algorithms

MDPI

Article

Attribute Relevance Score: A Novel Measure for Identifying Attribute Importance

Pablo Neirz †, Hector Allende † and Carolina Saavedra *,†

Departamento de Informática, Universidad Técnica Federico Santa María, Valparaíso 1680, Chile; pneira@usm.cl (P.N.); hector.allende@usm.cl (H.A.)

* Correspondence: carolina.saavedra@usm.cl

† These authors contributed equally to this work.

Abstract: This study introduces a novel measure for evaluating attribute relevance, specifically designed to accurately identify attributes that are intrinsically related to a phenomenon, while being sensitive to the asymmetry of those relationships and noise conditions. Traditional variable selection techniques, such as filter and wrapper methods, often fall short in capturing these complexities. Our methodology, grounded in decision trees but extendable to other machine learning models, was rigorously evaluated across various data scenarios. The results demonstrate that our measure effectively distinguishes relevant from irrelevant attributes and highlights how relevance is influenced by noise, providing a more nuanced understanding compared to established methods such as Pearson, Spearman, Kendall, MIC, MAS, MEV, GMIC, and Phi_k. This research underscores the importance of phenomenon-centric explainability, reproducibility, and robust attribute relevance evaluation in the development of predictive models. By enhancing both the interpretability and contextual accuracy of models, our approach not only supports more informed decision making but also contributes to a deeper understanding of the underlying mechanisms in diverse application domains, such as biomedical research, financial modeling, astronomy, and others.

Keywords: feature importance; feature selection; feature ranking; dependency measures; machine learning explainability; all-relevant problem

check for updates

Citation: Neirz, P.; Allende, H.; Saavedra, C. Attribute Relevance Score: A Novel Measure for Identifying Attribute Importance. *Algorithms* **2024**, *17*, 518. https://doi.org/10.3390/a17110518

Academic Editor: Muhammad Adnan Khan

Received: 27 September 2024
Revised: 4 November 2024
Accepted: 6 November 2024
Published: 9 November 2024

1. Introduction

In recent years, there has been significant advancement in methods for feature relevance and dependency assessment across a range of fields, from biomedical research to artificial intelligence and data engineering. Traditional approaches, such as those based on linear correlations, are well suited for capturing monotonic relationships but often fail in the presence of complex, nonlinear dependencies. Recent studies emphasize the limitations of these classical metrics and advocate for robust methods that adapt to diverse data characteristics. For instance, Khan et al. (2023) [1] highlight the challenges posed by high-dimensional data and underscore the need for feature selection methods that ensure computational efficiency without sacrificing relevance accuracy. Similarly, in biomedical contexts, methods that go beyond linear dependencies are essential for identifying complex, nonlinear relationships that could otherwise be overlooked by conventional techniques [2].

The rise of Big Data has intensified the need for scalable metrics capable of managing noisy and complex datasets [3]. Advanced methods, including metaheuristic algorithms, have demonstrated strong potential in unsupervised feature selection and clustering, offering robustness and adaptability for diverse data analysis needs. These methods are increasingly relevant for applications requiring reliable feature relevance assessment in complex, high-dimensional settings [4,5].

In the context of high-dimensional data analysis, identifying relevant attributes is critical for the effective implementation of machine learning methods and statistical analysis. This challenge can be approached from two primary perspectives: the "minimal optimal problem" and the "all-relevant problem". The former focuses on determining the smallest set of attributes that optimizes model performance, while the latter aims to identify all attributes that exhibit a relationship with the target variable, regardless of noise, redundancy, or indirect interactions [6]. The "all-relevant problem" is particularly valuable when seeking to understand the underlying mechanisms of the phenomenon under study, rather than merely building a predictive model. For example, in the medical field, when working with large datasets containing multiple attributes potentially related to cancer, the goal is to identify which of these attributes are truly relevant for prediction, even if they have been previously overlooked due to noise, complex interactions, or redundancy.

To address the "all-relevant problem", there is a growing preference for wrapper methods over classical statistical approaches. Wrapper methods combine a machine learning algorithm, a relevance criterion, and a search procedure that explores potential subsets of relevant features. These procedures are often heuristic and guided by the performance of an underlying model, frequently based on decision trees due to their interpretability and ability to handle complex variable interactions. However, wrapper algorithms have limitations, such as the absence of a measure that distinguishes the magnitude of relevance and the inability to clearly identify interactions among attributes or assess the likelihood of results being due to chance.

According to Kohavi and John (1997) [7], the relevance of a feature in a classification context should be defined in terms of its impact on the performance of an optimal Bayes classifier. A feature $\mathbf{x}$ is considered strongly relevant if its removal alone causes a deterioration in the performance of the optimal Bayes classifier. In other words, a strongly relevant feature is indispensable for maintaining the classifier's prediction accuracy. Conversely, a feature is considered weakly relevant if it is not strongly relevant, but there exists a subset of attributes S such that the performance of the Bayes classifier on S is worse than its performance on $S \cup \{\mathbf{x}\}$. This implies that while a weakly relevant feature may not always be critical, it can still contribute to improving the classifier's accuracy in specific contexts. A feature is deemed irrelevant if it is neither strongly nor weakly relevant; that is, its inclusion or exclusion does not affect the performance of the classifier. Understanding these distinctions is essential for determining the contribution and interaction of each feature within the dataset, which is a fundamental aspect of effective feature selection.

A notable issue with Kohavi and John's definition is that if we have two identical attributes $\mathbf{x_1}$ and $\mathbf{x_2}$, both strongly correlated with the target $\mathbf{y}$, removing one will not degrade the classifier's performance. Thus, according to their definition, $\mathbf{x_1}$ and $\mathbf{x_2}$ would not be considered strongly relevant, despite their strong relationship with the target. This limitation raises questions about whether Kohavi and John's definition intuitively captures relevance, particularly when an attribute that fully conveys information about the target could be categorized as weakly relevant [8].

Permutation feature importance, introduced by Breiman [9], is a fundamental technique for evaluating feature importance in supervised learning models. It involves randomly shuffling (i.e., randomly permuting) each feature's values and measuring the resulting performance drop, directly estimating each feature's relevance. A feature is relevant if its permutation significantly decreases the model's predictive accuracy, indicating its critical role. However, Strobl et al. [10] found that permutation-based measures can be biased, favoring high-cardinality features due to their higher chance of reducing impurity by chance. They proposed conditional permutation and tree-growing algorithm modifications to mitigate these biases, enhancing interpretability and validity in random forest models.

Stoppiglia et al. [11] introduced an innovative approach to the "all-relevant problem" in wrapper methods by incorporating a "test attribute" into the feature selection process. This randomly generated attribute, expected to be unrelated to the target variable, is added to the candidate attributes. The importance of other attributes is then evaluated based on their performance relative to the test attribute, determining relevance based on whether an attribute is ranked higher or lower than the test attribute. Unlike Kohavi and John's relevance definition, which may underestimate important attributes, Stoppiglia's approach considers an attribute relevant if its contribution exceeds that of a random attribute acting as a risk threshold, offering a more intuitive and practical perspective on relevance.

However, Stoppiglia's approach presents challenges, such as determining an acceptable risk threshold. In large datasets, more permissive risk thresholds may lead to the inclusion of irrelevant attributes. Additionally, heuristic-based selection may reduce the score of collinear attributes, and random attributes might, by chance, improve model performance, falsely indicating relevance. These issues can lead to the undervaluation or exclusion of potentially relevant attributes, complicating comprehensive identification of all relevant features. Such challenges highlight the limitations of Stoppiglia's method in effectively addressing the "all-relevant problem".

Stoppiglia's concept of outperforming test attributes has been adopted by algorithms designed to solve the "all-relevant problem", such as the Boruta algorithm [12]. These algorithms use admissible risk thresholds, addressing issues identified by Stoppiglia and reducing the arbitrariness introduced by random attributes.

Boruta, named after a Slavic forest god, was developed to address the "all-relevant problem" in multivariate classification. It combines concepts from Stoppiglia and Breiman. Stoppiglia's idea involved comparing predictor variables against a randomly generated benchmark attribute, while Breiman introduced permuting attribute values to assess feature importance. In Boruta, "shadow" attributes, created by permuting original attribute values, are used as benchmarks. These shadow attributes, unrelated to the target but retaining key statistical properties, ensure that both original and shadow attributes come from the same distribution. Boruta rigorously evaluates feature relevance by comparing original attributes against the highest-performing shadow attribute through statistical tests and multiple iterations of random forest. An attribute is deemed relevant if it consistently outperforms the best-performing shadow attribute, setting a high threshold and mitigating biases in feature selection processes [13]. While Boruta excels in multivariate contexts, in this work we focus on applying its principles in a univariate framework. In practical applications, relationships between variables are rarely symmetrical, and attributes with low cardinality often provide limited predictive power for those with high cardinality. For instance, an attribute with only three unique values is unlikely to effectively predict another with 100 unique values, whereas the attribute with higher cardinality could serve as a near-perfect predictor of the one with fewer values. This challenge, particularly evident in real-world scenarios, was highlighted by Wetschoreck (2020) [14].

1.1. Reproducibility

The challenges of reproducibility in machine learning, particularly in the context of high-dimensional data analysis, have been extensively examined. Goodman, Fanelli, and Ioannidis [15] categorize reproducibility into three essential types: methods reproducibility, which ensures that technical procedures can be precisely replicated; results reproducibility, which guarantees that reimplementations of a method produce statistically consistent outcomes; and inferential reproducibility, the most crucial of the three, which ensures that conclusions drawn from data are robust and generalizable across different experimental conditions. Bouthillier et al. [16] built upon this framework, emphasizing its importance in the context of machine learning, where these forms of reproducibility are essential for validating research findings and ensuring that they are not mere artifacts of specific experimental designs.

Moreover, factors such as data sampling, model initialization, and hyperparameter optimization introduce variability that can obscure true model performance. Addressing and modeling these sources of variability is essential for enhancing the reliability of machine learning benchmarks [17]. These principles are particularly related in the context of feature selection, where rigorous and reproducible methods are vital to ensure that the identified features genuinely contribute to the model's predictive power. Collectively, these approaches highlight the necessity of careful experimental design and robust evaluation criteria, which are indispensable for accurately determining the relevance of features in high-dimensional datasets.

