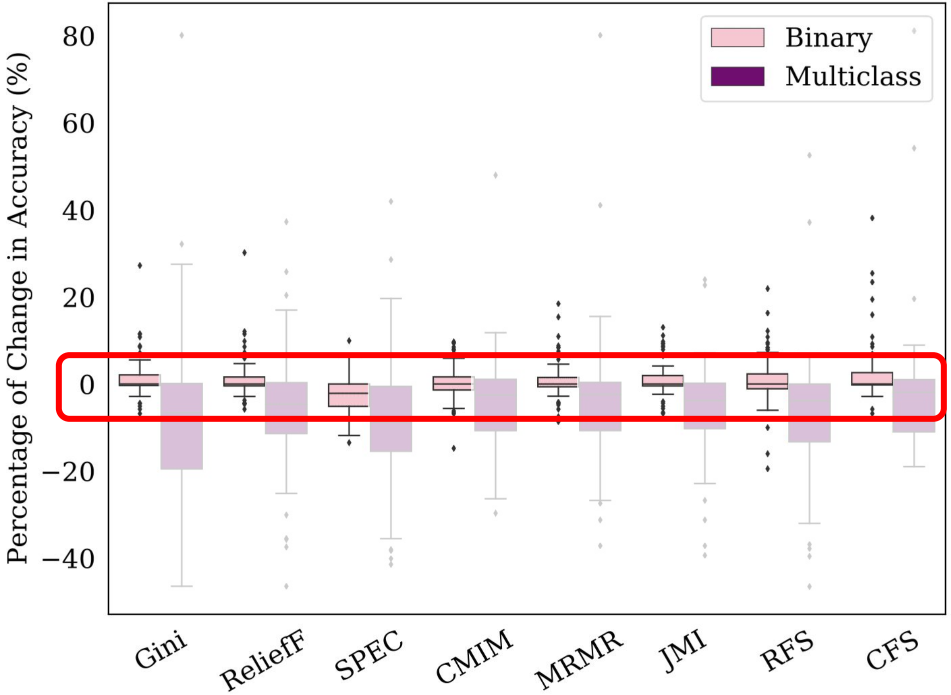
FS runtime



- Determined by various factors (number of instances, features, class, ...)
- Gini method (less than one second) is most time efficient
- CFS (over 80 minutes) is the least efficient method timewise due to the required computation of pairwise correlations between features
- Selected features can be re-used for model selection

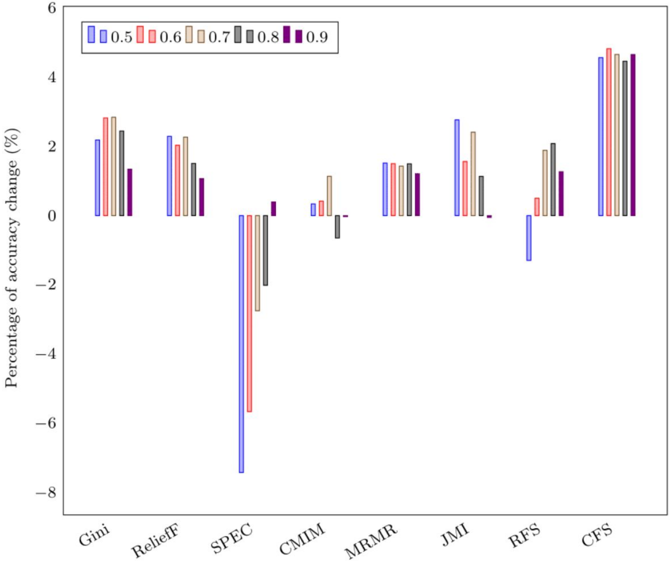
Classifier accuracy

The improvement in accuracy after feature selection is different for binary and multiclass classifications



