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

Florena Buse

Supervisors: Andra Ionescu and Asterios Katsifodimos

Faculty of Electrical Engineering, Mathematics & Computer Science, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 26, 2023

Name of the student: Florena Buse
Final project course: CSE3000 Research Project
Thesis committee: Andra Ionescu, Asterios Katsifodimos and Elvin Isufi

Abstract

Thus far the democratization of machine learning, which resulted in the field of AutoML, has focused on the automation of model selection and hyperparameter optimization. Nevertheless, the need for high-quality databases to increase performance has sparked interest in correlation-based feature selection, a simple and fast, yet effective approach to removing noise and redundancy in relational data. However, little to no attention has been paid to *what* correlation metric to choose in order to maximize the performance of ML systems. Our research investigates the effectiveness and efficiency of four widely-known correlation measures, in particular Pearson, Spearman, Cramér's V, Symmetric Uncertainty, in a manner that simulates an AutoML-like setting. We show that the exact theoretical assumptions of the methods do not always hold in practice, as well as shed light on the main aspects that need to be considered when integrating correlation-based feature selection in ML systems. Notably, the results indicate that the performance obtained by correlation-based methods is highly tied to the types and number of features present in the underlying database rather than the choice of ML algorithm. We devise promising conclusions that can further serve the advancement of AutoML systems by making feature selection fully automatic and computationally tractable.

1 Introduction

Lately, we have witnessed an ever-growing demand for the democratization of machine learning (ML): making customized state-of-the-art ML algorithms available to everyone [1, 2]. In response to this demand, the field of automatic ML (AutoML) aims to automate the process of designing and optimizing ML pipelines: from data engineering to model selection and hyperparameter tuning [3]. However, *how do AutoML systems identify the most relevant and high-performing features that enhance the algorithms?*

Feature engineering. In the context of ML models working with relational databases, performance highly relies on both the quantity and quality of the data [4]. Thus, any effective AutoML system incorporates two essential feature engineering tasks: (i) feature discovery, so that the potential of all available data is exploited, and (ii) feature selection, so that data redundancy and noise are minimized.

Feature discovery aims to search for candidate tables that can be joined with the base table on a common column (i.e. feature), as well as contain other columns that are correlated with the target [5]. As feature discovery augments the data with new features from the same domain, it can result in a database characterized by noise, irrelevancy and high dimensionality [6]. To further enhance the efficiency and effectiveness of AutoML systems, it is crucial that feature selection follows feature discovery in the ML pipeline, process that is depicted in Fig. 1.

Feature selection assesses the relevance of all columns in the base table, ranks them, and then returns an optimal subset of columns [7]. The retained column subset is able to efficiently describe the input data while removing noise and redundancy. The main objectives of feature selection are manifold: (i) keeping only the features useful in discriminating the target classes, (ii) defying the curse of dimensionality to improve the performance of ML algorithms, (iii) reducing computation and storage costs [8, 9], and (iv) improving the system's generalization to new data instances. While there are many types of feature selection, correlation-based selection holds particular importance due to its simplicity and effectiveness.

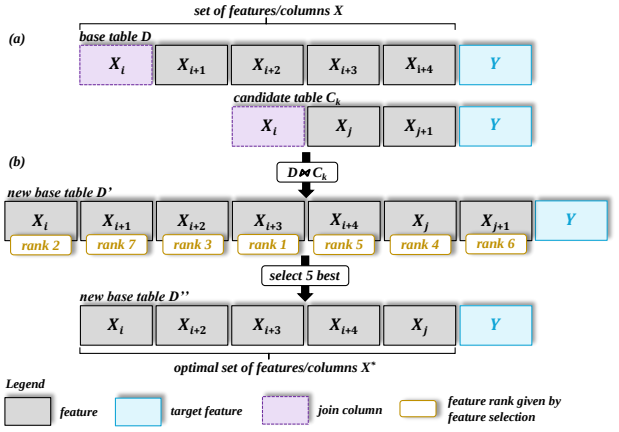

Figure 1: Visual representation of (a) feature discovery and (b) feature selection.

Correlation-based feature selection is a field that has already been extensively discussed and employed in practice, with many correlation metrics being proposed or devised [6, 8, 10, 11]. However, different correlation measures lead to different variants of the optimal feature subset, which can, in turn, positively or negatively impact the performance of the ML algorithm. To the best of our knowledge, no existing publication has evaluated which correlation measure, more specifically Pearson, Spearman, Cramér's V, Symmetric Uncertainty (SU), should be used for feature selection in order to maximize the performance of ML systems. To target this niche, our study investigates the following research question:

> *How do correlation-based feature selection techniques, in particular Pearson, Spearman, Cramér's V, Symmetric Uncertainty, influence the performance of Decision trees, Linear ML algorithms and Support vector machines?*

With a vision to discover what factors in the ML pipeline affect the performance of the proposed correlation measures, we aim to conduct our research with the following two subquestions in mind:

(i) *What is the best correlation-based feature selection technique to be used considering the dimensionality and feature types (discrete, continuous, nominal or ordinal) that characterize the data?*

(ii) *How much does the choice of ML algorithm that will consume the data influence the performance of correlation-based feature selection techniques?*

Pointwise, the main contributions of our research can be summarized as follows:

- The design and implementation of a pipeline that aims to simulate an AutoML-like setting and is able to efficiently analyze multiple configurations of ML tasks.

- The formalization and implementation of two heuristic approaches for correlation-based feature selection theoretically described in previous work [6, 8, 9] and used for our experiments.
- An empirical and statistical analysis showing how different dataset attributes (types of features, ratio of rows to columns) correlate with variations in the effectiveness and efficiency of correlation-based feature selection techniques.
- An assessment of whether data encoding can alleviate the downsides of breaking the theoretical assumptions of correlation measures.
- An empirical analysis evaluating whether feature selection, given the ML algorithm, can automatically infer a de facto correlation-based technique to apply in order to maximize performance.

Our findings add to the existing literature on feature selection presented in Section 2. Section 3 provides an overview of the proposed correlation techniques and other concepts involved in the evaluation. Following, the ML pipeline employed throughout our research is introduced in Section 4. Section 5 presents the results obtained by testing correlation-based feature selection on top of a plethora of algorithms and datasets. Complementarily, in Section 6 the key findings are analysed and novel conclusions are drawn. Finally, Section 7 concludes by summarizing our paper and suggesting some directions for further research.

2 Related work

In the field of AutoML, feature selection constitutes an underdeveloped literature topic in comparison to other stages of the ML pipeline, such as algorithm selection and hyperparameter optimization. Hence, no golden rules have been devised in prior work, and the exploration of many different methods best defines the literature's current state. Nevertheless, four categories of feature selection can be identified, with filter methods being the focus of our study.

2.1 Feature selection

Feature selection methods designed with different evaluation criteria fall into four broad categories:

(i) *Filter methods* [7–9, 11–15]. These assess the relevance of the features by relying only on the characteristics of the data, thus being independent of the ML algorithm.

(ii) *Wrapper methods* [4, 8, 9, 13, 14, 16]. They employ the ML algorithm as a black box to generate and score promising feature subsets according to the predictive power they give.

(iii) *Embedded methods* [4, 9, 13, 17]. These integrate feature selection in the entire process of training the ML algorithm.

(iv) *Hybrid methods* [6, 14]. They attempt to take advantage of the other methodologies by exploiting their specific evaluation criteria in different stages of the search for relevant features.

It has been argued that the filter-based techniques offer "a simple and powerful way to address the problem of variable selection" [9], which is reflected in the existing body of literature [10, 13, 18]. Thus, we make filter feature selection the focal point of our research.

2.2 Filter techniques

Filter-based feature selection techniques solely consider the association between the set of features and the target. In the first step, a suitable evaluation function is employed to assign all features with a relevance score, using either a univariate or a multivariate scheme. In the univariate case, each feature is evaluated individually based on its usefulness in discriminating the classes of the response feature [19]. In the multivariate case, multiple features are evaluated as a batch, thus additionally considering the interdependencies between features. In the second step, all features are ordered in ascending order of their relevance score, while the concluding step involves using a heuristic approach to select the best columns to be used for training [4, 7, 8, 13].