A/B testing, a widely used technique in experimental design and digital marketing, shares several key principles with reproducibility-focused methodologies. A/B testing involves comparing two versions of a product or feature (commonly referred to as A and B versions) to determine which performs better according to a predefined metric [18]. This process relies on statistical hypothesis testing to assess whether observed differences in performance are statistically significant, thereby informing decisions about which version should be adopted [19].

1.2. Phenomenon-Centric Explainability

In the rapidly evolving field of machine learning, the concept of explainability has become increasingly critical, particularly as models grow in complexity and are applied in high-stakes domains. Techniques such as SHAP (SHapley Additive exPlanations) [20], LIME (Local Interpretable Model-agnostic Explanations) [21], and the frameworks proposed by Sokol [22] exemplify this approach by providing insights into how individual predictions are made. These methods aim to enhance user trust and enable validation of complex models by attributing importance scores to individual features, thereby making the model's decisions more transparent and comprehensible.

However, recent work by Lapuschkin et al. [23], in "Unmasking Clever Hans predictors and assessing what machines really learn", highlights the significant limitations of relying solely on traditional explainability techniques. Their study demonstrates that models can generate seemingly accurate predictions based on spurious correlations within the training data, reminiscent of the "Clever Hans" effect observed in psychology [24].

Similarly, machine learning models may appear to perform well by leveraging incidental patterns in the data patterns that do not genuinely reflect the underlying phenomena. These issues become particularly evident in studies where models achieve high predictive accuracy by leveraging spurious correlations in the training data [25,26]. Similarly, in computer vision, research has shown that models can misclassify images due to subtle perturbations that are imperceptible to the human eye [27]. Additionally, recent work by Wang et al. [28] has demonstrated how dataset biases can influence the conclusions drawn from SHAP values. These examples underscore the necessity of ensuring that model predictions are grounded in genuine, relevant patterns that accurately capture the true underlying processes, rather than being influenced by misleading artifacts.

In light of these findings, we propose a paradigm shift towards *phenomenon-centric explainability*, which emphasizes understanding the underlying phenomena that models are designed to capture. This approach is particularly critical in scientific research and in addressing complex challenges such as the "all-relevant problem". In these contexts, the primary objective is to gain deep insights into domain-specific processes, recognizing the importance of uncovering relationships—whether causal, dependent, or instrumental—rather than merely optimizing predictive accuracy. This perspective suggests that phenomenon-centric explainability represents a logical and necessary evolution, aligning with the traditional goals of scientific inquiry, such as exploring complex relationships and underlying mechanisms, without relying on interpretations that may be influenced by spurious correlations in the training data.

2. Proposal: Attribute Relevance Score

In this work, within the framework of the all-relevant problem, we introduce a novel univariate approach for evaluating attribute relevance, designed to tackle the inherent challenges of feature selection in high-dimensional datasets. Our approach builds upon the concept of shadow attributes, introduced by L. Breiman and later used by M.B. Kursa in the Boruta algorithm, to create a more robust measure of individual attribute importance. Unlike traditional methods, which may produce biased results due to specific data splits or data assumptions, our technique systematically evaluates each attribute's contribution by comparing it against its permuted versions. This process ensures that the relevance identified for each feature truly reflects underlying patterns in the data rather than artifacts from specific data splits or model configurations, enhancing both the reliability and interpretability of the results.

Consider an attribute matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, where n represents the number of samples and m denotes the number of attributes, along with its corresponding labels $\mathbf{y} \in \mathbb{R}^n$. To evaluate a specific attribute $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$, where each x_i represents an observation, we begin by partitioning $\mathbf{x}$ into training and testing subsets, denoted as $\mathbf{x}_{\text{train}}$ and $\mathbf{x}_{\text{test}}$, respectively, such that $\mathbf{x} = \mathbf{x}_{\text{train}} \cup \mathbf{x}_{\text{test}}$ and $\mathbf{x}_{\text{train}} \cap \mathbf{x}_{\text{test}} = \emptyset$.

A model $\mathcal{M}$ (e.g., a decision tree) is then trained using the training subset $\mathbf{x}_{\text{train}}$ and its corresponding labels $\mathbf{y}_{\text{train}}$. Subsequently, a shadow model, $\mathcal{M}_s$, is trained using a permutation of $\mathbf{x}_{\text{train}}$, denoted as $\mathbf{x}_{\text{s}}$, with the same labels $\mathbf{y}_{\text{train}}$.

We then consider a performance measure F that results from evaluating the output of a model on the test subset $\mathbf{x}_{\text{test}}$, compared to the correct labels $\mathbf{y}_{\text{test}}$. The choice of F should be made according to the specific problem at hand.

The core premise of our approach is that the performance measure F of the model $\mathcal{M}$, trained on the original attribute $\mathbf{x}_{\text{train}}$, should consistently be superior to that of the shadow model $\mathcal{M}_s$, trained on the permuted attribute $\mathbf{x}_{\text{s}}$, when both are evaluated on the same test subset $\mathbf{x}_{\text{test}}$ with labels $\mathbf{y}_{\text{test}}$.

$$\text{Opt}(F(\mathcal{M})) > \text{Opt}(F(\mathcal{M}_s)), \tag{1}$$

where $\text{Opt}(\cdot)$ represents an optimization criterion for F. Depending on the nature of F, this criterion could aim to maximize or minimize its value.

To minimize biases from specific test subsets, which could distort evaluations, we employ random sampling across multiple test subsets. This ensures a comprehensive and robust assessment of attribute relevance, capturing the magnitude of the observed effects and articulating them with a level of detail that supports precise and actionable conclusions.

We propose Equation (2) to calculate the attribute relevance score (ARS) for performance metrics F, where "the closer to 1, the better", such as those commonly used in classification tasks (e.g., accuracy, F1 score, recall, and precision).

$$\text{ARS} = \max\left(0, \frac{\overline{F}(\mathcal{M}) - (\overline{F}(\mathcal{M}_s) + \epsilon)}{1 - (\overline{F}(\mathcal{M}_s) + \epsilon)}\right), \tag{2}$$

where $\overline{F}(\mathcal{M})$ represents the median performance of model $\mathcal{M}$ across multiple test subsets, $\overline{F}(\mathcal{M}_s)$ is the median performance of the shadow model over the same subsets, and ϵ is a small positive constant indicating the minimum acceptable effect size threshold. The ARS algorithm for classification is detailed in Algorithm 1.

The primary objective of our measure is to answer two fundamental questions:

1. Is there a significant dependency between an attribute $\mathbf{x}$ and another attribute or target variable $\mathbf{y}$, independent of factors such as cardinality, specific distributions, or artifacts from the data or model configuration? A score of $\text{ARS}(\mathbf{x}) > 0$ would indicate such a genuine dependency.
2. How effectively can $\mathbf{y}$ be approximated based on $\mathbf{x}$ using a specific error measure and a chosen algorithm? The attribute relevance score (ARS) quantifies the practical capability of $\mathbf{x}$ to predict or approximate $\mathbf{y}$, capturing not only the existence of a statistical dependency but also the magnitude of its predictive power in the context of the selected model and performance metric. This ensures that any detected relevance corresponds to a genuine and meaningful predictive capacity, rather than an artifact of the statistical properties of $\mathbf{x}$ or models based on chance.

To illustrate the behavior of the ARS and build intuition, consider the simple equation $\mathbf{y} = \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3$, where $\mathbf{x}_i$ are three independent, nonredundant predictor attributes within the same range. Suppose that, for the first attribute, we obtain $\overline{F}(\mathcal{M}_s) + \epsilon = 0.5$ and $\overline{F}(\mathcal{M}) = 0.75$; for the second attribute, $\overline{F}(\mathcal{M}_s) + \epsilon = 0.5$ and $\overline{F}(\mathcal{M}) = 0.65$; and for the third attribute, $\overline{F}(\mathcal{M}_s) + \epsilon = 0.5$ and $\overline{F}(\mathcal{M}) = 0.6$. Calculating the attribute relevance score (ARS) for each attribute yields scores of 0.5, 0.3, and 0.2, respectively. Notably, in the ideal case, the sum of these scores is 1.0, which suggests that these three independent variables, collectively, have the capacity to fully represent the target variable. This example illustrates how the ARS quantifies the cumulative relevance of independent attributes, capturing the extent to which a set of predictors can collectively account for the target variable.

Algorithm 1 Attribute relevance score for classification.

```
Set up model: (e.g., DecisionTreeClassifier)
Set up performance metric $F$: (e.g., $f_1$-score)
Set up acceptable error margin: $\epsilon$
Set significance level: $\alpha$ (e.g., $\alpha = 0.05$)
Initialize lists to store performance metrics of the original and shadow models
for iteration from 1 to $N$ do
    Split the data into training and testing sets: $\mathbf{X}_{\text{train}}, \mathbf{X}_{\text{test}}, \mathbf{y}_{\text{train}}, \mathbf{y}_{\text{test}}$
    Train the original model $\mathcal{M}$ on $\mathbf{X}_{\text{train}}$ and $\mathbf{y}_{\text{train}}$
    Generate shadow attributes by permuting the original attributes to obtain $\mathbf{X}_{\text{shadow}}$
    Train the shadow model $\mathcal{M}_s$ on $\mathbf{X}_{\text{shadow}}$ and $\mathbf{y}_{\text{train}}$
    Evaluate and store the performance metric $F$ of both models $\mathcal{M}$ and $\mathcal{M}_s$ on $\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}}$
end for
Compute the median performances $\overline{F}(\mathcal{M})$ and $\overline{F}(\mathcal{M}_s)$
Conduct statistical test: Perform a test of means (e.g., a $t$-test) on the performance
    metrics collected from $\mathcal{M}$ and $\mathcal{M}_s$ to obtain a $p$-value
if $p$-value $< \alpha$ then
    Compute the Attribute Relevance Score (ARS) using the following formula:
        $\text{ARS} = \max\left(0, \frac{\overline{F}(\mathcal{M}) - (\overline{F}(\mathcal{M}_s) + \epsilon)}{1 - (\overline{F}(\mathcal{M}_s) + \epsilon)}\right)$
else
    Set ARS $= 0$
end if
return ARS and the $p$-value from the statistical test
```

In Step 13 of Algorithm 1, we conduct a statistical test (e.g., a t-test) to compare the performance metrics of the original model $\mathcal{M}$ and the shadow model $\mathcal{M}_s$. This test provides a p-value that quantifies the likelihood that the observed difference in performance occurred by chance. If the p-value is less than the significance level α, we consider the difference in performance to be statistically significant. In this case, we proceed to compute the attribute relevance score (ARS) using the formula provided. If the p-value is greater than or equal to α, we conclude that there is no statistically significant difference between the models. Therefore, we set ARS $= 0$, indicating that the attribute is not considered relevant.

