WIS Web Information Systems TU Delft

Data-Driven Empirical Analysis of Correlation-Based Feature Selection Techniques

Florena Buse **Andra Ionescu** Supervisor **Asterios Katsifodimos** Responsible Professor

Web Information Systems (WIS) Group, Delft University of Technology

TU Delft

I. Introduction

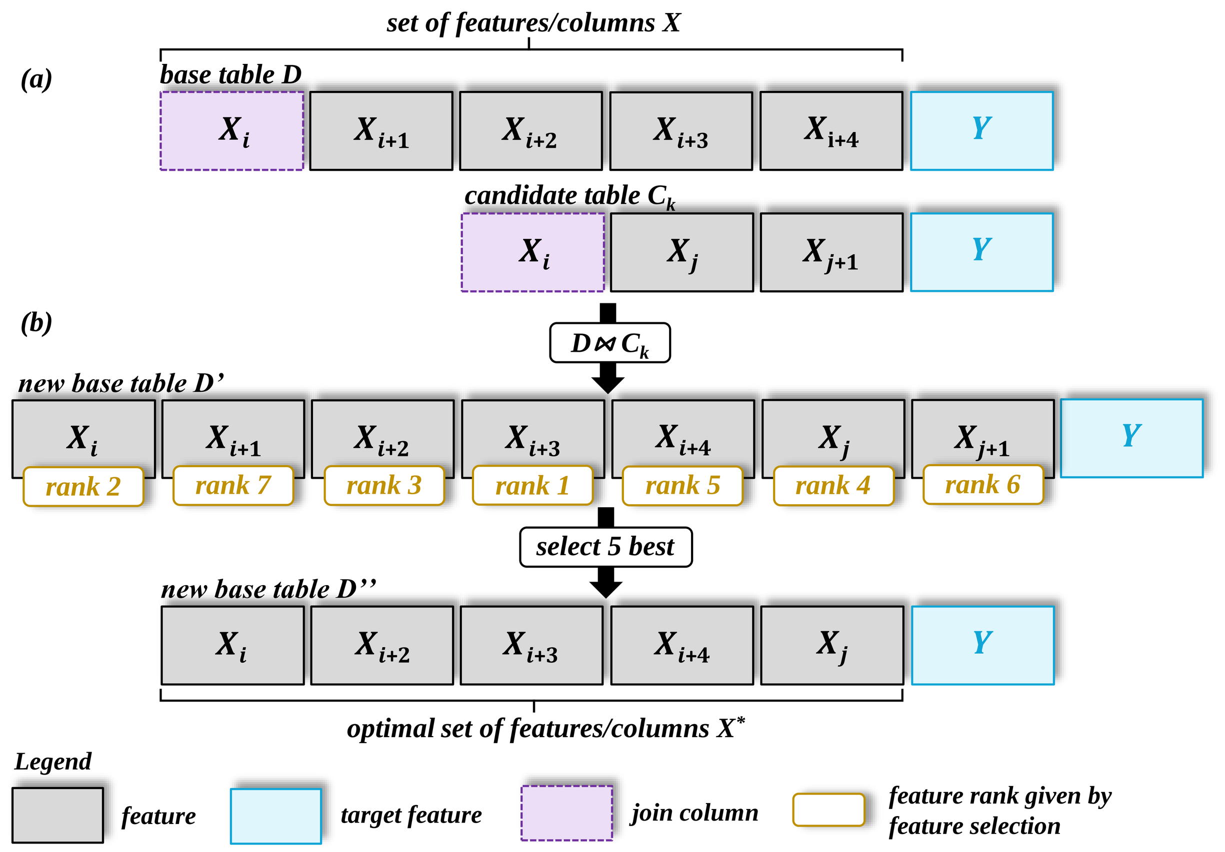

Figure 1. (a) Feature discovery vs (b) Feature selection.

II. Preliminaries

We propose four correlation measures that can be incorporated into feature selection to compute the relevancy of any feature X_i with regard to the target Y.

Pearson & Spearman

$$P(X_i, Y) = \frac{\sum_{j=1}^{N}(x_{ij} - \bar{X}_i)\cdot(y_j - \bar{Y})}{\sqrt{\sum_{j=1}^{N}(x_{ij} - \bar{X}_i)^2}\cdot\sqrt{\sum_{j=1}^{N}(y_j - \bar{Y})^2}}, \quad (1)$$

The Spearman correlation $S(X_i, Y)$ is computed in the same manner as $P(X_i, Y)$, except that X_i and Y are rank-transformed to values in $[1, N]$ [2].