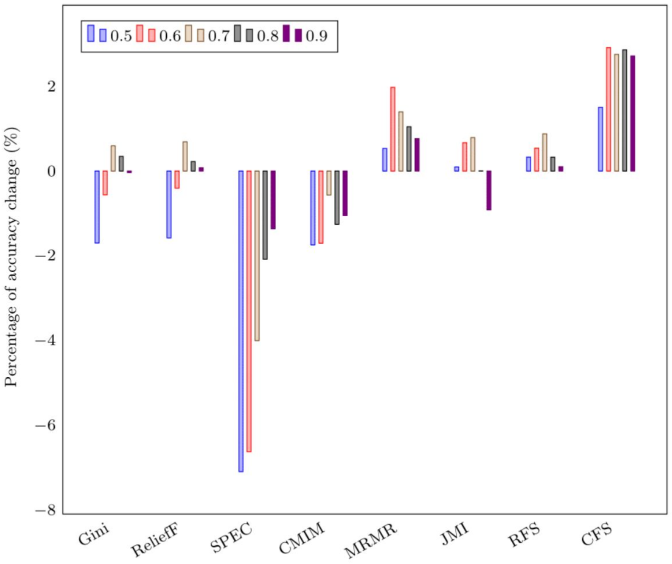
Binary classification

K-Nearest neighbour



K-Nearest Neighbour accuracy change for binary classification.

Naive Bayes

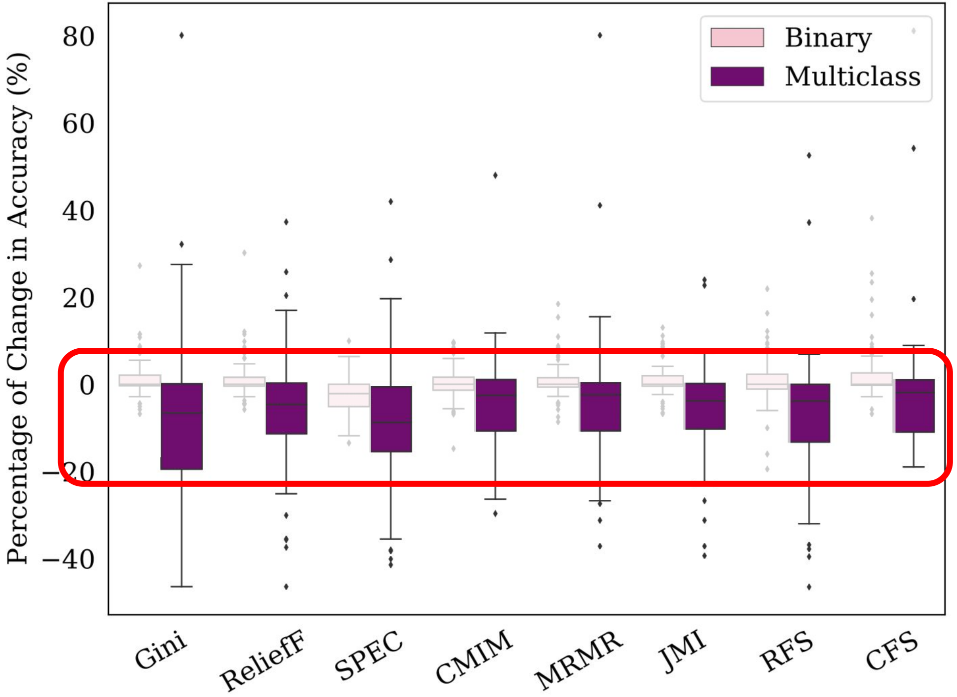


Naive Bayes accuracy change for binary classification.

Up to 5% accuracy improvement for binary classific

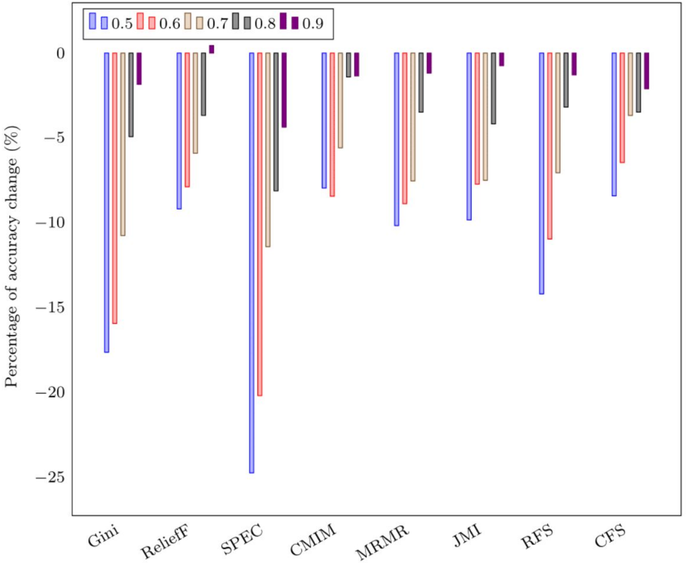
Classifier accuracy

The improvement in accuracy after feature selection is different for binary and multiclass classifications