The advantages of filter techniques are manifold, and we mention the ones empirically shown through our work: (i) The feature selection step only needs to be performed once for a certain dataset, and then multiple algorithms can be trained and evaluated in parallel. (ii) These methods easily scale to high-dimensional databases, making them indispensable in finding reduced subsets in those cases where databases possess a large number of features. (iii) They take little computational power to select the optimal subset of features, especially when compared to other feature selection methods. Nevertheless, filter techniques bring a prominent disadvantage: due to the absence of a specific ML model guiding feature selection, the selected column subset might give little predictive power for the target algorithm [4, 7, 13, 15].

Following the classification of [16], correlation-based techniques are a filter approach in the context of feature selection [4, 7] since they use a correlation metric to assess the importance of the features and, subsequently, filter out the columns that are useless in predicting the target. While feature selection based on correlation has been extensively considered before, a prevalent aspect of existing research is the proposition of various correlation measures, accompanied by an evaluation based on their capability to improve the performance of ML classifiers [13, 18, 20, 21]. However, these publications lack an analysis of the underlying factors in the ML pipeline that can contribute to the performance discrepancies observed when employing different correlation measures. Thus, they do not cover a crucial consideration, namely, which correlation metric to choose in order to obtain the most high-performing feature subset tailored to each specific ML problem. Our research aims to fill in this gap by investigating which factors of the ML pipeline, such as the characteristics of the data or the choice of ML algorithm, influence the performance obtained by some of the most common correlation metrics, in particular Pearson, Spearman, Cramér's V and Symmetric Uncertainty.

3 Preliminaries

In this section we provide (i) an overview of the proposed correlation measures, (ii) a formalization of two approaches for correlation-based feature selection defined in existing literature, (iii) a brief introduction to three well-known classes of ML algorithms and (iv) an outline of the evaluation metrics employed throughout our experiments.

In order to facilitate a consistent and coherent explanation of the formulas and algorithms that are involved in our work, for the rest of the paper we adopt a standard notation to represent relational data in two-dimensional (row and column) for-

Table 1: Notations.

Notation	Description or Definition
$D \in \mathbb{R}^{NXM}$	data table with N rows and M columns
M	number of columns in data table
N	number of rows in data table
X	full column set
X^*	selected column subset
X_i	i^{th} column, where $i \in [1, M]$
f	function to assess correlation
f_i	correlation score of i^{th} column
x_{ij}	j^{th} row of i^{th} column, where $i \in [1, M]$ and $j \in [1, N]$
$Y \in \mathbb{R}^N$	target column for all N rows
y_j	value of target column for j^{th} row, where $j \in [1, N]$
$\bar{y_j}$	predicted value of target column for j^{th} row, where $j \in [1, N]$

mat. The symbols are uniformly described in Table 1, where rows and columns correspond to instances and features in the data, respectively.

3.1 Correlation-based techniques

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

Pearson

Definition. Historically, the Pearson correlation has not only been the first formal measure of correlation but has also emerged as the prevailing method for computing the association between two continuous variables [22, 23]. The correlation coefficient of the population is formally defined in Equation 1 [8, 9]:

$$P(X_i, Y) = \frac{cov(X_i, Y)}{\sqrt{var(X_i) \cdot var(Y)}}, \tag{1}$$

where $cov()$ designates the covariance and $var()$ the variance between the two features X_i and Y.

Equivalently, the correlation coefficient estimate of a sample is defined in Equation 2 [9, 22–24]:

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

where $\bar{X_i} = \frac{1}{N} \cdot \sum_{j=1}^{N} x_{ij}$ and $\bar{Y} = \frac{1}{N} \cdot \sum_{j=1}^{N} y_j$, representing the mean values of X_i and Y. In the numerator, the raw values of the table cells are centered by subtracting the mean value of the respective column, and then the sum of cross-products of the centered columns is accumulated. By adjusting the scales of the columns in the denominator, the relationship between columns that have been measured in different units can be established [24].

Interpretation and assumptions. The Pearson correlation $P(X_i, Y)$ ranges from -1 to 1 and measures the linear component of the relationship between the two columns under consideration [9, 24]. If $P(X_i, Y) = 0$, then the features are considered to be uncorrelated. The closer the value of $|P(X_i, Y)|$ is to 1, the stronger the linear correlation between the two columns is [22, 25]. It is important to note that the Pearson correlation entails certain assumptions: X_i and Y should be continuous, follow a bivariate normal distribution and have a linear relationship [26]. Furthermore, because of its known sensitivity to outliers, data without outliers is preferred if Pearson is used for computing the correlation [23, 27].

Spearman

Definition. Widely used to compute the association of variables measured in an interval or ordinal scale [23, 26], the Spearman rank correlation is computed in the same manner as the Pearson correlation, given by Equations 1 and 2. However, a key distinction exists: while the calculation of $P(X_i, Y)$ involves using the actual sample values, the computation of $S(X_i, Y)$ involves transforming the sample values to ranks in the range [1, N]. To perform a rank transformation, the values in each of X_i and Y are ordered in ascending order and then assigned an integer $r \in [1, N]$ [26], with equal values being assigned the average rank [23, 28].

Interpretation and assumptions. By converting the values into ranks, Spearman correlation focuses on the relative magnitude of the values, allowing it to assess the strength of general monotonicity of the relationship between X_i and Y [26] and to be robust to outliers in the data [23]. Consequently, the assumptions are less restrictive: X_i and Y should be continuous or ordinal and have a monotonic relationship [27]. The range of values, as well as the corresponding interpretation, remains consistent with that of Pearson correlation.

Cramér's V

Definition and assumptions. When the features X_i and Y are nominal and have at least two categories (i.e. distinct values), their association can be assessed by cross-tabulating the data in a contingency table and computing Cramér's V correlation value. Cramér's V is derived from the chi-squared statistic, denoted by χ^2, which tests whether the association between two columns is significant, with the null hypothesis that the variables are independent. More specifically, it compares the observed frequencies, which are taken from the contingency table, to the expected frequencies generated following the null hypothesis [29]. The formula for Cramér's V is given in Equation 3:

$$C(X_i, Y) = \sqrt{\frac{\chi^2}{N \cdot min(C_{X_i} - 1, C_Y - 1)}}, \tag{3}$$

where C_{X_i} and C_Y denote the number of categories of X_i and Y, respectively [30–32].

Interpretation. Cramér's V correlation $C(X_i, Y)$ ranges from 0 to 1. A value of 0 for $C(X_i, Y)$ indicates little to no association between column X_i and target Y. Conversely, a value of 1 reflects a perfect relationship between the two features [30,31,33]. However, the interpretation of the measures of association (i.e. "weak", "moderate", "strong") is always relative to the domain and size of the data [32].

Symmetric Uncertainty

Definition. SU is a correlation measure based on the information-theoretical concept of entropy, which quantifies the uncertainty of a discrete random variable. The entropy of feature X_i can be computed from Equation 4:

$$H(X_i) = -\sum_{j=1}^{N} P(x_{ij}) \cdot log(P(x_{ij})), \tag{4}$$

where $P(x_{ij})$ refers to the prior probabilities for all values of X_i. After observing the values of the target column Y, the

TABLE 1: Feature types assumptions of the proposed correlation measures.