Cramér's V

$$C(X_i, Y) = \sqrt{\frac{\chi^2}{N \cdot min(C_{X_i} - 1, C_Y - 1)}}, \quad (2)$$

where χ^2 is the chi-squared test. C_{X_i} and C_Y denote the number of categories of X_i and Y [3].

Symmetric Uncertainty (SU)

$$SU(X_i, Y) = \frac{2 \cdot IG(X_i, Y)}{H(X_i) + H(Y)}, \quad (3)$$

where $IG(X_i, Y)$ is the information gain. $H(X_i)$ and $H(Y)$ refer to Shannon's entropy [4].

References

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[2] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, pp. 16–28, 01 2014.

[3] H. Akoglu, "User's guide to correlation coefficients," *Turkish Journal of Emergency Medicine*, vol. 18, pp. 91–93, 09 2018.

[4] B. Singh, N. Kushwaha, and O. Vyas, "A feature subset selection technique for high dimensional data using symmetric uncertainty," *Journal of Data Analysis and Information Processing*, vol. 02, pp. 95–105, 01 2014.

III. Research question & sub-questions

(RQ) How do correlation-based feature selection techniques, in particular Pearson, Spearman, Cramér's V, SU, influence the performance of Decision trees, Linear ML algorithms and Support vector machines?

(SQ1) What is the best correlation technique to be used considering the dimensionality and feature type(s) of the data?

(SQ2) How much does the choice of ML algorithm influence the performance of correlation-based feature selection techniques?