By incorporating this statistical test, we embed statistical significance into our attribute relevance score. However, it is important to recognize that a low p-value is a necessary but not sufficient condition for establishing attribute relevance. Additional factors such as effect size, consistency across iterations, and practical significance should also be considered to ensure that the identified relevance corresponds to meaningful patterns rather than random fluctuations. This holistic approach emphasizes the importance of not relying solely on statistical significance but also assessing the practical implications of the findings to confirm that they align with genuine data relationships.

This scoring system not only facilitates the identification of relevant attributes but also significantly enhances the reliability and validity of the constructed predictive models. By incorporating multiple test subsets and focusing on statistical robustness, our approach ensures that the results are both generalizable and reproducible across diverse experimental conditions, thereby contributing to the advancement of more robust and credible data analysis in high-dimensional settings.

Moreover, with slight modifications, Algorithm 1 can also be adapted for performance metrics F where "closer to zero is better" such as mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE), which are more common in regression problems, illustrated in Algorithm 2.

Algorithm 2 Attribute relevance score for regression.

```
Set up model: (e.g., DecisionTreeRegressor)
Set up performance metric $F$: (e.g., Mean Absolute Error)
Set up acceptable error margin: $\epsilon$
Set significance level: $\alpha$ (e.g., $\alpha = 0.05$)
Initialize lists to store performance metrics of the original and shadow models
for iteration from 1 to $N$ do
    Split the data into training and testing sets: $\mathbf{X}_{\text{train}}, \mathbf{X}_{\text{test}}, \mathbf{y}_{\text{train}}, \mathbf{y}_{\text{test}}$
    Train the original model $\mathcal{M}$ on $\mathbf{X}_{\text{train}}$ and $\mathbf{y}_{\text{train}}$
    Generate shadow attributes by permuting the original attributes to obtain $\mathbf{X}_{\text{shadow}}$
    Train the shadow model $\mathcal{M}_s$ on $\mathbf{X}_{\text{shadow}}$ and $\mathbf{y}_{\text{train}}$
    Evaluate and store the performance metric $F$ (e.g., MAE) of both models $\mathcal{M}$ and $\mathcal{M}_s$
    on $\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}}$
end for
Compute the median performances $\overline{F}(\mathcal{M})$ and $\overline{F}(\mathcal{M}_s)$
Conduct statistical test: Perform a test of means (e.g., a $t$-test) on the performance
metrics collected from $\mathcal{M}$ and $\mathcal{M}_s$ to obtain a $p$-value
if $p$-value $< \alpha$ then
    Compute the Attribute Relevance Score (ARS) using the following formula:

        $\text{ARS} = \max\left(0, \frac{\left(\overline{F}(\mathcal{M}_s) - \epsilon\right) - \overline{F}(\mathcal{M})}{\left(\overline{F}(\mathcal{M}_s) - \epsilon\right) - 0}\right)$

else
    Set ARS = 0
end if
return ARS and the $p$-value from the statistical test
```

By using models that directly accept categorical variables, such as CatBoost [29], and selecting appropriate performance metrics for multilabel and multioutput problems, the ARS method can be seamlessly extended to handle a wide variety of data types and problem settings. This flexibility ensures that the method remains robust and effective across different domains, making it a valuable tool for feature selection and attribute relevance evaluation in diverse applications.

3. Materials and Methods

3.1. Pearson's Correlation

Among the most widely used filtering methods for determining attribute relevance, we first encounter Pearson's correlation, denoted by r_{xy} (see Equation (3)).

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

where x_i and y_i are the individual observations of the variables, $\bar{x}$ and $\bar{y}$ denote the means, and n denotes the total number of observations.

This technique identifies attributes that exhibit a high linear correlation with the target variable, potentially suggesting attribute relevance. Pearson's correlation values range between -1 and 1. The closer the coefficient is to 1 or -1, the stronger the linear correlation.

A coefficient of 1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable increases in constant proportion. Conversely, a coefficient of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other decreases in constant proportion. A coefficient of 0 indicates the absence of a linear relationship between the two variables, implying no linear correlation.

It is important to note the limitations of Pearson's correlation in detecting nonlinear relationships between variables, as well as its sensitivity to outliers, which can lead to biased or inaccurate interpretations in the attribute selection process.

3.2. Spearman's Correlation

Another classical option is Spearman's correlation (Equation (4)), denoted by ρ. Spearman's correlation, or Spearman's rank correlation coefficient, is a statistical measure that evaluates the relationship between two variables. Unlike Pearson's correlation, which assesses a linear relationship, Spearman's coefficient examines monotonic relationships, meaning that as one variable increases or decreases, the other also tends to increase or decrease, without requiring a strictly linear relationship.

For its calculation, Spearman's correlation relies on the ranks of the data rather than the actual values. That is, each value in a variable is replaced by its position or rank in magnitude within its respective dataset.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{4}$$

where d_i denotes the rank differences between the observations. If the variables x_i and y_i have ranks r_{xi} and r_{yi}, respectively, then $d_i = r_{xi} - r_{yi}$, and n denotes the total number of observations.

Similar to Pearson's correlation, Spearman's correlation values also range between -1 and 1, but their interpretation is slightly different. A coefficient of 1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable also tends to increase in a monotonic relationship. Similarly, a coefficient of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other variable tends to decrease in a monotonic relationship. A coefficient of 0 indicates the absence of a monotonic correlation between the two variables.

3.3. Kendall's Correlation

Kendall's correlation, or Kendall's rank correlation coefficient, denoted by τ (Equation (5)), is a statistical measure that evaluates the association between two ordinal variables. It is based on the concept of concordance and discordance between pairs of observations. A pair of observations is concordant if the relative order between the two variables is the same (i.e., if one observation is higher in both variables or lower in both). A pair is discordant if the relative order is different. Similar to Spearman's correlation, τ is useful for measuring monotonic relationships, and its range of values and interpretation are equivalent, ranging between -1 and 1. A value of 1 indicates a perfectly positive monotonic relationship, and a value of -1 indicates a perfectly negative monotonic relationship. The main advantage of τ over Spearman's correlation is its reduced sensitivity to outliers.

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j) sgn(y_i - y_j) \tag{5}$$

where $sgn(x_i - x_j)$ and $sgn(y_i - y_j)$ are the sign functions that return 1 if $x_i > x_j$ or $y_i > y_j$, -1 if $x_i < x_j$ or $y_i < y_j$, and 0 if $x_i = x_j$ or $y_i = y_j$. The sum counts the concordant and discordant pairs of observations. A pair of observations (i, j) is concordant if both $(x_i > x_j$ and $y_i > y_j)$ or $(x_i < x_j$ and $y_i < y_j)$; otherwise, the pair is discordant.

3.4. Phi$_k$ Correlation

ϕ_k is a correlation coefficient developed by Baak et al. (2020) [30] to measure the association between two variables, including categorical and mixed variables (categorical and continuous), without the need for prior transformations. The aim of this coefficient is to provide a robust and more informative measure than traditional coefficients previously mentioned. Another difference is that its values range between 0 and 1, where 0 indicates no correlation and 1 indicates a perfect correlation. Since its calculation involves several rules, it does not have a closed-form formula:

1. Variable transformation: Categorical variables are handled using encoding, while numerical variables must be discretized if necessary. It is suggested to use 10 uniformly spaced bins per variable.
2. Contingency table: A contingency table is created with the transformed variables, where each pair contains N observations, r rows, and k columns.
3. Pearson's chi-square test (χ^2): The chi-square statistic is calculated for the contingency table, as described in Equation (6), to evaluate the independence between variables.
4. Calculation of ϕ_k: The value of χ^2 is interpreted as if it were derived from a bivariate normal distribution without statistical fluctuations, using Equations (7) and (8).
    - If $\chi^2 < \chi^2_{ped}$, then set ρ to zero.
    - Else, with fixed N, r, k, invert the $\chi^2_{b.n.}$ function and numerically solve for ρ in the range [0, 1].
    - The solution for ρ defines the correlation coefficient ϕ_k

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{6}$$

where O_{ij} represents the observed frequency in cell ij of the contingency table, E_{ij} represents the expected frequency in cell ij under the null hypothesis of independence of the variables, r is the number of rows in the contingency table, and k is the number of columns in the contingency table.

$$\chi^2_{b.n.}(\rho, N, r, k) = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(F_{ij}(\rho = \rho) - F_{ij}(\rho = 0))^2}{F_{ij}(\rho = 0)} \tag{7}$$

$$\chi^2_{b.n.}(\rho, N, r, k) = \chi^2_{ped} + \left(\frac{\chi^2_{max}(N, r, k) - \chi^2_{ped}}{\chi^2_{b.n.}(1, N, r, k)}\right) \cdot \chi^2_{b.n.}(\rho, N, r, k) \tag{8}$$

3.5. Mutual Information

From a perspective grounded in information theory, mutual information (MI) offers a distinct approach to understanding correlations by measuring how much information the knowledge of one variable provides about another. In other words, mutual information quantifies the reduction of uncertainty about one variable given information about the other variable (Equation (9)). The minimum value of MI is 0, indicating that two variables are completely independent, or that the knowledge of one variable provides no information about the other. It does not have a fixed upper limit; the maximum value can be as large as allowed by the entropy of each of the variables.

$$I(\mathbf{x}; \mathbf{y}) = \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)\log\left(\frac{P(x_i, y_j)}{P(x_i)P(y_j)}\right) \tag{9}$$

where $P(x_i, y_j)$ is the joint probability that $\mathbf{x}$ takes the value x_i and $\mathbf{y}$ takes the value y_j, while $P(x_i)$ and $P(y_j)$ are the marginal probabilities of $\mathbf{x}$ and $\mathbf{y}$, respectively. Mutual information is generally calculated over discrete probability distributions. Therefore, if the numerical variables are continuous, they may need to be discretized into intervals or categories. Since discretization can affect the accuracy of the calculation and lead to the loss of important information, in this work, we will use the scikit-learn implementation based on the work of Ross BC (2014) [31], which avoids discretization by using an entropy estimation based on K-nearest neighbor distances.