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

entropy of column X_i can be defined using Equation 5:

$$H(X_i|Y) = -\sum_{j=1}^{N} P(y_j) \cdot \sum_{j=1}^{N} P(x_{ij}|y_j) \cdot log(P(x_{ij}|y_j)), \quad (5)$$

where $P(y_j)$ denotes the prior probabilities for the values of Y, whereas $P(x_{ij}|y_j)$ refers to the posterior probabilities of X_i given Y. The amount by which the uncertainty of X_i decreases represents how much information the feature and target share together. This amount is referred to as information gain and is given by Equation 6:

$$IG(X_i, Y) = H(X_i) - H(X_i|Y). \quad (6)$$

Finally, by normalizing the information gain $I(X_i, Y)$ to the entropies $H(X_i)$ and $H(Y)$, the formula for computing SU is obtained in Equation 7 [7, 19, 34, 35]:

$$SU(X_i, Y) = \frac{2 \cdot IG(X_i, Y)}{H(X_i) + H(Y)}. \quad (7)$$

Interpretation and assumptions. SU is an extension to information gain that normalizes its values within the range [0, 1] and compensates for its bias when the features X_i and Y have many distinct values. $SU(X_i, Y) = 0$ means that the columns are independent, while $SU(X_i, Y) = 1$ denotes that knowledge of the value of X_i completely predicts the value of Y and vice versa [19, 34]. The entropy-based correlation of the SU measure requires nominal features [35, 36].

The main assumption of the proposed correlation methods refers to the type(s) of features. As this represents the focus of our study, we provide an overview of the types of variables the methods are designed for in TABLE 2.

3.2 Heuristics for filter-based feature selection

Established literature introduces a plethora of both univariate and multivariate heuristic approaches for feature selection which rely on statistical measures. Given the limited scope of our paper, we solely base our experiments on univariate approaches, and do not delve into multivariate approaches such as the well-known CFS (Correlation-Based Feature Selection) algorithm [10].

We now formalize two simple, yet effective, univariate heuristic approaches theoretically described in existing literature [6, 8, 9], which we further refer to as SELECT k BEST and SELECT ABOVE c. Both approaches can be used to select a subset of features X^* from any data table D by computing the absolute value of the correlation between each feature X_i and the target Y. However, while SELECT k BEST ranks the variables with regard to the correlation and then selects the k top-performing ones, SELECT ABOVE c chooses all features where the correlation is above a defined threshold c. Evidently, incorporating higher values of k or lower values of c will progressively add to the selected feature subset more and more variables of decreasing relevance to the target Y.

Heuristic approach 1: SELECT k BEST

Data: Full feature set X. Target feature Y.
Correlation function f. Number of features to select k
Result: Optimal feature subset X^*
$X^* \leftarrow \emptyset$
$M \leftarrow$ length of X
$S \leftarrow \{(X_1, f(X_1, Y)), (X_2, f(X_2, Y)), ..., (X_M, f(X_M, Y))\}$
sort S **in descending order of** $abs(f)$, where $(X, f) \in S$
for $i \leftarrow 1$ **to** k **do**
  add $S[i]$ **to** X^*
end
return X^*

Heuristic approach 2: SELECT ABOVE c

Data: Full feature set X. Target feature Y.
Correlation function f. Correlation threshold c
Result: Optimal feature subset X^*
$X^* \leftarrow \emptyset$
for $X_i \in X$ **do**
  if $abs(f(X_i, Y)) \geq c$ **then**
    add X_i **to** X^*
  end
end
return X^*

3.3 Machine learning algorithms

We introduce five widely-employed ML algorithms, in particular LightGBM, Random forest, XGBoost, Linear model, and Support vector machine, that will be utilised throughout our research due to their ability to handle high-dimensional data, robustness against overfitting and ability to capture complex data relationships [37].

(i) *Decision trees.* Gradient boosting decision tree (GBDT) [38] and Random forest (RF) [39] are instances of ensemble learning. While GBDT sequentially combines decision trees to achieve a robust model, RF constructs multiple decision trees in parallel and aggregates their predictions. Recent literature introduces quite a few high-performing, enhanced implementations of GBDT, among which we choose XGBoost [40] and LightGBM [41]. In comparison to traditional implementations, XGBoost and LightGBM incorporate sparse data optimization techniques, allowing them to handle data tables with a large number of rows and columns more efficiently.

(ii) *Linear model* [42] is a supervised learning model that aims to explain the target feature in terms of a linear combination of the explanatory features plus an error term, incorporating both Logistic regression and Linear regression. Whereas Logistic regression solely supports predictions for binary classification tasks by estimating the relationship between input features and class probabilities, Linear model extends this concept to support linear regression tasks, with the goal of predicting continuous target features.

(iii) *Support vector machines (SVMs)* [43] are combination of linear modeling and instance-based learning. SVMs select a number of critical boundary samples from each class and build a linear discriminant function that separates them as widely as possible. When no linear

Table 1: Experimental datasets.

Dataset	Rows	Columns ↓	Column type				Task
		(excluding target)	Discrete	Continuous	Nominal	Ordinal	
Census Income	48,842	14	0	6	8	0	Binary classification
Breast Cancer	569	31	0	31	0	0	Binary classification
Steel Plates Faults	1,941	33	8	25	0	0	Binary classification
Internet Advertisements	3,279	1,558	0	3	1,555	0	Binary classification
Gisette	6,000	5,000	5,000	0	0	0	Binary classification
Nursery	12,960	8	0	0	1	7	Multi-class classification
Connect-4	67,557	42	0	0	42	0	Multi-class classification
Arrhythmia	452	279	86	120	73	0	Multi-class classification
Bike Sharing	17,379	16	8	7	0	1	Regression
Housing Prices	1,460	80	3	33	27	17	Regression

separation is possible, the technique of "kernel" is employed to automatically inject the samples into a higher-dimensional space, and to learn a separator, called hyperplane, in that space [44].

3.4 Machine learning metrics

The assessment of ML algorithms is twofold: (i) efficiency-based: *how many computational resources are required to train the model?*, and (ii) effectiveness-based: *how accurate is the model in making classifications or predictions?* We provide a brief overview of the metrics that were deemed suitable to quantify both efficiency and effectiveness.

(i) *Effectiveness.* As our study concerns the two most common ML problems on tabular data (i.e. regression and classification), we rely on two regularly employed metrics in ML evaluation studies: (a) accuracy [45] for classification tasks, and (b) root mean square error [46] for regression tasks. Accuracy, computed using Equation 8, represents the ratio of correct predictions to the total number of predictions N for a dataset D; a higher value is desirable [47]. Complementarily, root-mean-square error (RMSE), formally defined in Equation 9, measures the average deviation between all N predictions of the model and the actual values of the target Y; a lower value is preferred.

$$\text{accuracy} = \frac{\sum_{j=1}^{N} \begin{cases} 1 & \text{if } y_j = \bar{y_j} \\ 0 & \text{otherwise} \end{cases}}{N} \tag{8}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{N} (y_j - \bar{y_j})^2}{N}} \tag{9}$$

(ii) *Efficiency.* In order to evaluate the efficiency and, implicitly, the usefulness of correlation-based feature selection in AutoML systems, we empirically compare the execution times of different tasks in the ML pipeline.

4 Methodology

In this section, we present the key components of our evaluation: (i) ten datasets with diverse characteristics, and (ii) a ML pipeline designed to streamline the experiments and analyse the four correlation-based feature selection techniques in a manner that simulates an AutoML-like setting.

4.1 Data

To conduct a robust evaluation of the correlation-based methods proposed in this study, we have selected ten datasets. Even though they all adhere to the tabular format introduced in Section 3, their properties vary in terms of (i) domain (medicine, education, economy, social science), (ii) ratio of rows to columns, (iii) number of numerical and categorical columns and (iv) task (binary classification, multi-class classification and regression). A general overview of the datasets, grouped by the type of task and ordered ascendingly by the number of columns, can be visualized in Table 3. Selecting these ten datasets has been deemed reasonable for our experiments, and space considerations keep us from presenting results with other data characteristics.

4.2 Machine learning pipeline

At its core, every effective ML system needs to address certain decisions: which ML algorithm to use and what hyperparameters to set for a given database, whether and how to preprocess the columns, and what evaluation metric to choose [48]. With a vision to automatically infer which correlation-based feature selection technique to apply given a data table and an algorithm, we design a ML pipeline that can efficiently analyze multiple configurations of ML tasks. The proposed pipeline, depicted in Fig. 2, can be divided into three stages that encompass all the necessary sub-tasks:

(i) **Pre-ML**, that attempts to achieve high-quality, appropriate data. In order to do so, it utilises the following techniques:

(a) *Imputation.* In the context of real-world data, missing values are frequently encountered, yet the existing implementations of correlation methods and ML algorithms typically do not support such missing values. Consequently, it becomes necessary to address this issue by imputing the missing values. Our preliminary results have shown that the imputation method does not affect the conclusion drawn; thus, in order to streamline our experiments, we proceed exclusively with the imputation of the most common value.

