Towards a Comprehensive Evaluation of Decision Rules and Decision Mining Algorithms Beyond Accuracy

Beate Wais[1,2] and Stefanie Rinderle-Ma[3]

[1] University of Vienna, Faculty of Computer Science, Research Group Workflow Systems and Technology; UniVie Doctoral School Computer Science DoCS, Vienna, Austria, beate.wais@univie.ac.at

[2] Technical University of Munich, Germany; TUM School of Computation, Information and Technology, Garching, Germany, stefanie.rinderle-ma@tum.de

Abstract. Decision mining algorithms discover decision points and the corresponding decision rules in business processes. So far, the evaluation of decision mining algorithms has focused on performance (e.g., accuracy), neglecting the impact of other criteria, e.g., understandability or consistency of the discovered decision model. However, performance alone cannot reflect if the discovered decision rules produce value to the user by providing insights into the process. Providing metrics to evaluate the decision model and decision rules comprehensively can lead to more meaningful insights and assessment of decision mining algorithms. In this paper, we examine the ability of different criteria from software engineering, explainable AI, and process mining that go beyond performance to evaluate decision mining results and propose metrics to measure these criteria. To evaluate the proposed metrics, they are applied to different decision algorithms on two synthetic and one real-life dataset. The results are compared to the findings of a user study to check whether they align with user perception. As a result, we suggest four metrics that enable a comprehensive evaluation of decision mining results and a more in-depth comparison of different decision mining algorithms. In addition, guidelines for formulating decision rules are presented.

Keywords: Process Mining · Decision Mining · Evaluation · Metrics · User Study · Explainability

1 Introduction

An important part of process discovery is decision mining, which provides algorithms to discover decision points in processes and the underlying decision rules guarding that decision based on event logs [22]. The discovered decision points and rules can enhance transparency by capturing the underlying logic of decisions and allowing users to understand the decisions in a process, i.e., *"making implicit decision information explicit"* [18]. Process and decision mining are increasingly gaining traction as transparency and standardization become crucial across different domains [9, 18]. Decision mining enables domain experts to detect potential deviations from underlying business logic (e.g., regulations) or discover changes in the process [30], as well as evaluate if these deviations and changes are intentional or due to errors (cf. [36]). This ability, in turn, can lead to fewer errors or a decrease in time until an error is detected, thereby minimizing the negative impact of an error. Decision mining algorithms[3] have been evaluated using common sense so far (see [3, 6, 31]) or concerning performance, e.g., accuracy and fitness [20, 23, 29]. Related fields such as rule induction also use performance-based criteria, e.g., coverage or error rates to evaluate results, cf. [4]. Similarly, in process mining, methods to evaluate process discovery algorithms include criteria such as fitness or precision [27]. While high performance is necessary, it is not sufficient to achieve valuable decision mining results. By valuable, we refer to the ability to accomplish the intended goal, e.g., to provide a profound understanding of process decisions and enable the user to take action.

Imagine a logistics use case where temperature-sensitive cargo is moved to a destination, where the cargo is unloaded and transferred to the customer. During transportation, the temperature is measured. As the destination is reached, it is checked if the temperature exceeds 25 degrees more than three times. If so, the goods are not OK ('NOK') and must be discarded. Otherwise, they will be transferred to the customer. The discovered decision rule could look like ①: IF $temperature.count(>= 26.0) >= 4.0$ THEN Discard Goods.

Another decision mining algorithm might discover the following decision rule ②: If $temperature\_quantile\_q\_0.8 > 25.90$ AND $temperature\_change\_quantiles\_f\_agg\_var\_isabs\_True\_qh\_1.0\_ql\_0.6\_ <= 27.67$ THEN Discard Goods. Rules ① and ② describe the same logic, which might not be visible at first sight. The second version is more complex, as the decision attributes are engineered, including more complex names and statements. Therefore, although the performance, measured using accuracy, is the same, the second version might provide less insight into the process. Insufficient insight might lead to a misinterpretation of the decision rules and, in turn, result in unfavorable business decisions. In the example, one could derive from ② to discard the goods for any temperature measurement above 25.90 degrees, which can result in either unnecessarily discarded goods or a violation of food safety regulations. This example shows that accuracy, or any performance-related metric, alone is insufficient to evaluate decision mining results. Having metrics to measure the performance and the ability to provide valuable insights can help decide which algorithm to apply for a specific use case.