3.6. Maximal Information-Based Nonparametric Exploration Suite

The Maximal Information-Based Nonparametric Exploration (MINE) suite is a collection of statistical tools designed to detect a wide range of relationships between variables in large datasets. Unlike traditional correlation measures that often focus on linear associations, MINE is tailored to capture both linear and nonlinear dependencies without making strong assumptions about the underlying data distribution. Developed by Reshef et al. (2011) [32], the MINE suite includes several measures, such as the maximal information coefficient (MIC), maximal asymmetry score (MAS), maximal edge value (MEV), and generalized mean information coefficient (GMIC), each designed to identify different types of associations in a data-driven manner.

One of the primary measures in the MINE suite, the maximal information coefficient (MIC), represents a significant advancement in detecting complex associations within data. MIC is specifically designed to identify both linear and nonlinear dependencies between two variables, making it particularly valuable in scenarios where relationships are complex or nonlinear and difficult to capture with conventional methods.

The methodology behind MIC involves optimizing the partitioning of the data space to maximize the normalized mutual information between variables (Equation (10)). It systematically explores all possible ways of dividing the data into grids and evaluates the strength of the relationship within each partition, selecting the configuration that maximizes normalized mutual information:

$$MIC(\mathbf{x};\mathbf{y}) = \max_{G\in \mathrm{Grids}(\mathbf{x},\mathbf{y})} \left\{ \frac{I(\mathbf{x};\mathbf{y} \mid G)}{\log(\min\{|\mathbf{x}|,|\mathbf{y}|\})} \right\} \tag{10}$$

where $I(\mathbf{x};\mathbf{y} \mid G)$ is the mutual information conditioned on a specific grid partition G, and $|\mathbf{x}|$ and $|\mathbf{x}|$ denote the cardinalities of variables $\mathbf{x}$ and $\mathbf{y}$. The normalization ensures that MIC values range between 0 and 1, where 1 indicates a perfect association.

MIC has been recognized for its ability to uncover complex patterns in data, and software tools like "minepy" version 1.2.6, created by Davide Albanese under a GNU General Public License (GPL). Refs. [33,34] have made it accessible for large-scale analyses. However, the exhaustive evaluation of multiple data partitions can be computationally intensive. To address this, Reshef et al. introduced optimizations to improve computational efficiency, especially in high-dimensional datasets.

The MINE suite also includes the maximal asymmetry score (MAS), which extends the capability of detecting asymmetrical relationships that symmetric measures often overlook. MAS quantifies the degree of asymmetry in the relationship between two variables, offering a more nuanced perspective:

$$MAS(\mathbf{x};\mathbf{y}) = \max_{G\in \mathrm{Grids}(\mathbf{x},\mathbf{y})} \left\{ \frac{|(\mathbf{x};\mathbf{y} \mid G) - I(Y;X \mid G)|}{\log(\min\{|\mathbf{x}|,|\mathbf{y}|\})} \right\} \tag{11}$$

where $(\mathbf{x};\mathbf{y} \mid G)$ and $I(Y;X \mid G)$ are mutual information measures calculated in opposite directions. MAS values between 0 and 1 reflect the extent of asymmetry in the association.

Additionally, the maximal edge value (MEV) focuses on identifying the strongest dependencies within subsets of data, which may not be uniformly strong across all points but exhibit significant relationships in specific regions:

$$MEV(\mathbf{x};\mathbf{y}) = \max_{G\in \mathrm{Grids}(\mathbf{x},\mathbf{y})} \{\mathrm{EdgeValue}((\mathbf{x};\mathbf{y} \mid G))\} \tag{12}$$

where $\mathrm{EdgeValue}((\mathbf{x};\mathbf{y} \mid G))$ is the mutual information at the boundaries of a grid partition G, highlighting local dependencies that might be averaged out in global measures like MIC.

3.7. Generalized Mean Information Coefficient

The generalized mean information coefficient (GMIC), proposed by Luedtke and Tran [35], extends the maximal information coefficient (MIC) by introducing a more flexible framework for detecting associations, particularly in finite sample settings. The GMIC enhances the sensitivity of MIC by incorporating a generalized mean approach, which allows for a more nuanced detection of statistical dependencies between variables.

The core innovation of GMIC lies in its use of a maximal characteristic matrix, defined as follows:

$$C^*(\mathbf{x},\mathbf{y})_{i,j} = \max_{kl \leq ij}\{C(\mathbf{x},\mathbf{y})_{kl}\}, \tag{13}$$

where $i, j \geq 2$. This matrix captures the highest normalized mutual information that can be obtained with grid sizes up to ij, providing a robust measure of the strongest possible association achievable for each grid resolution.

GMIC itself is calculated using a generalized mean, which is defined as follows:

$$GMIC_p(\mathbf{x},\mathbf{y}) = \left(\frac{1}{Z}\sum_{ij \leq B(n)} (C^*(\mathbf{x},\mathbf{y})_{i,j})^p\right)^{1/p}, \tag{14}$$

where p is a tuning parameter that adjusts the measure's sensitivity to different aspects of the data, and $Z = \#\{(i,j) : ij \leq B(n)\}$ denotes the number of grid sizes considered. By varying p, GMIC can emphasize different features of the data's structure; for example, when $p \to \infty$, it approaches the original MIC, while for $p \to -\infty$, GMIC converges to the minimal value from the maximal characteristic matrix

The introduction of the tuning parameter p provides flexibility, allowing GMIC to detect a broader range of complex associations across various data types. In simulations, GMIC has demonstrated improved power over MIC for detecting associations, particularly in finite samples. It retains desirable properties, such as convergence to zero for independent variables as the sample size increases, and convergence to one for noiseless, monotonic relationships.

Overall, GMIC enhances the capabilities of the MINE suite by offering a more adaptive and sensitive approach to uncovering diverse relationships in complex datasets, making it a valuable tool for exploratory data analysis in high-dimensional contexts.

4. Simulation Results

4.1. Experiment #1: Computer-Generated Data

This study utilizes a synthetically generated dataset to evaluate the performance of the proposed methods under various types of bivariate relationships, including both linear and nonlinear dependencies. The data were created using multiple transformations and multivariate normal distributions to simulate distinct scenarios that reflect complexities commonly observed in real-world datasets. The goal is to assess the robustness and effectiveness of the proposed attribute relevance score (ARS) (using decision trees as the base model), and other established dependency measures across diverse data configurations.

- **Multivariate normal distributions**: Seven datasets, each containing 1000 samples, were generated using multivariate normal distributions with correlation coefficients ranging from −1.0 to 1.0. These datasets represent different degrees of linear relationships between variables, labeled from (A) to (G) in Figure 1.
- **Rotated normal distributions**: To examine the effect of rotational transformations on normally distributed data, seven datasets of 1000 samples each were created by applying specific rotation angles (ranging from 0 to $\pi/2$) to a bivariate normal distribution with equal variance and high covariance. These are labeled from (H) to (N) in Figure 1. A special case, labeled (K) with unique horizontal distribution,

was included to demonstrate a scenario with zero correlation. These transformations model linear relationships oriented in diverse directions.

- **Other nonlinear and complex distributions**: This category includes seven distinct datasets representing various nonlinear, asymmetric, and complex dependencies. They are labeled from (O) to (U) in Figure 1. Key examples are the following:
  - (O): Displays a quadratic dependency with added noise.
  - (P) and (Q): Represent sequential rotations applied to uniformly distributed data.
  - (R) and (S): Exhibit parabolic and partially symmetric patterns.
  - (T): Simulates a trigonometric relationship combining sinusoidal and cosinusoidal transformations.
  - (U): Comprises a clustered structure formed by four multivariate normal distributions centered at distinct locations, resembling spatially separated clusters.

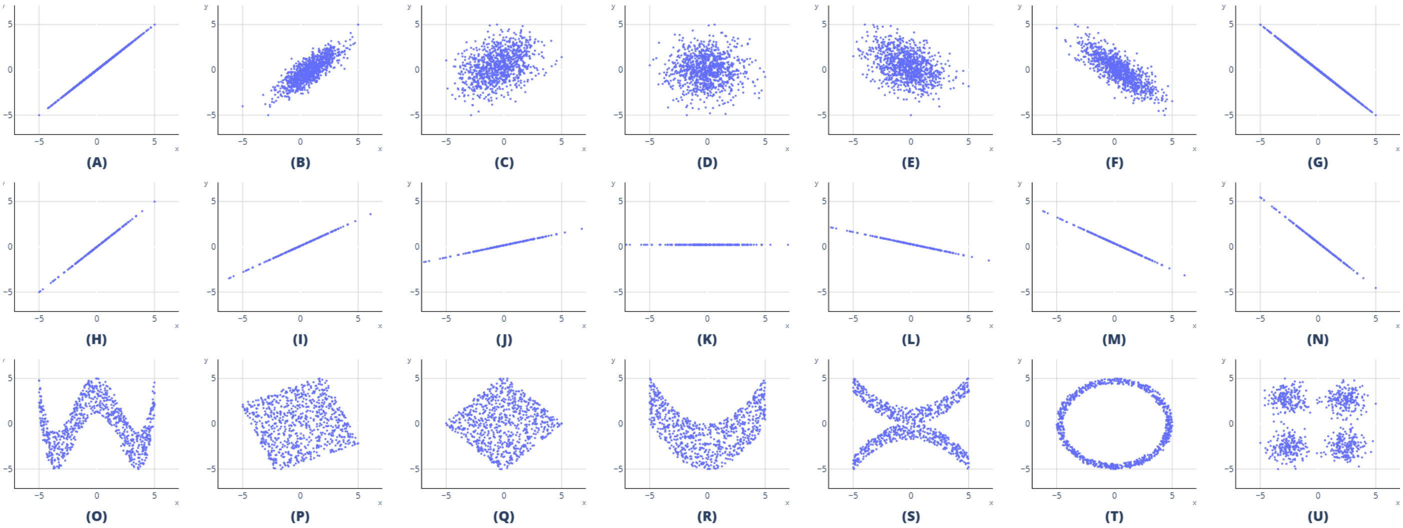

Figure 1. Scatter plots illustrating various bivariate relationships generated. Each subplot represents a different synthetic dataset: (**A**–**G**) are datasets generated from multivariate normal distributions with varying correlations; (**H**–**N**) are datasets generated from rotated normal distributions, illustrating different linear relationships; (**O**–**U**) represent other complex, nonlinear patterns.