IV. Methodology

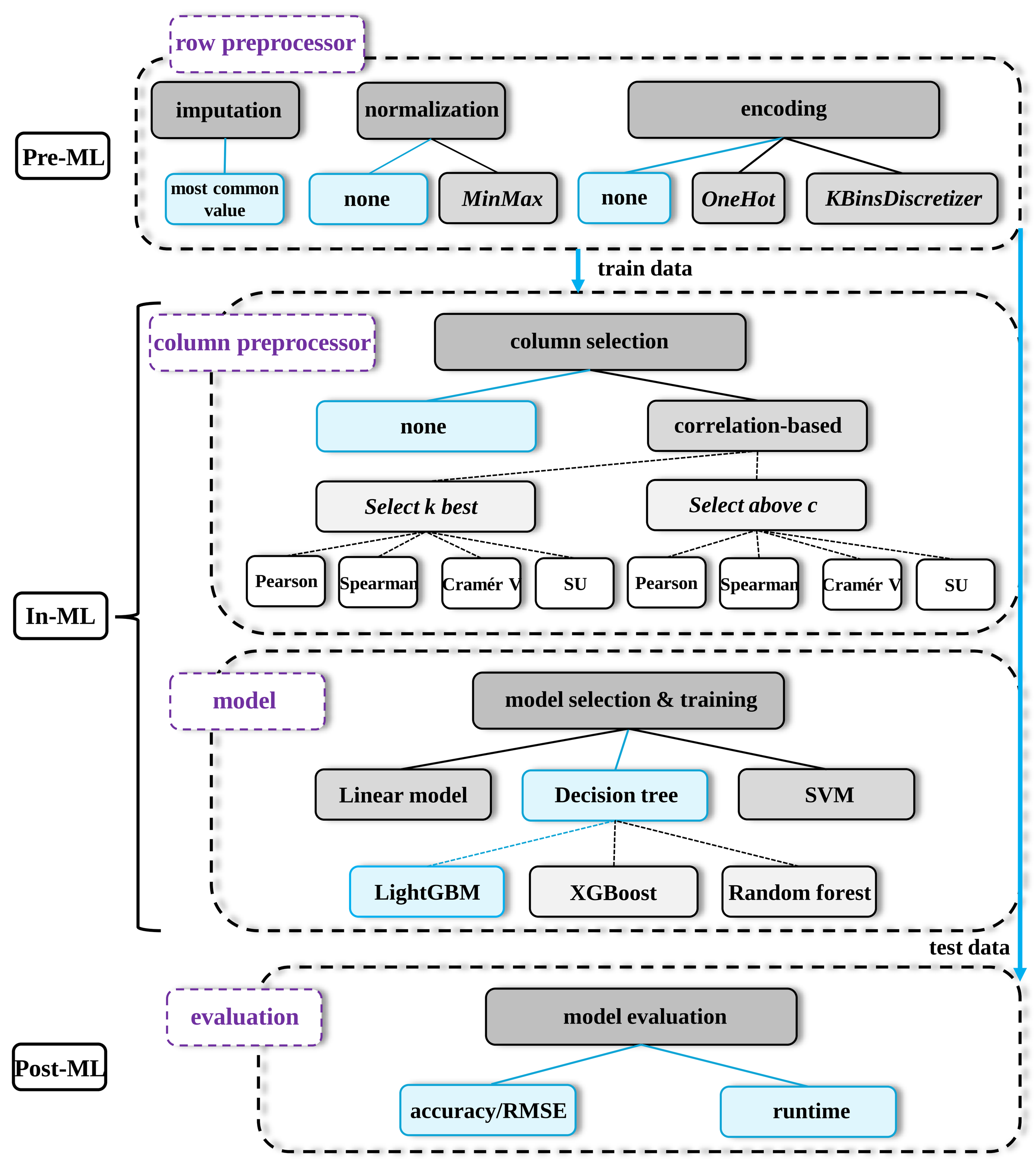

Figure 2. ML pipeline for the empirical evaluation. Blue hyperparameters form the *baseline* configuration.

V. Empirical results

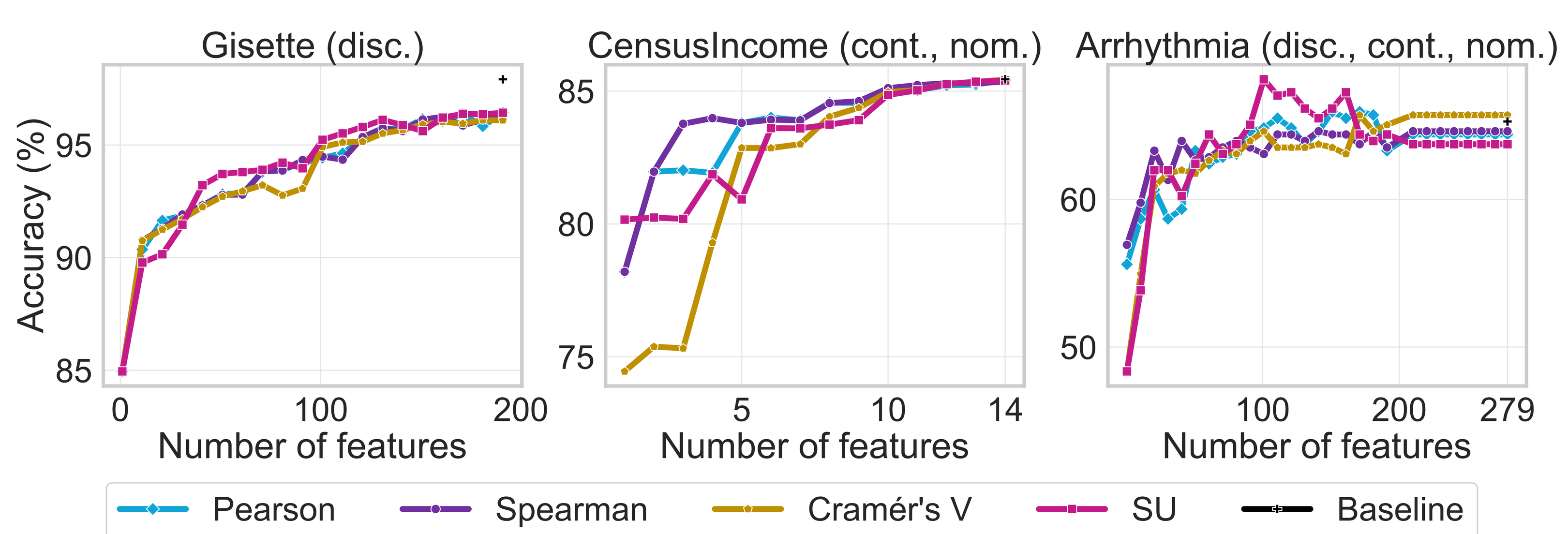

Figure 3. **Effectiveness** of the correlation-based feature selection techniques averaged over all ML algorithms.