(b) *Normalization.* In the case of SVMs, the prediction results are dramatically influenced by the feature scales [49]. Consequently, in order to avoid suboptimal results, it is necessary to normalize the features, for which we employ min-max scaling.

(c) *Encoding.* Given that most real-world datasets do not contain a singular type of feature, our research aims to explore the potential benefit of employing data encoding techniques to enhance correlation-based feature selection and alleviate the potential negative effects of breaking the assumptions listed

in Table 2. By default, we investigate all datasets keeping the original columns. If it is deemed necessary to further investigate whether the performance of a correlation method could be improved, we can investigate the datasets by (i) transforming all columns to continuous using one-hot encoding or (ii) transforming all columns to nominal through K-bins discretization.

(ii) **In-ML** aims to perform feature selection with the proposed techniques and train the selected model on the adjusted data table but only using the train set. It comprises the following two stages: (a) *Feature selection*, where for each pipeline configuration we choose one heuristic approach (Select k best or Select above c) and one correlation method (Pearson, Spearman, Cramér's V or SU). (b) *Model selection and training*, where we deploy an ML algorithm (LightGBM, RF, XGBoost, Linear model or SVM) on the dataset after feature selection.

(iii) **Post-ML**, that assesses performance with regard to other model variants. It consists of the following two sub-tasks: (a) *Model evaluation*, where we look at the accuracy or RMSE on the test set and the runtime of the training stage of the ML algorithm. (b) *Feature selection evaluation*, where we analyse the runtime of the feature selection stage, in order to gain insights into how each correlation technique is affected by the dimensionality of the data.

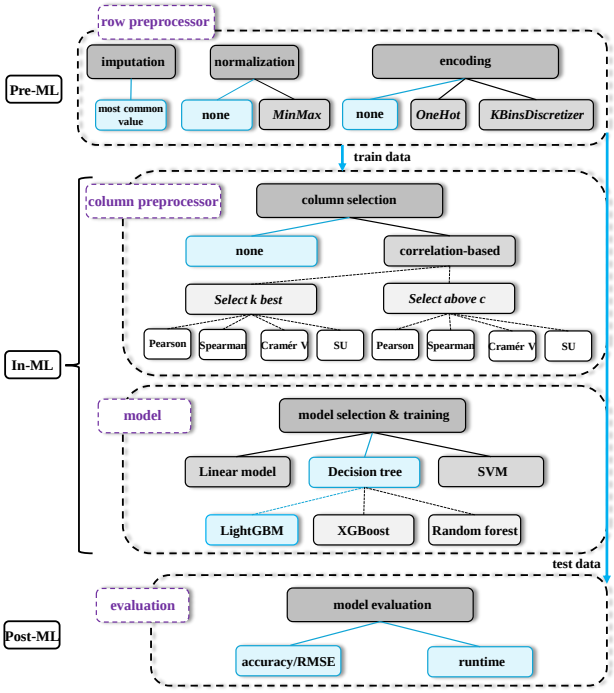

Figure 2: Configuration space for the ML pipeline. The blue-coloured hyperparameters form an instance of the *baseline* ML pipeline (no feature selection). Each configuration of the pipeline comprises up to two row preprocessors, none or one column preprocessor, one model and two evaluation metrics.

5 Evaluation

This section incorporates together the different aspects of a data-driven empirical evaluation: (i) the technical setup of the experiments, and (ii) an overview of the results, accompanied by the configuration of the ML pipeline employed for conducting the experiments.

It is worth mentioning that, due to space considerations, only a subset of the datasets and one heuristic approach will be used to report and reason about each experiment. Nonetheless, the conclusions derived within the scope of an experiment extend to all the datasets that are not mentioned.

5.1 Experimental setup

We ran our experiments on a server provided by the EEMCS Faculty at Delft University of Technology. Two 64-core AMD Epyc 7H12 processors were made available to us, along with 511 GiB of memory. Besides the hardware resources, we built the ML pipeline employed for conducting the experiments in Python 3.10. The correlation-based feature selection techniques were developed using the SciPy package and the Scikit-feature GitHub repository [7]. The SVM algorithm was evaluated using the implementation from the Scikit-learn library, whereas the other ML algorithms were employed using their AutoGluon implementations [37].

5.2 Empirical results

Data characteristics

Effectiveness. The experiments visualized in Fig. 3 and Fig. 4 examine the ability of the considered correlation methods to deal with the different types of features. We consider the accuracy achieved by each correlation method averaged over all five ML algorithms, thus eliminating potential algorithm dependency. We employ the Select k best heuristic approach, with increasing values of k, the value that corresponds to the number of variables retained after feature selection. Additionally, two-tailed independent t-tests are employed to assess the significance of the difference in accuracy among two correlation methods for the different values of the number of features. A p-value below 0.05 is considered indicative of a significant difference in accuracy.

(i) Numerical features. On one hand, when we consider datasets containing only *discrete* features (Gisette), computing the optimal feature subset with the SU measure leads to the best accuracy for most values of k. However, we note that the difference in accuracy for the four methods is insignificant here. On the other hand, for datasets with *continuous* features (Breast Cancer), Pearson and Spearman exhibit higher accuracy compared to the other methods. Even more so, with only half the features, Pearson achieves a 1.57% higher accuracy than the baseline. We note, however, the surprisingly good results given by Cramér's V and SU, considering that they are normally suitable for categorical features. When we combine *discrete and continuous* features (Steel Plates Faults), the accuracy obtained deviates significantly when Cramér's V and SU are employed for feature selection. We investigate the effects of encoding all features to categorical, and we notice that the performance of these two methods is greatly improved on the encoded dataset, as depicted in Fig. 4.

(ii) Categorical features. In the case of datasets with predominantly *nominal* features (Internet Advertisements, Connect-4) Cramér's V and SU are the better choices for

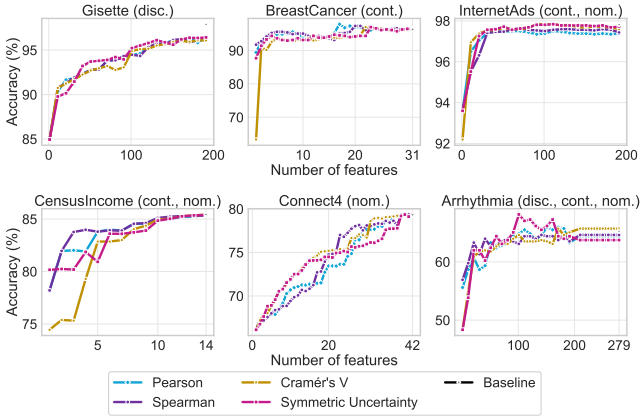

FIGURE 3: Accuracy obtained by the four correlation-based feature selection techniques averaged over all five ML algorithms and computed for an increasing number of features retained by feature selection. For each dataset, the feature types present in the data are mentioned next to the title.

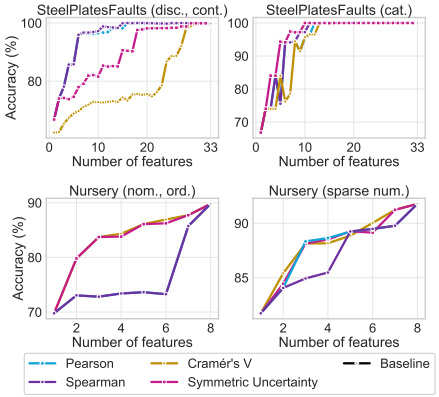

FIGURE 4: Accuracy obtained by the four correlation-based feature selection methods on the original datasets (left) and the encoded datasets (right). In the lower figure on the left, the Pearson line exactly overlaps the Spearman one.

computing the correlation between features and target at different levels of k. However, the difference between the four methods is not statistically significant. When it comes to predominantly *ordinal* data (Nursery), the statistical correlation measures of Pearson and Spearman perform significantly worse, as can be observed in FIG. 4. That being said, this is a surprising result for Spearman, which is designed for ordinal features. Nevertheless, when applying one-hot encoding, the accuracy obtained by Pearson and Spearman improves, as shown in FIG. 4.