To ensure comparability across subsets, all data were scaled using MinMax scaling within the range of -5 to 5, preserving the inherent structural relationships while normalizing their ranges, as illustrated in Figure 1.

Results #1

Table 1 presents the performance of various dependency measures applied to the synthetically generated datasets described in the previous section. Each dataset was designed to capture different types of bivariate relationships, including both linear and nonlinear dependencies, to assess the robustness and effectiveness of the proposed attribute relevance score (ARS) and other correlation measures.

The results demonstrate that for datasets based on **multivariate normal distributions (A–G)**, the ARS and other linear measures such as Pearson and Spearman coefficients perform as expected when the relationship between variables is linear. In particular, datasets (A) and (G), which exhibit perfect linear correlation, yield maximum values (1.000) for most measures, including Pearson, Spearman, Kendall, and Phi_k, with corresponding high ARS scores of 0.996. Similarly, datasets with moderate to low correlation levels (e.g., (B) and (F)) show a consistent decline in these measures, with ARS scores reflecting these variations (0.429 and 0.394, respectively).

Table 1. Comparison of various dependency measures across different datasets. The columns represent different metrics: r_{xy} is the Pearson correlation coefficient, ρ is the Spearman rank correlation coefficient, and τ is Kendall's tau coefficient. These metrics assess the linear and monotonic relationships between variables, while other columns represent measures of mutual information (MI), maximal information coefficient (MIC), maximal asymmetry score (MAS), maximum edge-value (MEV), generalized mean information coefficient (GMIC), and the proposed attribute relevance score (ARS).

Set	r_{xy}	ρ	τ	ϕ_k	MI	MIC	MAS	MEV	GMIC	ARS Proposed
(A)	1.000	1.000	1.000	1.000	5.984	1.000	0.000	1.000	1.000	0.996
(B)	0.805	0.783	0.587	0.859	0.575	0.519	0.030	0.519	0.461	0.429
(C)	0.378	0.378	0.258	0.373	0.081	0.200	0.017	0.200	0.148	0.049
(D)	0.009	0.016	0.010	0.157	0.007	0.126	0.007	0.126	0.054	0.000
(E)	0.394	0.383	0.261	0.407	0.054	0.215	0.015	0.215	0.158	0.089
(F)	0.809	0.800	0.606	0.817	0.508	0.512	0.015	0.512	0.465	0.394
(G)	1.000	1.000	1.000	1.000	5.984	1.000	0.000	1.000	1.000	0.996
(H)	1.000	1.000	1.000	1.000	4.373	1.000	0.000	1.000	1.000	0.980
(I)	1.000	1.000	1.000	1.000	4.373	1.000	0.000	1.000	1.000	0.980
(J)	1.000	1.000	1.000	1.000	4.373	1.000	0.000	1.000	1.000	0.980
(K)	0.606	0.624	0.483	0.784	0.697	0.498	0.038	0.498	0.412	0.000
(L)	1.000	1.000	1.000	1.000	4.373	1.000	0.000	1.000	1.000	0.980
(M)	1.000	1.000	1.000	1.000	4.373	1.000	0.000	1.000	1.000	0.980
(N)	1.000	1.000	1.000	1.000	4.373	1.000	0.000	1.000	1.000	0.980
(O)	0.052	0.046	0.033	0.732	0.733	0.773	0.626	0.773	0.670	0.477
(P)	0.014	0.014	0.006	0.525	0.240	0.195	0.026	0.195	0.136	0.046
(Q)	0.036	0.044	0.023	0.526	0.315	0.163	0.025	0.138	0.121	0.024
(R)	0.028	0.028	0.018	0.688	0.534	0.416	0.245	0.416	0.382	0.297
(S)	0.020	0.021	0.018	0.783	0.902	0.450	0.037	0.262	0.208	0.135
(T)	0.025	0.020	0.003	0.836	1.421	0.566	0.079	0.376	0.443	0.215
(U)	0.013	0.025	0.017	0.000	0.012	0.134	0.009	0.134	0.065	0.000

For the **rotated normal distributions (H–N)**, the ARS effectively captures the directional dependencies across different orientations. Datasets (H) to (J) and (L) to (N) maintain high scores across all measures (1.000), including ARS values around 0.980, indicating the robustness of the ARS in identifying dependencies even when the data are rotated. Notably, dataset (K), which presents no linear or clear directional correlation, shows an ARS value of 0.000, while other measures still return nonzero values. This highlights the ARS's unique sensitivity to the absence of any meaningful relationship, effectively distinguishing it from traditional dependency measures.

The results for the **other nonlinear and complex distributions (O–U)** highlight the limitations of traditional linear measures in capturing intricate dependencies. For instance, in dataset (O), which exhibits a quadratic dependency, the ARS score (0.477) is notably higher than those of other measures, such as Pearson (0.052) and Spearman (0.046), demonstrating ARS's ability to detect nonlinear patterns. Similarly, datasets (R) and (T), featuring parabolic and trigonometric relationships, yield ARS values (0.297 and 0.215, respectively) that surpass those of the linear measures, suggesting that ARS offers a more refined assessment of these complex dependencies. Notably, datasets (D) and (K), which do not exhibit any significant linear or nonlinear relationships, have ARS scores of 0.000, whereas other measures still return nonzero values. This distinction underscores the ARS's unique sensitivity to the absence of meaningful relationships, effectively setting it apart from traditional dependency measures. However, for dataset (U), which represents a clustered structure formed by four multivariate normal distributions centered at distinct locations, it may be arguable whether any of the measures presented, including ARS, effectively capture the inherent clustering pattern. In this case, all measures, including ARS, tend to produce the lowest values observed, indicating the potential limitations of these metrics in identifying complex clustering structures.

Overall, the results indicate that ARS is a robust metric for capturing both linear and nonlinear dependencies across diverse types of relationships. In comparison to traditional measures such as Pearson, Spearman, and Kendall—which are effective at identifying monotonic dependencies but limited with more complex nonlinear relationships—the ARS consistently outperformed these classical methods, particularly on datasets (O) to (U). By using ARS alongside conventional metrics, a more comprehensive framework is established for evaluating bivariate dependencies, fulfilling the objectives of this study and providing valuable insights across a broad range of real-world scenarios.

4.2. Experiment #2: Benchmark Dataset

This section aims to compare our proposed method against established predictive approaches using a synthetic dataset that has been used in prior research [12,36]. The selected benchmark dataset is particularly suitable for this evaluation as it incorporates both informative and noninformative variables, with varying degrees of noise, thus providing a comprehensive framework for assessing model performance under realistic conditions.

The dataset is generated using a set of equations designed to create correlated predictor variables that combine both signal and noise components, allowing for a critical evaluation of each method's ability to discern relevant features in the presence of spurious or irrelevant variables. The data generation process is defined as follows:

$$\mathbf{y} = 0.25\exp(4\mathbf{x}_1) + \frac{4}{1+\exp(-20(\mathbf{x}_2 - 0.5))} + 3\mathbf{x}_3 + \eta, \tag{15}$$

where $\mathbf{y}$ is the output variable, η is normally distributed with mean 0 and standard deviation 0.2, and the predictor variables $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ are uniformly distributed between 0 and 1 ($U(0,1)$).

Each relationship in Equation (15) contributes uniquely to $\mathbf{y}$, reflecting the distinct characteristics of their functional forms. The exponential dependency on $\mathbf{x}_1$ acts as the primary driver of variability in $\mathbf{y}$. Due to its exponential nature, even small increases in $\mathbf{x}_1$ can produce substantial shifts in $\mathbf{y}$, especially when $\mathbf{x}_1$ reaches higher values within its range. This makes $\mathbf{x}_1$ a dominant factor influencing the output variable. In contrast, the logistic dependency on $\mathbf{x}_2$ provides a significant but bounded effect on $\mathbf{y}$. The logistic function has its strongest impact around the midpoint ($\mathbf{x}_2 = 0.5$), where it introduces notable changes in $\mathbf{y}$. However, this effect stabilizes outside of that range, introducing a controlled nonlinear influence that is considerable near the midpoint but limited in other regions. The linear dependency on $\mathbf{x}_3$ has a more modest contribution to $\mathbf{y}$. Although this linear relationship ensures proportional adjustments in $\mathbf{y}$ with changes in $\mathbf{x}_3$, its effect can be easily overshadowed by the more pronounced variations introduced by $\mathbf{x}_1$ and $\mathbf{x}_2$, as well as by the error term η. This suggests that, in the presence of more dominant influences and noise, the contribution of $\mathbf{x}_3$ may be less perceptible and could, therefore, receive a lower score in certain attribute importance measures.

4.2.1. Results #2.1

Examining the results without the noise term η offers additional insight into the ARS measure's effectiveness in isolating informative dependencies. To highlight this, we present the ARS behavior derived from Equation (15) after excluding η. This "noise-free" version of ARS is evaluated using three distinct base models: decision trees, linear regression, and k-nearest neighbors (k = 5). By removing the noise component, we aim to reveal how well each model captures the inherent relationships between $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ and the target variable $\mathbf{y}$, thereby underscoring the relative strengths of each approach in detecting true signal (see Figure 2).