Measuring criteria beyond performance-related criteria is also a goal of explainability in AI (XAI). XAI generates explanations for black-box models, thereby providing more information to the users. The extent of explainability achieved by the provided explanations can be measured using different metrics. Mostly, the understandability of the generated explanation, as well as how accurately the explanation fits the underlying model, are evaluated; see [24, 37] for a general overview and [32] for explainability metrics with regards to predictive process monitoring. The criteria used in XAI can be a start to evaluating decision

[3] Note that the result of a decision mining algorithm comprises the decision model as well as the textual decision rules stemming from the decision model.

mining results. However, further criteria may be of interest. Software engineering and process mining domains provide additional criteria beyond performance-related criteria and might be valuable in evaluating decision mining results.

Therefore, this work aims to discover which criteria are suitable to evaluate decision mining results in addition to performance-based criteria by addressing the following research questions. **RQ1:** Which non-performance related criteria are relevant in the context of decision mining? **RQ2:** How to measure non-performance related criteria in decision mining? **RQ3:** How to achieve decision mining results that enable users to make informed decisions?

To answer RQ1-RQ2, software engineering, XAI, and process mining literature is analyzed to find suitable criteria in Sect. 2. Section 3 proposes metrics to evaluate these criteria regarding decision mining results. The metrics are applied to different data sets and compared to user perception in Sect. 4, followed by proposing guidelines (↦ RQ3). A conclusion is given in Sect. 5.

2 Literature Review on Evaluation Criteria

Evaluating decision mining results should encompass performance-based and non-performance-based criteria to ensure a comprehensive evaluation. In this section, we look at literature from the domains of software engineering (SE), XAI, process mining (PM), and business process quality (QoBP) to analyze which criteria are used to evaluate methods and created artifacts. The literature analysis is not exhaustive. However, it provides comprehensive criteria covering many aspects of decision mining results.

We start by discussing SE literature as there exists an abundance of criteria to evaluate different aspects of created software, e.g., [19], and because SE criteria have already been applied to process modeling research. [34], for example, analyze how SE quality metrics can be applied to process modeling and [11] in the context of business processes in general. A concept in the SE domain related to decision mining is *useful transparency*. It refers to the goal of *"enabling stakeholders to make decisions based on the provided information and act upon them"*, capturing the challenge to get from information being available to information being useful [12]. The authors suggest that *information quality* is essential in achieving useful transparency. The information quality criteria used in [12] are originally defined by [14] and encompass 16 criteria including consistency, free-of-error, and understandability. Other information and data quality frameworks[4] such as the literature review on data quality in [10] often contain similar and overlapping criteria. The first conclusion is that data and information quality criteria provide a reasonable basis for our goal of comprehensive evaluation in decision mining.

Data quality in decision mining can be analyzed at multiple levels. Firstly, the underlying data, i.e., *event logs*, can be evaluated as the quality of the event

[4] Data and information are distinct concepts; data consists of the raw data points and requires some interpretation to become information [10]. However, the quality criteria are strongly overlapping.

logs affects the quality of decision mining results. We consider event logs as out of scope for this work and instead refer to [2]. Secondly, the *decision model* as primary decision mining result can be evaluated. Thirdly, the textual *decision rules*, which are generated using the decision model and presented to the user, can be examined. We select data quality dimensions relevant to decision mining and address the quality of the decision model and decision rules. The 16 information quality criteria provided in [14] are used as basis and compared to quality criteria from literature in SE [10], XAI [24], PM [16], and QoBP [11]. The results are summarized in Table 1 and discussed in the following. The last column indicates whether the criterion refers to the quality of the event log, the decision model, or the decision rule. All criteria will be analyzed, and the criteria relevant to decision mining are highlighted in grey.

Table 1. Quality criteria in literature.

Information Quality [14]	Data Quality [10]	XAI [24]	Qualitative PM Criteria [16]	QoBP, Input/ Output [11]	Level
Accessibility	Accessibility			Accessibility	L(og)
Appropriate Amount of Information				Amount of Data	L
Believability				Believability	R(ule)
Completeness	Completeness	Fidelity (Completeness)	Quality (Completeness)	Completeness	M(odel)
Concise Representation		Interpretablity (Parsimony)	Quality (Conciceness)		R
Consistent Representation	Concistency	Interpretability (Clarity)	Quality (Concistency)		R, M
Ease of Manipulation					N/A
Free-of-Error	Accuracy, Validity, Reliability	Fidelity (Soundness)	Quality (Correctness)	Accuracy	M
Interpretability		Interpretability	Understandability (Readability)		R
Objectivity				Objectivity	L
Relevancy	Relevancy			Relevancy	R
Reputation				Reputation	L
Security				Security	L
Timeliness	Timeliness, Currentness			Timeliness	L, M
Understandability		Interpretability (Parsimony)	Understandability (Complexity)		R
Value-Added			Usability	Value-Added	R

XAI research provides criteria relating to explainability as discussed in Sect. 1. In XAI, explainability consists of two main criteria: understandability of the generated explanation and model fidelity, i.e., how well the explanations represent the underlying model. The two criteria can be further split into clarity, parsimony, completeness, and soundness. Understandability is relevant for decision mining. Model fidelity is relevant for non-transparent methods, e.g., neural networks. Typically, inherently transparent decision trees are used in decision mining, and a proxy model is unnecessary to generate an explanation. However, in the context of the correctness of the discovered decision model, the sub-criteria completeness and soundness are relevant.