(iii) Mixed features. Upon combining *continuous and nominal* features (Census Income), we notice, once again, the lower performance of Cramér's V on datasets that contain continuous data. When we also consider *discrete* features (Arrhythmia), all methods are able to perform the best at different levels of k, and on average their behaviour is similar. However, SU is able to achieve a 0.65% higher accuracy than the baseline with only a third of the total number of features.

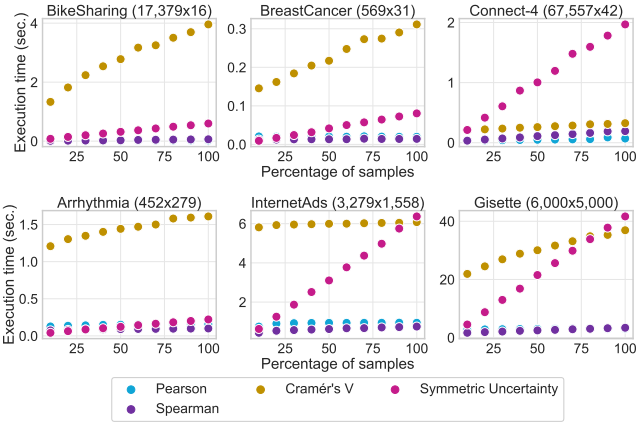

FIGURE 5: Execution time of the feature selection stage using the four correlation methods computed for increasing percentage of the number of samples. For each dataset, the ratio of rows to columns is mentioned next to the title.

Efficiency. To assess the variability of the execution time of feature selection for each correlation method, we examine the computational time required by the SELECT ABOVE c heuristic to compute all correlation values and select the features with a correlation above 0.5 with the target. Additionally, we systematically vary the percentage of data instances utilized by feature selection. The results, depicted in FIG. 5, show that Pearson and Spearman are quite insensitive to both the number of columns and the number of rows in the data, being able to achieve the lowest runtime for all datasets. In fact, in most cases, they require less than 0.1 seconds to compute the ranked set of features and to return the top-performing ones. SU is the most sensitive method to the number of instances used for feature selection, with the computational cost growing exponentially. Here we notice a trade-off between effectiveness and efficiency, as our findings show that SU gives higher accuracy as the number of samples is increased. Lastly, Cramér's V, though less sensitive to the number of samples than SU, is the most time-expensive correlation measure for computing correlation when compared to the other three. Lastly, in our experiments, we observe that the types of features do not greatly influence the execution time of the correlation-based techniques.

Machine learning algorithm

Effectiveness. The experiments obtained by running SELECT k BEST on all datasets and algorithms show that the behaviour of the correlation-based techniques is consistent across all algorithms employed on the same dataset, as can be seen in the figures included in Appendix A. Nevertheless, as previously mentioned, the behaviour across different datasets fluctuates significantly. Thus, we conclude that the effectiveness of correlation-based methods is primarily influenced by the inherent characteristics of the dataset rather than the choice of algorithm.

Efficiency. To analyse the effects of including correlation-based feature selection in the training stage of ML algorithms, we investigate the execution time of the **In-ML** stage of the pipeline in FIG. 6. We compare the training time with feature selection, averaged across all five models, to the baseline. We observe that, in the case of datasets with lower numbers of features (Breast Cancer, Steel Plates Faults), including feature

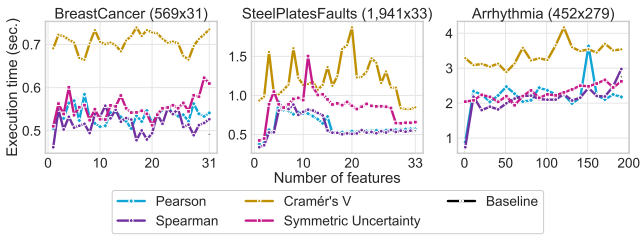

FIGURE 6: Execution time of the **In-ML** stage using the four correlation methods computed for an increasing number of features.

selection in the training stage is always more computationally costly than the baseline. However, with more features (Arrhythmia), the execution time actually benefits from employing feature selection.

6 Discussion and limitations

Notably, the results indicate that the values of effectiveness obtained when applying the proposed correlation-based methods are highly tied to the ratio of numerical and categorical columns and the feature types characteristic to the data. Additionally, we have empirically demonstrated that the theoretical assumptions of the correlation measures regarding the nature of features do not always hold in practice. Consequently, it is necessary that we devise new ones, that are shown in TABLE 4. Moreover, no specific correlation method has been identified to exhibit superior performance exclusively for a particular algorithm. This behaviour is unsurprising since filter methods do not take into consideration the ML algorithm in the feature selection process. In regard to the correlation measures, we note the key findings:

(i) *SU* is the only correlation measure that has performed relatively well on all datasets, being suitable for all types of features. However, it fails to outperform statistical measures of correlation for numerical features, with the most significant outlier being Steel Plates Faults. We notice a trade-off in the case of datasets of high dimensionality: even though the estimated probabilities and, hence, entropy values involved in the calculation of SU are more accurate in the presence of more data instances, the computational cost can grow exponentially.

(ii) *Pearson and Spearman* exhibit similar behaviour, with a common limitation of not effectively improving the accuracy when used on ordinal data. Between the two methods, Spearman emerges as the preferred choice due to its advantage of requiring less computational time.

(iii) *Cramér's V* is the lowest-performing measure of correlation, achieving on multiple datasets the lowest accuracy, as well as the highest runtime. We deem it unsuitable for continuous features due to its tendency to yield low accuracy when the number of features is limited.

There exist a few limitations worth exploring in the evaluation presented thus far. These include (i) *K-fold cross-validation omission.* Due to the inherent complexity of the proposed ML pipeline and the multitude of potential configurations, k-fold cross-validation was not employed during the execution of the experiments. However, we implemented this technique in our codebase, so that it can be incorporated in future experiments. (ii) *Limited scope of feature preprocessing techniques* that were utilized throughout our research.

TABLE 6: Feature types suitable in practice for the proposed correlation measures.

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

Our choices of imputation, normalization and encoding techniques may not fully capture the variability and impact of employing other preprocessing approaches.

7 Concluding remarks and future directions

In this paper, we provided a data-driven empirical analysis of four widely-known correlation-based feature selection techniques: Pearson, Spearman, Cramér's V and Symmetric Uncertainty. In particular, we focused on investigating which correlation measure should be used for feature selection in order to maximize the performance of ML systems, while considering the choice of dataset and algorithm. To this end, we devised an ML pipeline that resembles an AutoML-like setting, and we formalized two heuristic approaches for feature selection in order to conduct the experimental evaluation.

By addressing the first research sub-question, we discovered that the effectiveness and efficiency of the correlation-based feature selection techniques are highly influenced by two main data characteristics, in particular, the ratio of rows to columns and the feature types characteristic to the dataset. As future recommendations for AutoML systems, if effectiveness is desired, SU should be employed, as it usually works well with all types of features. However, if efficiency is prioritized, Spearman is the best choice that can handle high-dimensional datasets without compromising a significant degree of effectiveness.

The exploration of the second research sub-question revealed that the performance of correlation-based techniques is independent of the ML model, remaining consistent across different algorithms on the same dataset. Nonetheless, our empirical work demonstrates that incorporating feature selection in AutoML systems, in spite of a small increase in execution time, can yield similar or higher accuracy compared to the baseline when retaining as few as one-third of the total number of features.