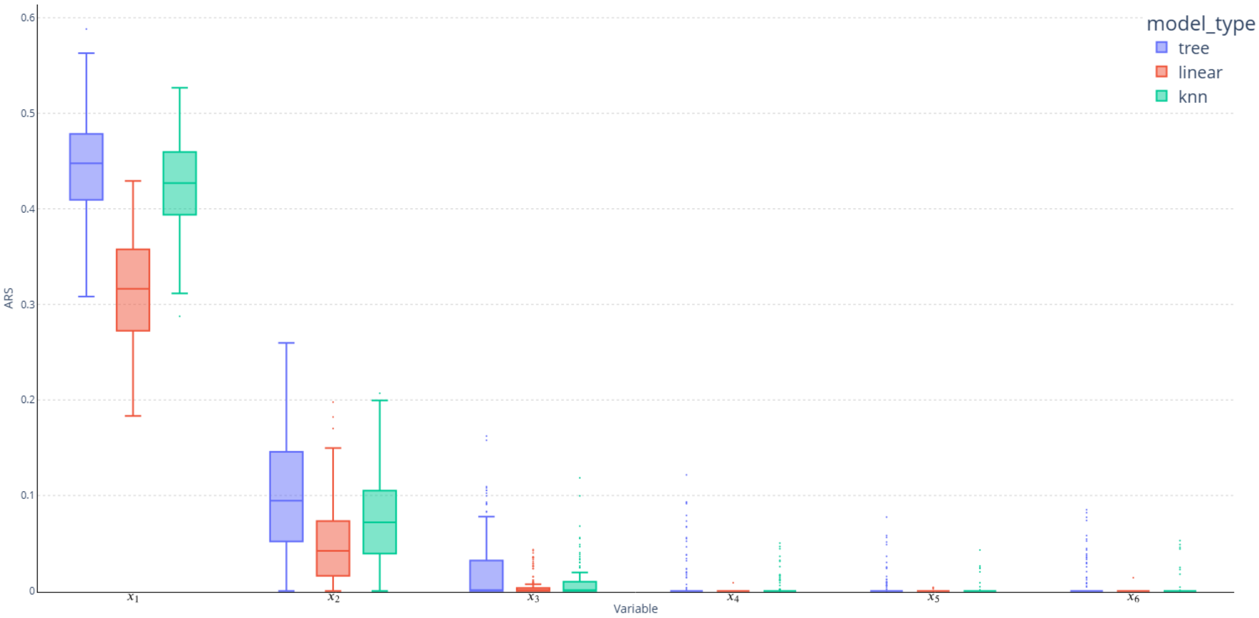

Figure 2. Performance of ARS with different base models on Equation (15) without the noise term η on the benchmark dataset. The informative variables ($\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$) and noninformative variables ($\mathbf{x}_4$, $\mathbf{x}_5$, and $\mathbf{x}_6$) are evaluated across three base models: decision trees, linear regression, and k-nearest neighbors.

Performance on benchmark dataset without noise for different base models: Figure 2 illustrates the ARS measure's sensitivity to both informative variables ($\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$) and noninformative variables ($\mathbf{x}_4$, $\mathbf{x}_5$, and $\mathbf{x}_6$) under different base models (decision trees, linear regression, and k-nearest neighbors) with the noise term η excluded. The decision tree and k-nearest neighbors models demonstrate a stronger capacity to capture the inherent dependencies, particularly for the nonlinear relationships in $\mathbf{x}_1$ and $\mathbf{x}_2$. In contrast, the linear regression model struggles to detect these dependencies, reflecting its limitations with nonlinear data. For the noninformative variables $\mathbf{x}_4$, $\mathbf{x}_5$, and $\mathbf{x}_6$, ARS consistently returns zero across all quartiles, with only occasional nonzero outliers, highlighting ARS's robustness in distinguishing relevant features from noise. This behavior underscores ARS's reliability in filtering out noninformative attributes, even in the absence of noise.

4.2.2. Results #2.2

To increase the dataset's complexity, additional noninformative predictor variables $\mathbf{x}_4$, $\mathbf{x}_5$, and $\mathbf{x}_6$ are generated uniformly within the range 0 to 1 ($U(0,1)$). These variables do not contribute directly to $\mathbf{y}$, simulating the presence of noise and further challenging the model's ability to differentiate between relevant and irrelevant features. To further enrich the dataset, we simulates the presence of progressively weaker signal-to-noise ratios, allowing the method to be tested on its ability to identify relevant variables in the presence of noise. To achieve this, we generate noise versions of the predictor variables using the following equation:

$$\mathbf{v}_i^{(j)} = \mathbf{x}_i + \left(0.01 + \frac{0.5(j-1)}{10-1}\right) N(0;0.3), \tag{16}$$

for $j = 1, \ldots, 10$ the level of noise, where the correlation between $\mathbf{x}_i$ and $\mathbf{v}_i^{(j)}$ systematically decreases as j increases. $i = 1, \ldots, 6$ corresponds to the index values of the original $\mathbf{x}_i$. This formula generates multiple noisy versions of the base variables $\mathbf{x}_i$. This experimental setup enables a robust comparison of our method against alternative techniques by specifically evaluating their capacity to distinguish between relevant and nonrelevant features, even in the presence of noise. By conducting 100 repetitions with 100 observations each, we aim to thoroughly assess the robustness, accuracy, and adaptability of the methods when faced with scenarios that include a mixture of informative and noninformative variables.

The evaluation of our proposed attribute relevance score (ARS) method, using decision trees as the base model, conducted alongside established dependency measures, was performed using the benchmark dataset described above. This dataset provides a challenging environment, containing both informative and noninformative variables with varying levels of noise, designed to test the robustness and discriminative ability of each method.

Performance on informative variables: For the variable $\mathbf{v}_1^{(j)}$, both versions $\mathbf{v}_1^{(1)}$ and $\mathbf{v}_1^{(10)}$ exhibit high mean ARS scores (0.435 and 0.264, respectively) compared to the other predictors, indicating strong relevance (see Figure 3). This aligns with the role of $\mathbf{x}_1$ as the primary driver of variability in $\mathbf{y}$ due to its exponential dependency. Furthermore, the low standard deviation of the ARS across both versions highlights the stability and robustness of this measure in consistently capturing the predictive power of $\mathbf{v}_1^{(j)}$.

In the case of $\mathbf{v}_2^{(j)}$, represented by $\mathbf{v}_2^{(1)}$ and $\mathbf{v}_2^{(10)}$, the mean ARS scores indicate relevance (0.087, std: 0.063 and 0.069, std: 0.064, respectively); however, these scores are notably lower than those obtained for $\mathbf{v}_1^{(j)}$ and other methods. This discrepancy is expected, considering the second objective of our study, which assesses the effectiveness of approximating $\mathbf{y}$ from $\mathbf{x}$ using a specific error measure and algorithm. ARS quantifies this practical capability, capturing not only the statistical dependence but also the magnitude of predictive power within the context of the selected model and performance metric. In this context, despite the added structured noise in $\mathbf{v}_2^{(j)}$, the mean ARS remains relatively stable, suggesting that the relevance of $\mathbf{v}_2^{(j)}$ is not significantly diminished by noise alone. This stability may imply that the ARS effectively captures a fundamental relationship between $\mathbf{v}_2^{(j)}$ and $\mathbf{y}$ that remains informative despite the noise corruption. The resilience of $\mathbf{v}_2^{(j)}$'s ARS score under noise indicates that the relationship retains predictive utility, though at a much lower magnitude compared to $\mathbf{v}_1^{(j)}$.

For $\mathbf{v}_3^{(j)}$, both versions $\mathbf{v}_3^{(1)}$ and $\mathbf{v}_3^{(10)}$ register low mean ARS scores (0.019, std: 0.033 and 0.018, std: 0.039, respectively), which is consistent with the weaker dependency of $\mathbf{x}_3$ on $\mathbf{y}$, particularly in the presence of external variability and more dominant dependencies.

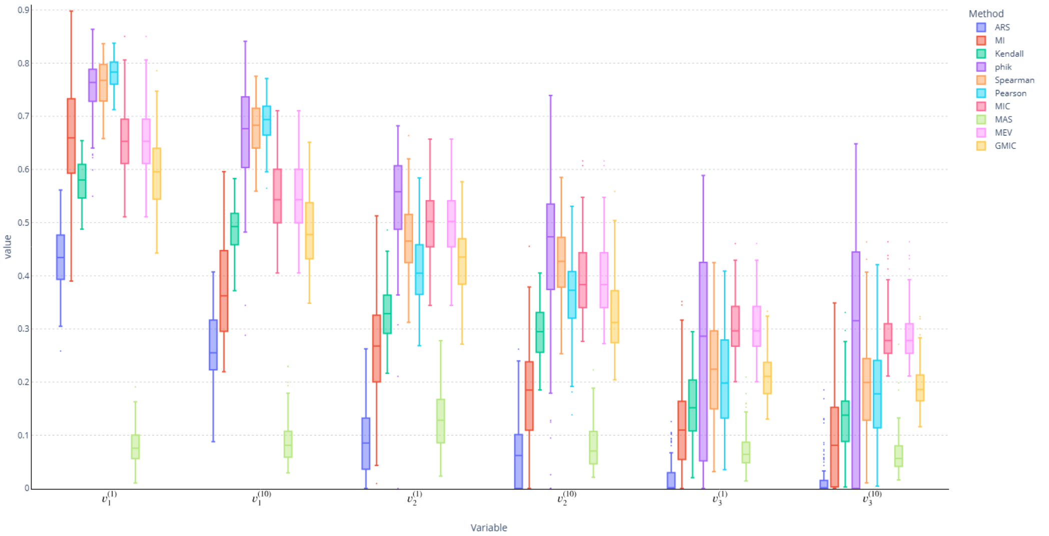

Figure 3. Performance of various dependency measures on informative variables $\mathbf{v}_1^{(j)}$, $\mathbf{v}_2^{(j)}$, and $\mathbf{v}_3^{(j)}$ across different noise levels (j). This figure illustrates the mean and standard deviation of each measure, highlighting the stability of ARS compared to other metrics.