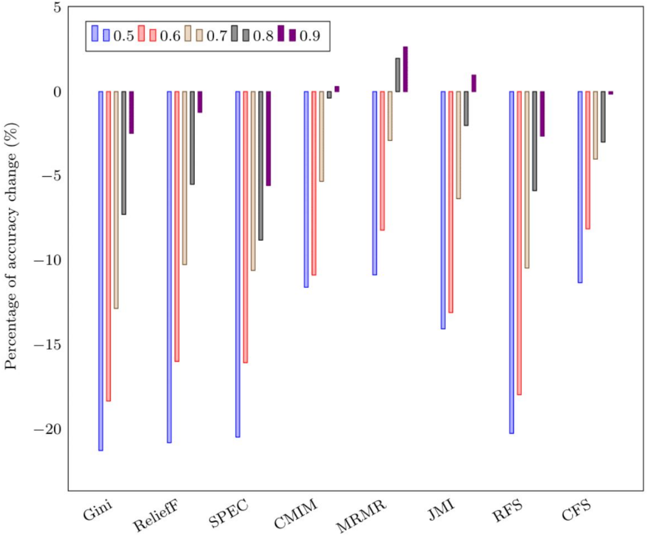
Multiclass classification

K-Nearest neighbour



K-Nearest Neighbour accuracy change for multi-class classification.

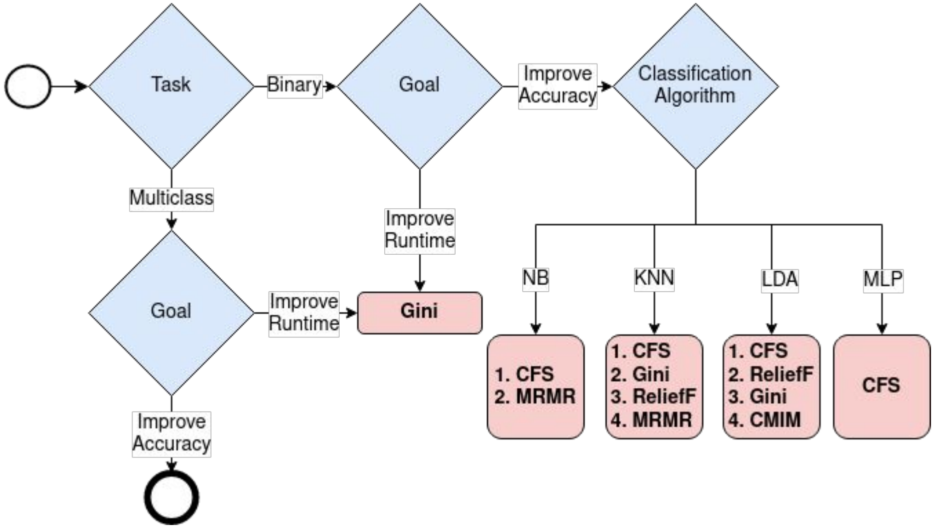
Naive Bayes



'aive Bayes accuracy change for multi-class classification.

Feature selection mostly leads to accuracy degradation for multiclass classification

Recommendation



Recommendation based on:

- ○ Goal
- ○ Task
- ○ Classification algorithm

No one size fits all FS method

Further work

- Extend work to larger datasets with more clusters of dataset factors and repositories
- Investigate the dependence of multiclass classification performance degradation after FS on the multiclass classification strategy
- Extend work to regression and clustering tasks
- Propose more efficient implementations of FS methods aiming for faster runtime

Thank You!

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

DTIM
Database Technologies and
Information Management

Contact

Email: unjoku@essi.upc.edu

Website: https://www.essi.upc.edu/dtim

Github: https://github.com/F-U-Njoku/filter-fs-impact-on-classification