Finally, some propositions of future work include (i) *The exploration of other existing heuristic approaches for correlation-based feature selection.* In particular, based on preliminary results, we propose the investigation of a promising novel heuristic approach that takes the union of the optimal feature subsets returned by each of the correlation measures. (ii) *The exploration of other correlation metrics.* Our work can be extended with other promising correlation measures, such as Kendall's Tau [20], for which the devised ML pipeline could still be employed for their evaluation. (iii) *The evaluation of the techniques on a diverse collection of real-world datasets.* Given that our research lies the foundational work for determining the most suitable correlation metric based on the type(s) of features present in the data, we encourage future researchers to explore more datasets and undertake further investigations on other data characteristics, such as the existence of outliers or the type of distribution.

8 Responsible research

To maintain the integrity of our results and to uphold high ethical standards of our research, several precautions and concerns have been considered. We review our process based on two criteria: (i) ethical implications in ML, and (ii) reproducibility of our research outcomes.

8.1 Ethical implications in machine learning

Our research aims to analyse how correlation-based feature selection techniques, in particular Pearson, Spearman, Cramér's V, Symmetric Uncertainty, can influence the performance of ML systems, considering the dataset characteristics and the ML algorithm. We strongly believe that our work can increase explainability, interpretability, and justifiability of decision-making in AutoML systems for two main reasons. (i) By using correlation-based feature selection, the ML pipeline can provide a transparent explanation of the relevant features, their relationship with the target variable, and how they contribute to the model's predictions. (ii) By examining the selected features, researchers can identify potential biases or errors in the ML pipeline, therefore addressing any unintended consequences. In this sense, we acknowledge the possibility of a particular feature that has been historically discriminated against, such as ethnicity, age or gender, to be found highly correlated with the response feature. Consequently, users may draw conclusions that correlation-based feature selection reinforces this discrimination. Nevertheless, the opposite might also happen: feature selection can uncover and bring to attention certain types of discrimination, allowing data engineers to take action against them.

It is worth noting that the choice of datasets encompasses various fields, such as medicine, economy and engineering, to ensure the generalization of the results across domains. The data used in our research is publicly available, and was obtained from widely-used sources: OpenML, Kaggle and UC Irvine. In addition, while selecting the data sources, several factors were considered: consent, copyright, re-identification, data storage and manipulation. As part of these considerations, all the datasets are available under a permissive license and, for datasets involving human data, strict anonymization protocols have been implemented in order to ensure the confidentiality and privacy of the participants.

We note that our study was conducted without external funding or any conflicts of interest. While the results in and of themselves are not malicious, this does not prevent the proposed techniques and heuristics from being utilised by malicious users. However, according to the tripartite model [50], engineers can only be held ethically liable for the technical decision and engineering choices that they make. Thus, while we strongly condemn malicious actions and emphasize that the purpose of our research is to improve the performance of ethical AutoML systems, we cannot be held ethically responsible for the actions of those making use of our research.

8.2 Reproducibility of results

To ensure that the integrity of the results is guaranteed, the entire experimental process and results discussed in our paper are reproducible. In this sense, the following aspects have been considered:

(i) *Data availability.* Ten datasets were employed throughout our paper. To enhance the reusability of the datasets by future researchers, they adhere to the FAIR principles [51]: (i) **findability**, as all the data is available on our GitHub repository, (ii) **accessibility**, because the GitHub repository is public, no authentication is required to have access to the data, (iii) **interoperability**, as all datasets are stored in a common CSV format, and (iv) **reusability**, as the repository contains a summary of the datasets, which details on their characteristics (ratio of columns to rows, distribution of feature types, type of ML task) and contains easy-to-understand explanations.

(ii) *Source code availability and quality.* We provide the code required to run the exact versions of our experiments as a branch on our GitHub repository. We have additionally included all the files with results that have been generated throughout our research. In order to facilitate future researchers who may want to reproduce or extend our work, we provide: (i) a comprehensive overview of the ML pipeline that has been employed for running the experiments, and (ii) a detailed ReadMe file, which can be found on our GitHub repository. Lastly, the implementation of the ML pipeline adheres to the best practices of software development, incorporating explanatory comments to enhance comprehensibility for users without a technical background. While we cannot guarantee that our software is completely bug-free since the underlying libraries (e.g. AutoGluon, Scikit-learn et cetera) are under active development, we did conduct thorough manual testing to mitigate any risks posed by software bugs.

(iii) *Effectiveness metric reproducibility.* All pseudo-random number generators used in the experimental setup use a fixed seed, with a value of 0. That being said, we expect future runs of the experiments to generate identical outcomes for the metrics of effectiveness (i.e. accuracy and RMSE) or other parameters involved in our study, such as the number of selected features or the correlation values.

(iv) *Efficiency metric reproducibility.* It is important to acknowledge that biases with regard to the reported metric of efficiency (i.e. execution time) may have occurred during the evaluation process due to factors such as, but not limited to, the shared nature of the server employed for the experiments. To minimize the potential impacts of the limitation created by concurrent activities on the hardware resources, all experiments have been repeated three times and the presented results are averaged over all runs. Nevertheless, we note that the execution times might differ when the experiments are rerun by other researchers.

(v) *Citations.* All ideas borrowed from other works are properly cited in our paper, thus allowing readers to dive further in-depth into the topics of interest.

9 Acknowledgements

First and foremost, I would like to acknowledge the tireless and prompt help of my supervisor, Andra Ionescu. She has always challenged me to push beyond my limits and strive for excellence. Furthermore, I would like to express my utmost gratitude to my responsible professor, Asterios Katsifodimos, for his remarkable expertise and invaluable guidance that helped me think outside the box. Lastly, I would

like to thank my peers (Andrei Manastireanu, Andrei Udila, Kiril Vasilev and Duyemo Anceaux) for their feedback and help, which significantly improved the quality of this paper.

References

[1] F. Hutter, L. Kotthoff, and J. Vanschoren, *Automated Machine Learning: Methods, Systems, Challenges*. Springer Publishing Company, Incorporated, 1st ed., 05 2019.

[2] L. Tuggener, M. Amirian, K. Rombach, S. Lorwald, A. Varlet, C. Westermann, and T. Stadelmann, "Automated machine learning in practice: State of the art and recent results," in *6th Swiss Conference on Data Science (SDS)*, IEEE, 06 2019.

[3] R. S. M. Lakshmi Patibandla, V. S. Srinivas, S. N. Mohanty, and C. Ranjan Pattanaik, "Automatic machine learning: An exploratory review," in *9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pp. 1–9, 09 2021.

[4] C. Chai, J. Wang, Y. Luo, Z. Niu, and G. Li, "Data management for machine learning: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, pp. 4646–4667, 02 2023.

[5] F. Nargesian, A. Asudeh, and H. V. Jagadish, "Responsible data integration: Next-generation challenges," in *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, (New York), p. 2458–2464, Association for Computing Machinery, 06 2022.

[6] A. A. Gnana Singh, S. Balamurugan, and J. L. EPIPHANY, "Literature review on feature selection methods for high-dimensional data," *International Journal of Computer Applications*, vol. 136, 02 2016.

[7] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, 12 2017.

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

[9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, pp. 1157–1182, 03 2003.

[10] M. Hall, *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 06 2000.

[11] H. Wang, T. M. Khoshgoftaar, and J. Van Hulse, "A comparative study of threshold-based feature selection techniques," in *2010 IEEE International Conference on Granular Computing*, pp. 499–504, IEEE, 08 2010.

[12] S. Rajagopal and K. Subburathinam, "Performance assessment of feature selection methods using k-means on adult dataset," *CCIIT*, vol. 4, pp. 606–612, 01 2013.

[13] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, pp. 2507–2517, 08 2007.

[14] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, pp. 491–502, 03 2005.

[15] N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection – a comparative study," pp. 178–187, 12 2007.

[16] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–324, 12 1997.

[17] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," *IEEE/CAA Journal of Automatica Sinica*, vol. PP, pp. 1–13, 03 2019.

[18] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, vol. 38, pp. 1200–1205, 07 2015.

[19] 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.

[20] J. Van Hulse, T. Khoshgoftaar, A. Napolitano, and R. Wald, "Threshold-based feature selection techniques for high-dimensional bioinformatics data," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 1, 06 2012.