Qualitative criteria to evaluate PM results are explored by [16]. PM artifacts are usually examined concerning three qualitative aspects, i.e., understandability, quality, and usability. Understandability and quality can be further split into sub-dimensions, overlapping with criteria named in [14], e.g., completeness and conciseness. Usability is defined as being beneficial to the user, which aligns with the criteria “Value-Added’ in [14]’. Dimensions of QoBP have been defined based on SE quality criteria [11]. The Input/Output quality aspect is relevant to this work, as decision models and rules could be seen as output. These have substantial overlap with [14].

Table 1 shows the mapping of criteria from XAI, PM, and QoBP onto information quality criteria. All criteria correspond or overlap with information quality dimensions. The terms used in literature are inconsistent, e.g., interpretability and understandability are sometimes synonymous and sometimes seen as different concepts. “Believability” refers to the extent to which the presented rule is credible. It is important for the presented decision rule to be accepted as credible to be useful to users. The extent to which all necessary data values are included is described by the “Completeness” criterion. This questions whether the decision model covers all cases and data ranges. Rules should be represented as simply as possible, captured by “Concise Representation”. “Consistent Representation” can refer to the rule format, but also if the model is free of contradiction. Both interpretations are potentially relevant to evaluating decision mining results. The criterion “Ease of Manipulation” captures the ability to change information. In decision mining, this would apply to cases where a user feedback mechanism exists, which is generally not the case; therefore, this criterion is not applicable. “Free-of-Error” refers to the correctness, i.e., in decision mining to the correctness of the discovered decision model and decision rules and therefore relates to performance-based criteria such as accuracy. “Interpretability” refers to the extent to which appropriate language, symbols, units, and definitions are used, which relates to attributes and conditions used in the textual rules in decision mining. The degree to which a rule meets the expectations and requirements of a user relates to “Relevancy”. The extent to which data is sufficiently up-to-date is captured by “Timeliness”. This can refer to the log data and the decision model if it is updated regularly to incorporate changes. This criterion is relevant for online decision mining. As up to now, only one online decision mining algorithm exists, see [30], this criterion is not included in the analysis presented here. The extent to which a user easily comprehends a rule is described by the criterion “Understandability”, which relates to the complexity of a decision rule. Note that in the definition by [14] “Interpretability” and “Understandability” cover different aspects, the first one describing the comprehensibility of the contained variable names, symbols, etc., whereas the second term covers the comprehensibility of the overall rule, which are related but not identical issues. “Value-Added” can be equated to the overall goal of decision mining, i.e., providing not just information but useful information. In total, the following criteria have been defined as relevant for decision mining, i.e., “Believability”,“Completeness”, “Conciseness”, “Consistency”, “Free of Error”, “Interpretability’, “Relevancy”, “Understandability”’ and “Value-Added”.

The following section will analyze existing metrics and, if necessary, propose new metrics to quantify the specified criteria.

3 Metrics in Decision Mining

Metrics to assess non-performance-based criteria enable evaluating decision mining results and comparing different decision mining algorithms. Data and information quality criteria can be used as proxy measurements to measure overall achieved quality. This relates to a functionally grounded evaluation strategy [5]. This section contains an analysis of related literature and a proposition of metrics to measure the defined criteria in decision mining.

Information and data quality frameworks such as [8, 10, 35] propose metrics for measuring quality criteria. The literature on information quality often refers to databases, web pages, and search engines. Therefore, some metrics are too broad or too specific for the context of decision mining. [10] give an overview of data quality frameworks and related measurements for, e.g., completeness, which is calculated by dividing all available items by the number of expected items, i.e., missing values in a database. Information quality regarding search engines can be assessed using a mix of quantitative metrics such as consistency using the number of style guide deviations and user surveys for, e.g., comprehensiveness or clarity [15]. Searching for missing values in a database and looking at web page style guides do not apply to decision mining.