In comparison, traditional methods such as Pearson and Spearman correlations displayed higher mean values for informative variables. For instance, the Pearson correlation for $\mathbf{v}_1^{(1)}$ was 0.781 (std: 0.028), indicating strong linear relationships. This pattern persisted across other informative variables, with measures like GMIC and MIC showing similarly elevated means (e.g., GMIC for $\mathbf{v}_1^{(1)}$ was 0.595, std: 0.065), apparently reflecting substantial associations. However, it is important to note that this situation also arises for noninformative variables, as we will show in the next results.

Performance on noninformative variables: In the case of noninformative variables $\mathbf{v}_4^{(j)}$, $\mathbf{v}_5^{(j)}$, and $\mathbf{v}_6^{(j)}$, ARS effectively identified the lack of relevance by consistently returning scores of zero across all instances, with only occasional nonzero values attributable to random chance (see Figure 4). The ARS scores for these variables were zero in the vast majority of cases, resulting in extremely low mean values and minimal standard deviations. For example, for $\mathbf{v}_4^{(1)}$, ARS returned a mean score of 0.009 with a standard deviation of 0.024, indicating that most scores were zero except for rare outliers. This demonstrates ARS's robustness in correctly identifying noise and nonrelevant attributes, as the minimal nonzero scores observed can be attributed to expected statistical fluctuations in practical datasets.

However, other dependency measures exhibited higher variability and occasionally elevated scores, even for noninformative attributes, suggesting spurious associations. Metrics such as GMIC and MIC presented mean values of 0.139 (std: 0.027) and 0.240 (std: 0.035), respectively, for $\mathbf{v}_4^{(1)}$, suggesting a tendency to overestimate the relevance of noisy attributes. Similarly, traditional correlation measures, such as Spearman and Pearson also showed notable variability (e.g., mean Spearman for $\mathbf{v}_4^{(1)}$ was 0.088, std: 0.068), indicating their limitations in distinguishing between informative patterns and noise.

Overall, the attribute relevance score (ARS) outperformed traditional and information-based metrics by providing a more stable and reliable assessment of attribute relevance across both informative and noninformative variables. While measures like Pearson, Spearman, and MIC can capture strong relationships when present, their high values—especially in noninformative scenarios—reduce their effectiveness in complex, noisy datasets. The fact that ARS returned zero scores for all noninformative variables, except for occasional outliers, underscores its robustness and practical utility in feature selection processes. By effectively filtering out noise and avoiding false positives, ARS stands out as a particularly powerful tool for distinguishing between meaningful and spurious relationships, thereby enhancing the reliability of feature selection in predictive modeling.

5. Discussion

This study introduces the attribute relevance score (ARS), a novel metric designed to robustly identify relevant attributes in complex datasets exhibiting both linear and nonlinear relationships with varying levels of noise. For instance, ARS consistently returned scores close to zero for noninformative variables, with mean scores below 0.01 and minimal standard deviations, whereas traditional metrics often yielded higher scores with greater variability. This distinguishes ARS from traditional dependency measures, which often exhibit significant fluctuations, especially when assessing noninformative variables.

A key finding from our experiments is ARS's robustness to noise. Traditional metrics such as Pearson, Spearman, and Kendall correlations struggle to maintain consistency in the presence of noise, frequently showing inflated scores for noninformative variables. This issue is particularly problematic in high-dimensional settings, where distinguishing between signal and noise is crucial. In contrast, ARS consistently produces low scores for noninformative attributes $\mathbf{v}_4^{(j)}$, $\mathbf{v}_5^{(j)}$, and $\mathbf{v}_6^{(j)}$ with minimal variability, as shown in Figure 4. This stability suggests that ARS more reliably filters out irrelevant features, enhancing the quality and interpretability of predictive models. Further supporting ARS's robustness, the analysis of ARS in the noise-free scenario (Figure 2) yields results closely aligned with those observed in noisier conditions. ARS maintains consistent sensitivity to informative variables ($\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$) across both scenarios, with decision trees and k-nearest neighbors models effectively capturing the dependencies. The consistency of these results, even in the absence of noise, underscores ARS's ability to reliably identify relevant features regardless of noise levels. Likewise, noninformative variables ($\mathbf{x}_4$, $\mathbf{x}_5$, and $\mathbf{x}_6$) continue to exhibit near-zero ARS scores, reinforcing ARS's capacity to filter out irrelevant features accurately. These findings demonstrate that ARS provides a stable assessment of attribute relevance, with minimal influence from random noise—essential for feature selection applications that demand reliability despite fluctuations in data quality.

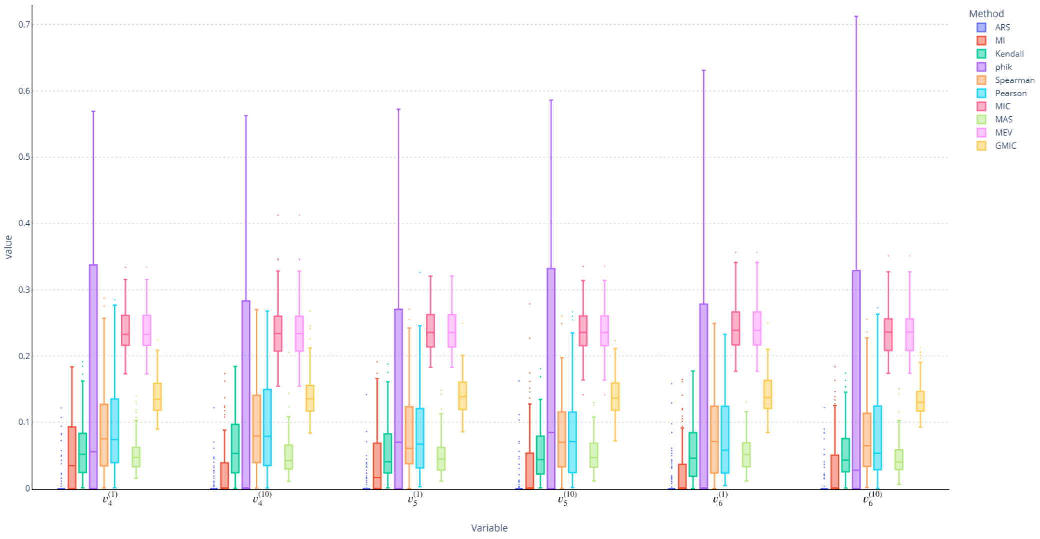

Figure 4. Performance of various dependency measures on noninformative variables $\mathbf{v}_4^{(j)}$, $\mathbf{v}_5^{(j)}$, and $\mathbf{v}_6^{(j)}$ across different noise levels (j). This figure shows the mean and standard deviation for each measure, demonstrating the high variability of traditional metrics compared to the consistently low scores of ARS. Absolute values are considered for readability.

Our comparative analysis against established metrics such as GMIC, MAS, and MIC further underscores the strengths of ARS. While these metrics aim to capture complex nonlinear relationships, they exhibit higher variability and occasionally assign moderate relevance scores to noninformative attributes, reflecting their susceptibility to noise. The high standard deviations observed indicate a lack of robustness, potentially leading to inconsistent feature selection. In contrast, ARS accurately identifies relevant attributes such as $\mathbf{v}_1^{(j)}$, $\mathbf{v}_2^{(j)}$, and $\mathbf{v}_3^{(j)}$ while maintaining low scores for noise variables, effectively balancing sensitivity to true relationships with resilience to random fluctuations.

The ability of ARS to consistently differentiate between relevant and irrelevant attributes has significant implications for high-dimensional data analysis, particularly in fields where data complexity and noise pose challenges, such as genomics, finance, and environmental modeling. By integrating statistical rigor with practical applicability, ARS contributes to more reliable feature selection, enhancing reproducibility and reducing the risk of overfitting. This is crucial in scientific research, where the goal is not only to build predictive models but also to uncover meaningful relationships that drive further understanding.

The significance of reproducibility in machine learning cannot be overstated, particularly in high-dimensional data analysis. Reproducibility concerns often arise when results depend heavily on specific data splits or model configurations, leading to conclusions that may not generalize. In feature selection, these issues are critical, as identifying relevant attributes is pivotal for subsequent analysis.

Our methodology incorporates concepts from reproducibility and A/B testing. Similar to A/B testing, we compare models trained on original attributes with those trained on *shadow attributes*, which are permuted versions of the original attributes. By evaluating whether the model using the original attributes consistently outperforms the one using shadow attributes across various data samples, we assess the relevance of each attribute. This comparison grounds attribute selection in robust statistical principles, ensuring that the selected attributes are genuinely informative and that the results are reproducible under diverse conditions.

By employing a probabilistic approach to assessing feature relevance, ARS reduces the risk of findings being artifacts of particular experimental setups. The use of shadow attributes and random sampling provides a rigorous framework for quantifying feature importance while controlling for variability introduced by data sampling and model configurations. Moreover, the ARS framework strengthens inferential reproducibility by ensuring that identified relevant attributes consistently contribute to model performance across various scenarios. This aligns with principles of methods reproducibility, results reproducibility, and inferential reproducibility.

We acknowledge that ARS computational complexity increases linearly with the number of input features. Specifically, for datasets with a large number of attributes, the need to train separate models for each attribute can become computationally intensive, posing practical challenges in high-dimensional machine learning applications. However, because each attribute is evaluated independently, ARS can leverage parallel computing resources by distributing computations across multiple processors or computational nodes. This parallelization significantly reduces total processing time, enabling ARS to remain scalable and maintain practical performance even with a large number of attributes.

While ARS offers clear strengths, it is not without limitations. The current implementation relies on decision trees, which may not capture all forms of attribute interactions, especially in highly nonlinear or interaction-heavy scenarios. Moreover, ARS is inherently a univariate approach, evaluating each attribute's relevance independently. This means it lacks the capability to detect redundancies and interactions between attributes. Synergistic effects, where combinations of attributes contribute significantly to the target variable, may be overlooked. This limitation can be critical in domains where attribute interactions play a crucial role. Future research should explore extending ARS to incorporate multivariate analysis techniques or models that can account for attribute interactions.