[21] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome informatics. International Conference on Genome Informatics*, vol. 13, pp. 51–60, 01 2002.

[22] H. Zhou, Z. Deng, Y. Xia, and M. Fu, "A new sampling method in particle filter based on Pearson correlation coefficient," *Neurocomputing*, vol. 216, pp. 208–215, 12 2016.

[23] J. de Winter, S. Gosling, and J. Potter, "Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data," *Psychological Methods*, vol. 21, pp. 273–290, 09 2016.

[24] J. L. Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, pp. 59–66, 06 1988.

[25] D. Nettleton, "Chapter 6 - Selection of variables and factor derivation," in *Commercial Data Mining*, pp. 84–85, Boston: Morgan Kaufmann, 01 2014.

[26] R. Artusi, P. Verderio, and E. Marubini, "Bravais-Pearson and Spearman correlation coefficients: Meaning, test of hypothesis and confidence interval," *The International Journal of Biological Markers*, vol. 17, pp. 148–151, 04 2002.

[27] N. S. Chok, "Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data," Master's thesis, University of Pittsburgh, 09 2010.

[28] S. Kokoska and D. Zwillinger, "Chapter 14 - Nonparametric statistics," in *CRC Standard Probability and Statistics Tables and Formulae*, pp. 372–376, 05 2000.

[29] M. J. Fisher, A. P. Marshall, and M. Mitchell, "Testing differences in proportions," *Australian Critical Care*, vol. 24, pp. 133–138, 05 2011.

[30] R. K. Prematunga, "Correlational analysis," *Australian Critical Care*, vol. 25, pp. 195–199, 08 2012.

[31] M. Kearney, *Cramér's V*. 12 2017.

[32] J. Cohen, "Chapter 7 - Chi-square tests for goodness of fit and contingency tables," in *Statistical Power Analysis for the Behavioral Sciences* (J. Cohen, ed.), pp. 215–227, Academic Press, 11 1977.

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

[34] S. Bakhshandeh, R. Azmi, and M. Teshnehlab, "Symmetric uncertainty class-feature association map for feature selection in microarray dataset," *International Journal of Machine Learning and Cybernetics*, vol. 11, 02 2020.

[35] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," vol. 2, pp. 856–863, 01 2003.

[36] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *International Joint Conference on Artificial Intelligence*, 09 1993.

[37] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, and A. Smola, "Autogluon-tabular: Robust and accurate automl for structured data," 03 2020.

[38] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, 10 2001.

[39] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 11 2001.

[40] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, p. 785–794, Association for Computing Machinery, 08 2016.

[41] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 12 2017.

[42] W. Penny, K. Friston, J. Ashburner, S. Kiebel, and T. Nichols, "Chapter 8 - The general linear model," in *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, pp. 101–103, 01 2007.

[43] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, p. 144–152, Association for Computing Machinery, 07 1992.

[44] M. H. Nguyen and F. de la Torre, "Optimal feature selection for support vector machines," *Pattern Recognition*, vol. 43, pp. 584–591, 03 2010.

[45] M. Junker, R. Hoch, and A. Dengel, "On the evaluation of document analysis components by recall, precision, and accuracy," in *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99 (Cat. No.PR00318)*, vol. 5, pp. 713–716, 08 1999.

[46] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?," *Geoscientific Model Development*, vol. 7, pp. 1247–1250, 06 2014.

[47] G. Handelma, H. Kok, R. Chandra, A. Razavi, S. Huang, M. Brooks, M. Lee, and H. Asadi, "Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods," *American Journal of Roentgenology*, vol. 212, pp. 1–6, 10 2018.

[48] M. Feurer, A. Klein, K. Eggensperger, J. T. Springenberg, M. Blum, and F. Hutter, *Auto-sklearn: Efficient and Robust Automated Machine Learning*, pp. 113–134. Springer International Publishing, 05 2019.

[49] A. Graf and S. Borer, "Normalization in support vector machines," pp. 277–282, 03 2007.

[50] I. Poel, van de and L. Royakkers, "Chapter 1 - The responsibilities of engineers," in *Ethics, technology, and engineering: an introduction*, pp. 21–22, Wiley-Blackwell, 01 2011.

[51] M. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. Boiten, L. Da Silva Santos, P. Bourne, J. Bouwman, A. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. Evelo, R. Finkers, A. Gonzalez-Beltran, A. Gray, P. Groth, C. Goble, J. Grethe, J. Heringa, P. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. Lusher, M. Martone, A. Mons, A. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The fair guiding principles for scientific data management and stewardship: Comment," *Scientific Data*, vol. 3, 03 2016.

A Feature selection and machine learning algorithm analysis results

This appendix presents a visual summary of the results obtained by running the SELECT K BEST feature selection approach on each dataset and algorithm. For each one of the ten datasets, five algorithms are considered: (i) decision tree algorithms: LightGBM, Random forest, XGBoost; (ii) Linear model; and (iii) Support vector machine. Additionally, each algorithm is evaluated in five forms: (i) without feature selection, denoted as *baseline*; (ii) with feature selection using Pearson correlation technique; (iii) with feature selection using Spearman correlation technique; (iv) with feature selection using Cramér's V correlation technique; and (v) with feature selection using Symmetric Uncertainty correlation technique. The number of features chosen by feature selection, k, takes all the possible subset sizes in the dataset. The exceptions are Internet advertisements and Gisette datasets, where the maximum sizes considered are 250 and 200, respectively, due to space considerations. Further analysis was performed on the figures in this appendix, but it was decided to be summed up enough to support the results and design of Section 5.

A.1 Census Income dataset

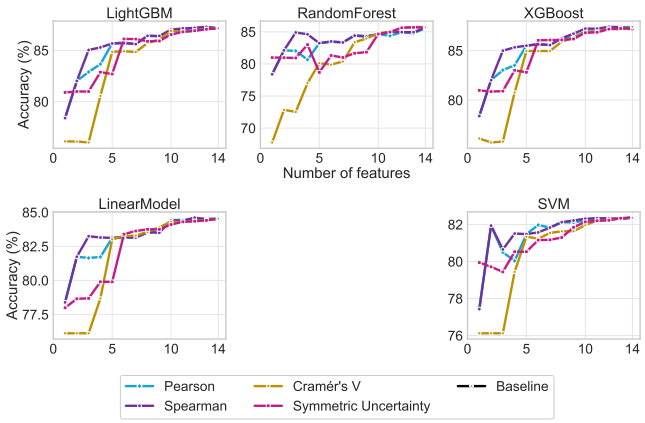

FIGURE 7: Accuracy for Census Income dataset.

A.2 Breast Cancer dataset

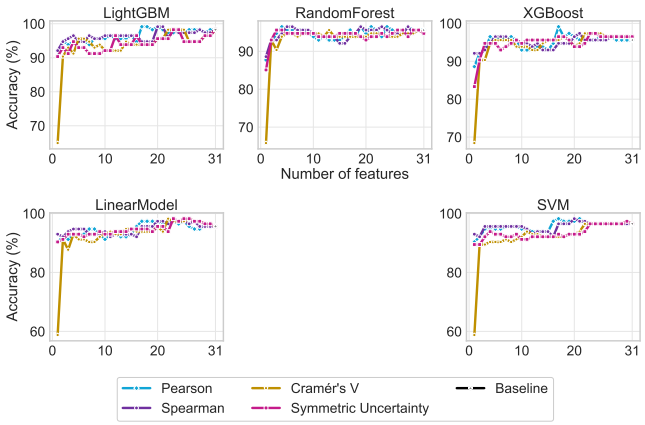

FIGURE 8: Accuracy for Breast Cancer dataset.