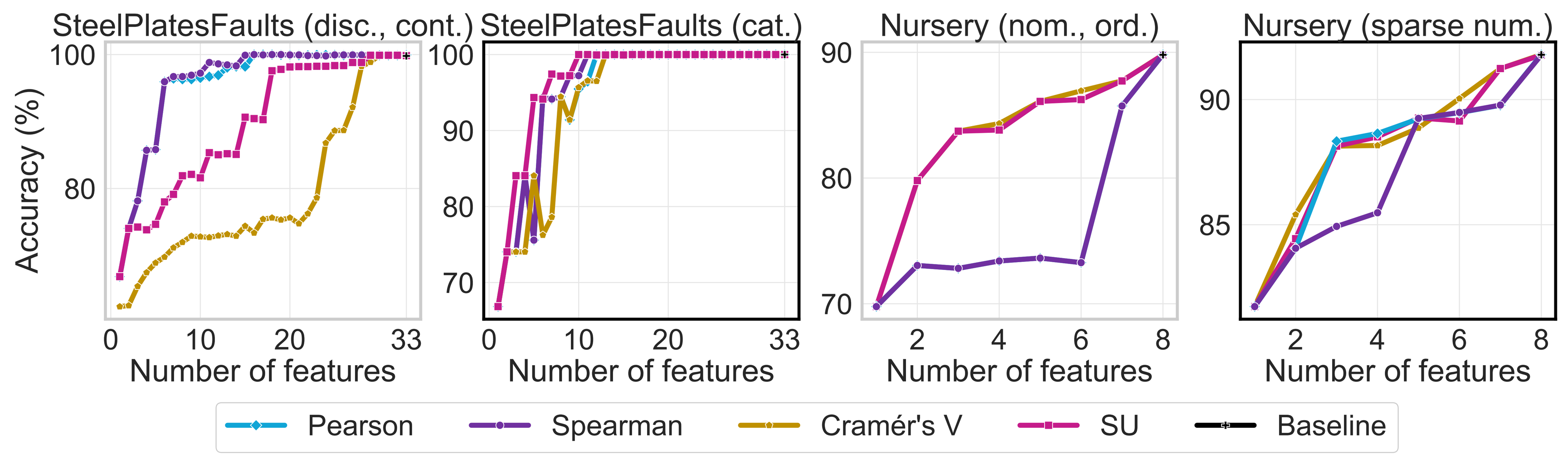

Figure 4. **Effectiveness** of the methods on the original datasets (grey outline) and the encoded datasets (black outline).

