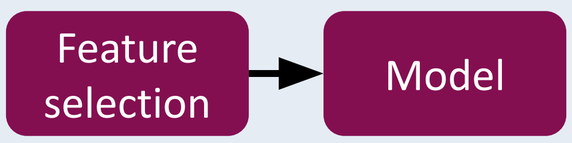
DISTRIBUTION AND REPLICATION FOR FEATURE SELECTION

01 Preliminaries

Feature selection is the process of detecting **relevant features** and discarding irrelevant or redundant ones from a dataset.

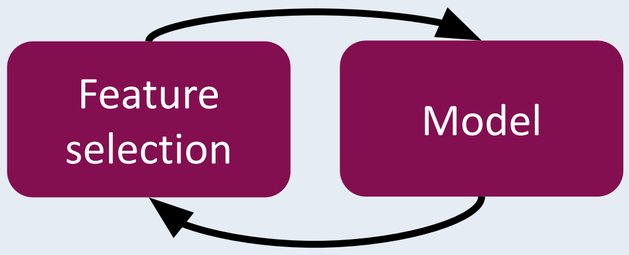
Filter



Relatively fast Execution

Non 'optimal' selection

Wrapper



High accuracy

Computationally intensive

Embedded

Moderate execution

Poor generalization

02 Motivation

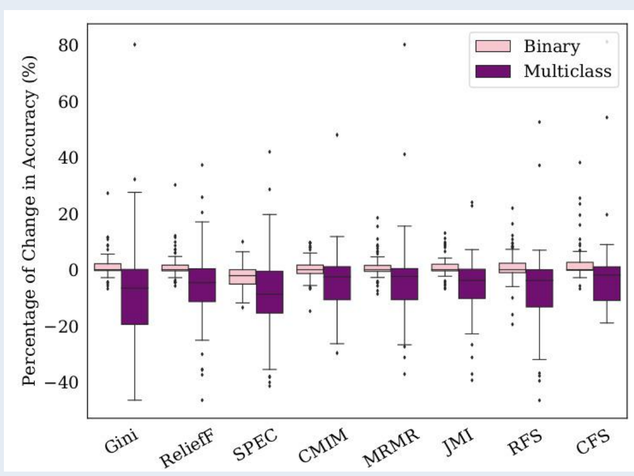
Characteristics
- High dimensionality
- Poor explainability
- Lengthy training

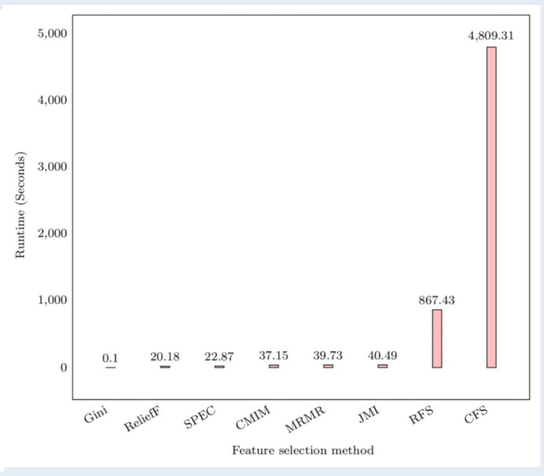
FS Challenges
- Scalability
- Interpretability
- Distributed data

Gaps
- Single criteria
- Focus on filter
- Insufficient tools
- Lack of explainability

03 Impact of filter methods

More impact on binary classification and fairly scalable





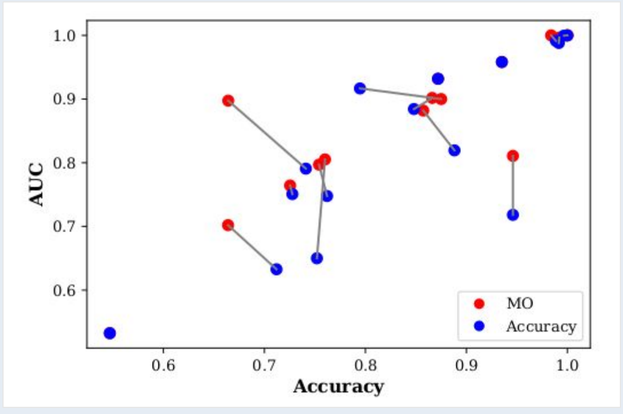
04 Wrapper methods – better?

Wrapper methods yield superior results

Method	NB	KNN	DT	SVM	Average Rank
SFS	10	9	4	9	1
JMI	4	4	8	6	2
Gini	5	2	6	5	3
CMIM	3	2	4	5	4
MRMR	3	3	4	3	5
Relief	2	1	4	3	6
SPEC	0	2	1	2	7

05 Current focus - Multi-Objective feature selection

A novel multi-criteria wrapper method

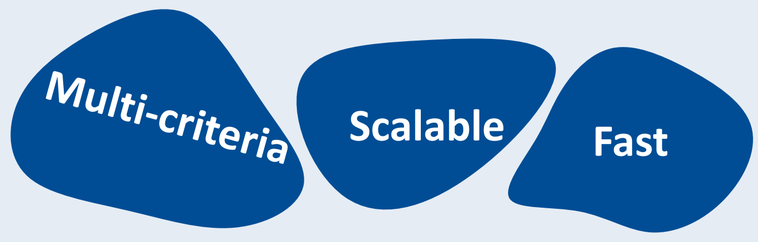


Scalarization approach

- Pareto approach
- Genetic algorithm
- More than two objectives
- Explainability

06 Outcome

A scalable feature selection tool with a suite of methods and multiple criteria for selection and post-selection evaluation to enable faster feature selection, interpretability, and explainability.



Further reading:

Njoku, U., Abelló, A., Bilalli, B., & Bontempi, G. Impact of Filter Feature Selection on Classification: An Empirical Study, DOLAP 2022

Authors: Contact unjoku@essi.upc.edu
Uchechukwu Fortune, NJOKU
Alberto ABELLO
Besim BILALLI
Gianluca BONTEMPI

Eleventh European Big Data Management & Analytics Summer School (eBISS 202) July 3 – 7, 2023, Barcelona, Spain