A.3 Steel Plates Faults dataset

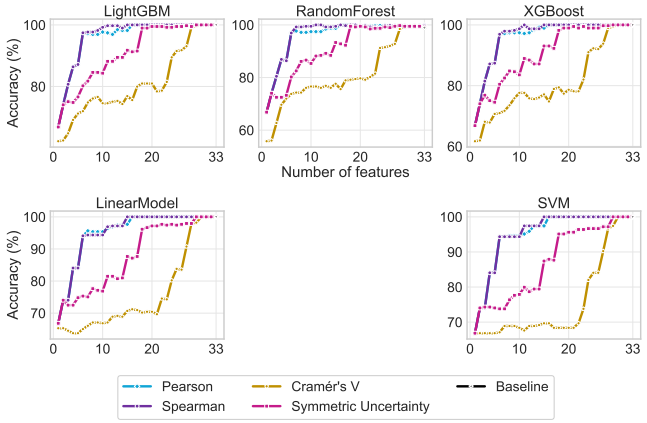

FIGURE 9: Accuracy for Steel Plate Faults dataset.

A.4 Arrhythmia dataset

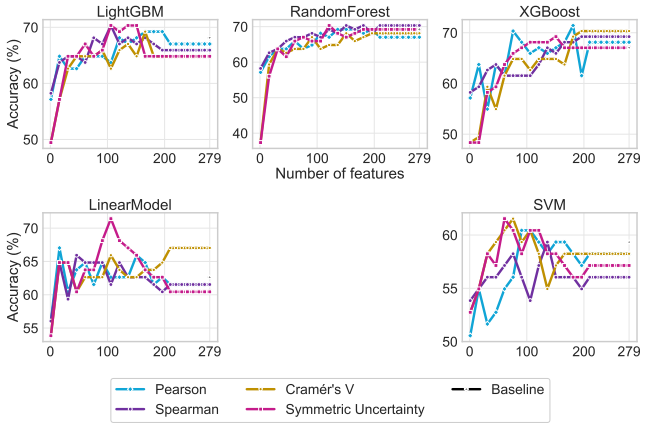

FIGURE 10: Accuracy for Arrhythmia dataset.

A.5 Internet Advertisements dataset

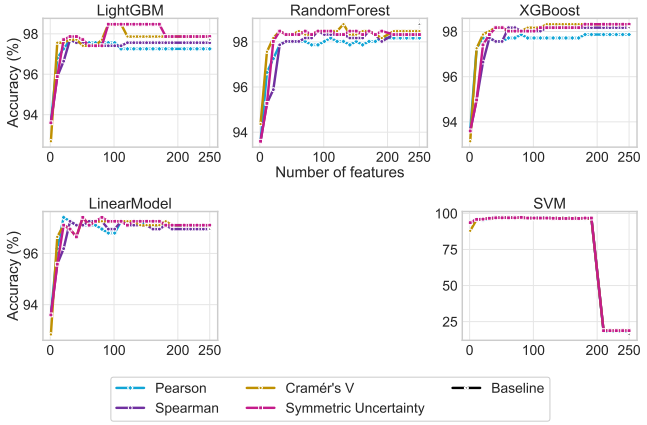

FIGURE 11: Accuracy for Internet Advertisements dataset.

A.6 Gisette dataset

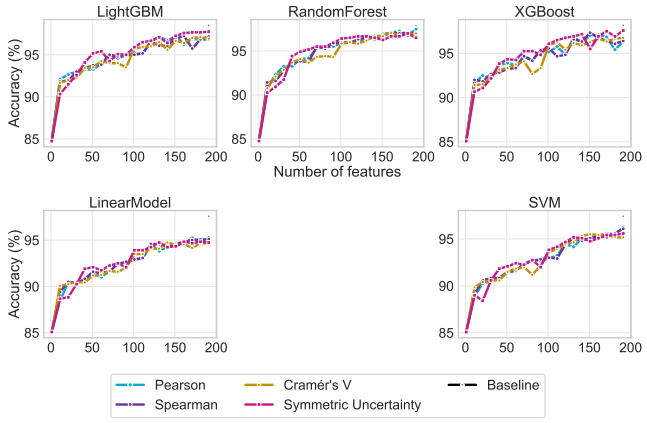

FIGURE 12: Accuracy for Gisette dataset.

A.7 Nursery dataset

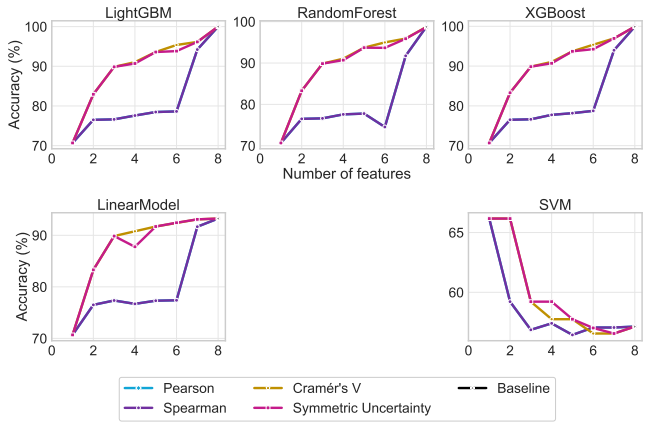

FIGURE 13: Accuracy for Nursery dataset.

A.8 Connect-4 dataset

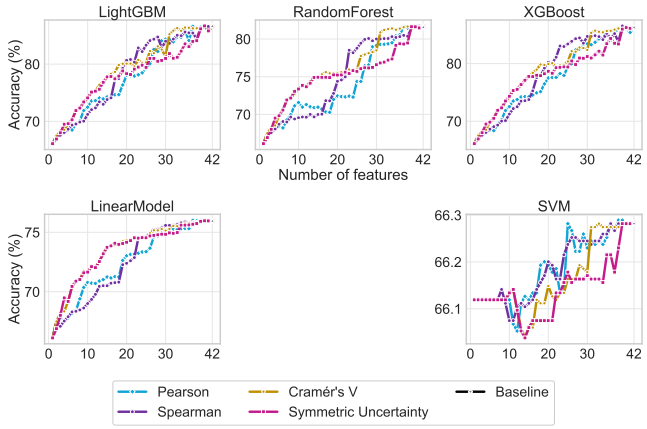

FIGURE 14: Accuracy for Connect-4 dataset.

A.9 Bike Sharing dataset

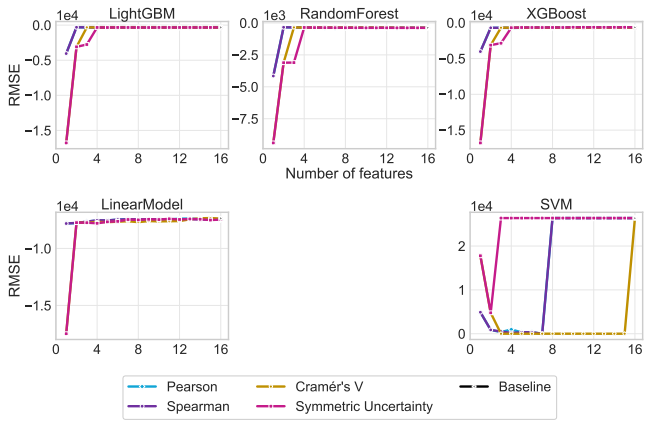

FIGURE 15: RSME for Bike Sharing dataset.

A.10 Housing Prices dataset

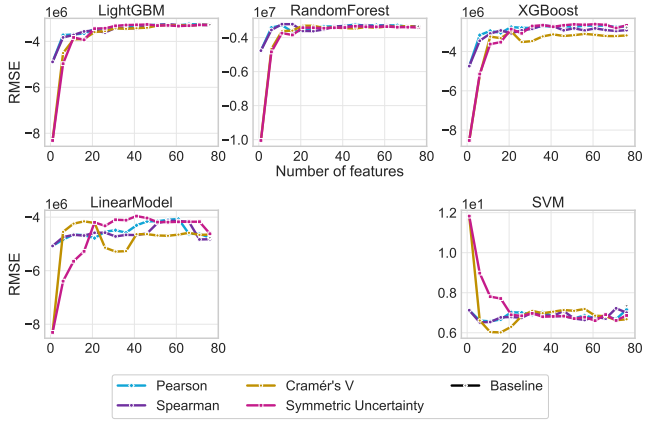

FIGURE 16: RSME for Housing Prices dataset.