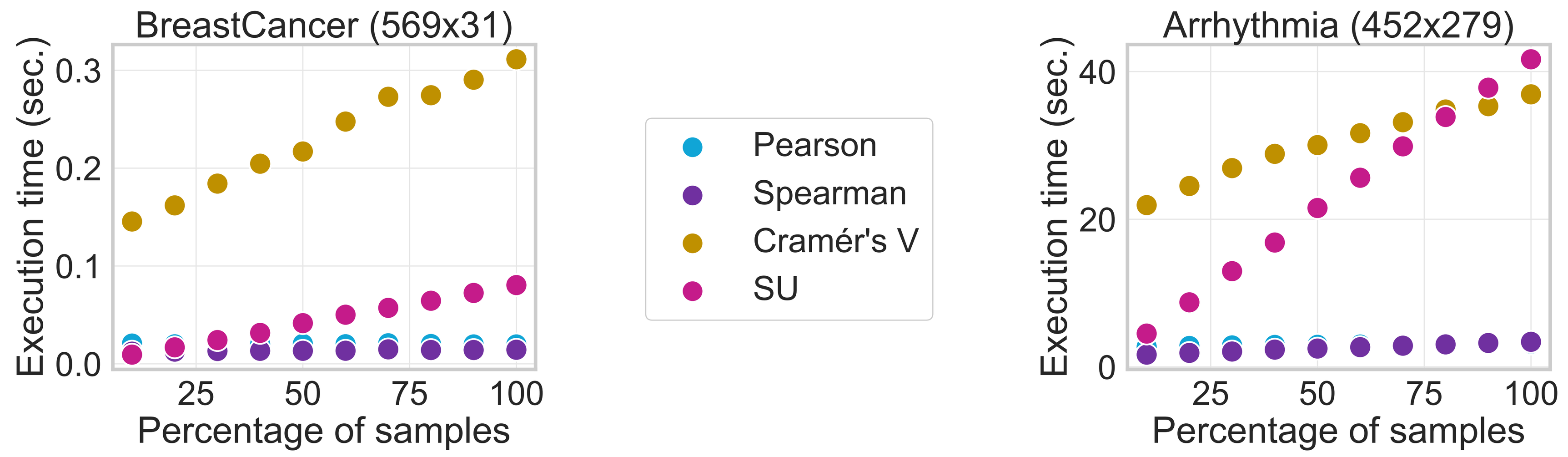

Figure 5. **Efficiency** of the feature selection stage computed for an increasing % of the samples.

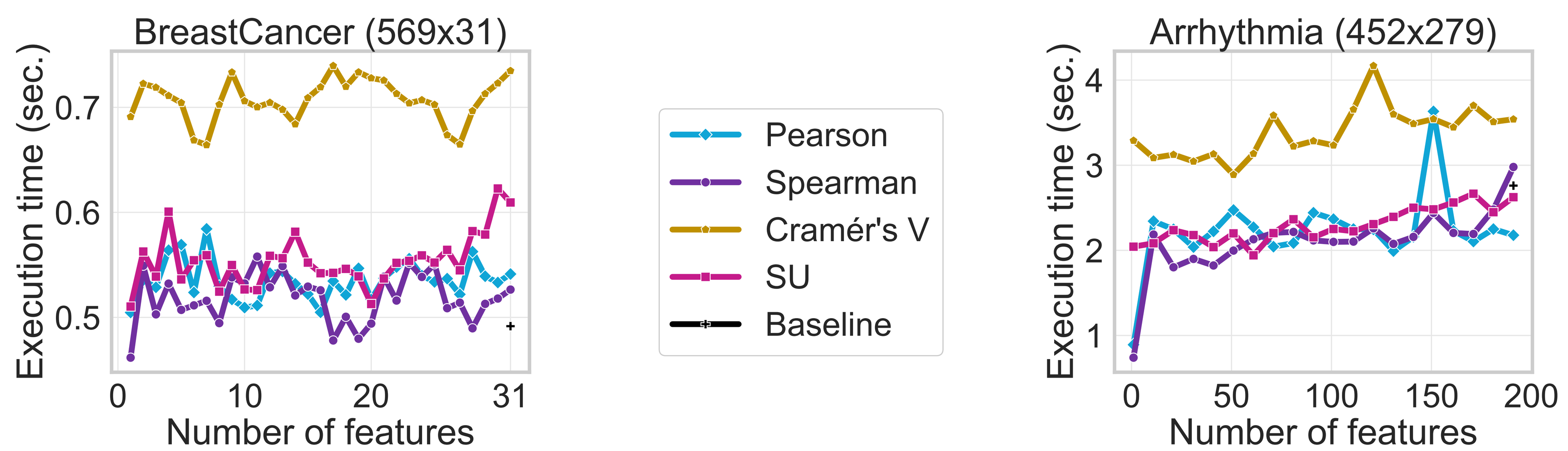

Figure 6. **Efficiency** of the **In-ML** stage of the pipeline computed for an increasing # of features.

VI. Conclusions

1. (SQ1) (i) **Effectiveness** of methods is highly tied to the type(s) of features. Theoretical assumptions do not hold in practice and we devise new ones in Table 1.
   (ii) **Efficiency** of feature selection is dependent on the dimensionality of the data.
2. (SQ2) (i) No correlation measure has been identified to exhibit superior **effectiveness** exclusively for a particular algorithm.
   (ii) **Efficiency** of the ML system decreases with feature selection, but it is worth the trade-off to obtain increased effectiveness.

Table 1. Feature types suitable to the correlation measures. Purple represents the types assumed **in theory**. Red denotes the types that were found to work **in practice**.

	Numerical		Categorical	
	Discrete	Continuous	Nominal	Ordinal
Pearson	✓	✓ ✓	✓	
Spearman	✓	✓ ✓	✓	✓
Cramér's V	✓		✓ ✓	✓
Symmetric Uncertainty	✓	✓	✓ ✓	✓

i.buse@student.tudelft.nl a.ionescu-3@tudelft.nl a.katsifodimos@tudelft.nl