In summary, our work advances feature selection methodologies in high-dimensional data analysis and enhances the reproducibility and generalizability of results. By embedding statistical significance into a measure of attribute relevance and ensuring robustness across different conditions, we contribute to developing more reliable and reproducible machine learning models. Future work will focus on expanding the applicability of ARS to a broader range of models and exploring its integration into automated machine learning frameworks.

6. Conclusions and Future Work

The ARS method represents a versatile approach that can be adapted to various types of data and problem settings. Its application extends beyond standard regression and classification tasks to exploratory data analysis and feature engineering, where the identification of significant predictors is crucial. ARS's ability to provide consistent and reliable relevance scores, particularly in noisy and complex environments, aligns with the increasing need for robust, interpretable, and reproducible methods in data science. Its stability under noisy conditions and superior performance in distinguishing informative attributes position it as a valuable tool for researchers and practitioners. By addressing key challenges in feature selection for complex datasets, ARS improves predictive model accuracy and interpretability, paving the way for more robust data-driven decision making.

However, the relevance of an attribute should not depend solely on its individual contribution to model performance but also on its interactions with other attributes. This perspective highlights the need for a more holistic evaluation of attribute relevance within complex predictive models. In many scenarios, attributes do not function in isolation; their true value emerges through interactions with other features. Ignoring these interactions can lead to underestimating the importance of certain attributes that may appear irrelevant when considered univariately but have a significant impact in combination with others. Moreover, attribute redundancy should not unduly diminish an attribute's univariate relevance or intrinsic predictive capability. Two redundant attributes might share valuable information, and removing one could result in a loss of crucial insights. Therefore, it is essential to consider redundancy in a way that does not unduly penalize the relevance of attributes that, despite being redundant, are predictively valuable.

To address these considerations, future research should explore integrating ARS with other feature selection frameworks capable of effectively evaluating both attribute interactions and redundancies. Additionally, adapting ARS to more advanced learning models, such as deep neural networks, could enhance its applicability across different contexts. Expanding the evaluation of ARS on real-world datasets from various domains would further validate its effectiveness and uncover additional insights into its utility.

In conclusion, a robust evaluation of attribute relevance that carefully considers both interactions among attributes and the potential for redundancy is essential for developing predictive models that are not only accurate but also interpretable and reliable. Furthermore, this strategy emphasizes the critical distinction between a model that consistently improves performance due to meaningful insights and one that does so merely by chance. By focusing on selecting nonredundant variables, even at the expense of optimal performance in some metrics, we align the model's outcomes more closely with the realities of the underlying phenomena. This approach moves us towards a more robust definition of understanding within the field of machine learning. Through the development of a multivariate relevance measure, we can assess when a model is not just learning correlations but is genuinely comprehending the underlying phenomena. This deeper level of understanding enables models to generalize more effectively, providing reliable interpretations and fostering the model's capacity to make informed decisions across diverse scenarios. Consequently, this advancement enhances the fidelity and reliability of model interpretations, contributing to a more nuanced and scientifically sound understanding of the model's ability to generalize and adapt.

Author Contributions: Conceptualization, P.N. and C.S.; methodology, C.S.; software, P.N.; validation, P.N. and C.S.; resources, C.S.; writing—original draft preparation, P.N.; writing—review and editing, C.S. and H.A.; visualization, P.N.; supervision, C.S.; project administration, C.S. and H.A.; funding acquisition, C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Project USM PI_LIR_24_14.

Data Availability Statement: The data can be found at the following link: https://github.com/Pneirz/Attribute-Relevance-Score/ (accessed on 5 November 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ullah, S.; Mahmood, Z.; Ali, N.; Ahmad, T.; Buriro, A. Machine Learning-Based Dynamic Attribute Selection Technique for DDoS Attack Classification in IoT Networks. *Computers* **2023**, *12*, 115. [CrossRef]
2. Kang, I.A.; Njimbouom, S.N.; Kim, J.D. Optimal Feature Selection-Based Dental Caries Prediction Model Using Machine Learning for Decision Support System. *Bioengineering* **2023**, *10*, 245. [CrossRef] [PubMed]
3. Kiratsoudis, S.; Tsiantos, V. Enhancing Personnel Selection through the Integration of the Entropy Synergy Analysis of Multi-Attribute Decision Making Model: A Novel Approach. *Information* **2024**, *15*, 1. [CrossRef]
4. AL-Gburi, A.F.J.; Nazri, M.Z.A.; Yaakub, M.R.B.; Alyasseri, Z.A.A. Multi-Objective Unsupervised Feature Selection and Cluster Based on Symbiotic Organism Search. *Algorithms* **2024**, *17*, 355. [CrossRef]
5. Ðurasević, M.; Jakobović, D.; Picek, S.; Mariot, L. Assessing the Ability of Genetic Programming for Feature Selection in Constructing Dispatching Rules for Unrelated Machine Environments. *Algorithms* **2024**, *17*, 67. [CrossRef]
6. Nilsson, R.; Peña, J.M.; Björkegren, J.; Tegnér, J. Consistent Feature Selection for Pattern Recognition in Polynomial Time. *J. Mach. Learn. Res.* **2007**, *8*, 589–612.
7. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]
8. Mnich, K.; Rudnicki, W.R. All-relevant feature selection using multidimensional filters with exhaustive search. *Inf. Sci.* **2020**, *524*, 277–297. [CrossRef]
9. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
10. Strobl, C.; Boulesteix, A.L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **2007**, *8*, 25. [CrossRef]

11. Stoppiglia, H.; Dreyfus, G.; Dubois, R.; Oussar, Y. Ranking a Random Feature For Variable And Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1399–1414. [CrossRef]
12. Degenhardt, F.; Seifert, S.S. Evaluation of variable selection methods for random forests and omics data sets. *Briefings Bioinform.* **2019**, *20*, 492–503. [CrossRef] [PubMed]
13. Kursa, M.B.; Rudnicki, W.R. Feature selection with the Boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
14. Wetschoreck, F. RIP Correlation. Introducing the Predictive Power Score. Towards Data Science, Medium. 2020. Available online: https://towardsdatascience.com/rip-correlation-introducing-the-predictive-power-score-3d90808b9598 (accessed on 7 November 2024).
15. Goodman, S.; Fanelli, D.; Ioannidis, J. What does research reproducibility mean? *Sci. Transl. Med.* **2016**, *8*, 341ps12. [CrossRef] [PubMed]
16. Bouthillier, X.; Laurent, C.; Vincent, P. Unreproducible Research is Reproducible. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019.
17. Bouthillier, X.; Delaunay, P.; Bronzi, M.; Trofimov, A.; Nichyporuk, B.; Szeto, J.; Mohammadi Sepahvand, N.; Raff, E.; Madan, K.; Voleti, V.; et al. Accounting for Variance in Machine Learning Benchmarks. *Proc. Mach. Learn. Syst.* **2021**, *3*, 747–769.
18. Kohavi, R.; Longbotham, R. Online Controlled Experiments and A/B Testing. In *Encyclopedia of Machine Learning and Data Mining*; Sammut, C., Webb, G.I., Eds.; Springer: Boston, MA, USA, 2017; pp. 922–929. [CrossRef]
19. Deng, A.; Xu, Y.; Kohavi, R.; Walker, T. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, New York, NY, USA, 4–8 February 2013; WSDM'13; pp. 123–132. [CrossRef]
20. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017; NIPS'17; pp. 4768–4777.
21. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?". In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, CA, USA, 13–17 August 2016. Available online: https://arxiv.org/pdf/1602.04938.pdf (accessed on 7 November 2024). [CrossRef]
22. Sokol, K.; Flach, P. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, New York, NY, USA, 27–30 January 2020; FAT*'20; pp. 56–67. [CrossRef]
23. Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; Müller, K.R. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* **2019**, *10*, 1096. [CrossRef]
24. Pfungst, O. *Clever Hans (The Horse of Mr. von Osten): A Contribution to Experimental, Animal, and Human Psychology*; Holt, Rinehart, and Winston: New York, NY, USA, 1911.
25. Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A.; Etmann, C.; McCague, C.; Beer, L.; et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **2021**, *3*, 199–217. [CrossRef]
26. DeGrave, A.J.; Janizek, J.D.; Lee, S.I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **2021**, *3*, 610–619. [CrossRef]
27. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
28. Thomas, T.; Straub, D.; Tatai, F.; Shene, M.; Tosik, T.; Kersting, K.; Rothkopf, C.A. Modelling dataset bias in machine-learned theories of economic decision-making. *Nat. Hum. Behav.* **2024**, *8*, 679–691. [CrossRef]
29. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6638–6648. [CrossRef]
30. Baak, M.; Koopman, R.; Snoek, H.; Klous, S. A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Comput. Stat. Data Anal.* **2020**, *152*, 107043. [CrossRef]
31. Ross, B.C. Mutual information between discrete and continuous data sets. *PLoS ONE* **2014**, *9*, e87357. [CrossRef] [PubMed]
32. Reshef, D.N.; Reshef, Y.A.; Finucane, H.K.; Grossman, S.R.; McVean, G.; Turnbaugh, P.J.; Lander, E.S.; Mitzenmacher, M.; Sabeti, P.C. Detecting novel associations in large data sets. *Science* **2011**, *334*, 1518–1524. [CrossRef] [PubMed]
33. Albanese, D.; Riccadonna, S.; Donati, C.; Franceschi, P. A practical tool for maximal information coefficient analysis. *GigaScience* **2018**, *7*, giy032. [CrossRef] [PubMed]
34. Albanese, D.; Filosi, M.; Visintainer, R.; Riccadonna, S.; Jurman, G.; Furlanello, C. minerva and minepy: A C engine for the MINE suite and its R, Python and MATLAB wrappers. *Bioinformatics* **2012**, *29*, 407–408. [CrossRef]
35. Luedtke, A.; Tran, L.H. The Generalized Mean Information Coefficient. *arXiv* **2013**, arXiv:1308.5712. [CrossRef]
36. Chen, Z.; Zhang, W. Integrative analysis using module-guided random forests reveals correlated genetic factors related to mouse weight. *PLoS Comput. Biol.* **2013**, *9*, e1002956. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.