Looking at XAI literature such as [24, 37], different measures for explainability dimensions are defined. Understandability[5], for example, is measured using parsimony, i.e., the complexity of explanations. Parsimony is calculated using the number of attributes for different attributes (e.g., control flow or event attributes). Another metric is the effective complexity, which calculates the dependency of a prediction on specific attributes. For explainable predictive process monitoring, [32] discusses evaluation metrics. Parsimony and functional complexity are used to evaluate the understandability of predictions. Functional complexity measures the model complexity, similar to effective complexity, by permutating the possible values for each attribute and measuring the change in predictions.

Based on the literature analysis, we propose six metrics covering the criteria defined in Sect. 2: Accuracy (Free-of-Error), Model Completeness (Completeness), Effective Complexity (Conciseness), Interpretability (Interpretability), Parsimony (Understandability), and Remine Consistency (Consistency).

[5] Please note that the terms interpretability and understandability are often used interchangeably in the literature. However, interpretability in our context refers to the attribute names, not the overall understanding of the decision rule, building on the definition by [14]. Therefore, we use the term understandability, even if the related work uses the term interpretability.

The possible value range for all metrics is $[0, 1]$; higher values indicate “better”, e.g., more accurate or less complex, results. Criteria “Believability”, “Relevancy” and “Value-Added” do not have an associated metric. ‘Believability” and “Relevancy” require a user survey to be evaluated appropriately, as they are inherently subjective dimensions. The criterion “Value-Added” refers to the overall benefit of the decision mining results, which depends on the use case and requires a different evaluation strategy, see [5].

Accuracy evaluates if the discovered decision rule can classify instances correctly. The following definition is used:

$$Accuracy := \frac{Number\ of\ correctly\ classified\ instances}{Total\ number\ of\ instances}$$

Accuracy is used to evaluate the criterion “Free-of-Error”. A rule is assumed to be correct, i.e. free-of-error, if the accuracy is high. There still might be cases where the accuracy is high, but the rule is incorrect, for example, due to noisy data or overfitting. For the evaluation, a second, broader, performance-based metric is added, the *F1 Score*, which considers precision and recall [33].

Completeness can refer to different aspects and is used differently in literature, e.g. [10]. For decision mining, completeness can be defined as all possible classes, i.e., paths, are covered by the decision model and can be measured by:

$$Model\_Completeness := \frac{Number\ of\ classes\ in\ decision\ model}{Total\ number\ of\ classes}$$

If only two classes exist and the decision model covers one class, we assume the other class is the default.

Effective Complexity - EC is defined by [26] as the minimum number of attributes that can meet an expected performance measure; lower values indicate simple and less complex models. Similarly, [32] use functional complexity to measure understandability. Functional complexity is calculated by permutating attributes of an explanation, measuring the resulting prediction changes using the Hamming Distance. We adapt these definitions and calculate the effective complexity for decision mining by looking at the contained conditions in a decision rule and measuring the change in results if one condition is dropped at a time, using the Hamming Distance, see Alg. 1.

Algorithm 1 Effective Complexity in Decision Mining

```
Input: Decision Rule R,Output: Effective Complexity
Change = 0, Split rule R in Conditions C, delimiter: “AND”,“OR”
for c in C do
    Make new rule r without c, make prediction p with r
    Calculate Normalized Hamming Distance(p, original prediction)
    Change += Hamming Distance
end for
EffectiveComplexity = 1-(Change/#Conditions)
```

Effective complexity relates to the criterion "Conciseness", as higher values indicate that the results are considerably altered if a condition is removed, and therefore, the rule only contains necessary conditions.

Interpretability for decision rules covers the comprehensibility of attribute names, symbols, and used units, as defined in Sect. 2. A new metric is proposed that evaluates the interpretability of each used attribute by considering if the attribute name contains special characters and can be found in the dictionary, i.e., if it is an understandable word[6]. Note that this definition also includes syntactic accuracy, which leads to classifying a word as not understandable if there are spelling mistakes, for example; this is reasonable as spelling mistakes can make it more difficult to comprehend names and conditions. The overall length of the rule is also taken into account, as longer rules make it more difficult to interpret the contained attributes and conditions. Units are not included, as this strongly depends on the underlying data, i.e., if the units are part of the log. The calculation can be seen in Equation 1.

$$\begin{aligned} SpecialChars &= 1 - \frac{\#Special\ Characters}{\#Characters} \\ NonWords &= 1 - \frac{\#Words\ not\ in\ Dictionary}{\#Words\ in\ Rule} \\ Interpretability &= \frac{1}{Length(Rule)} * x + SpecialChars * y + NonWords * z \end{aligned} \tag{1}$$

The three conditions are scaled to 1 using weights, *x, y, z*, with $x+y+z=1$ and $x, y, z \in [0,1]$. The weights are optimized using the results of the pre-test user study, see Sect. 4.

Parsimony is often used to evaluate the understandability of an explanation, e.g., [37]. It measures the complexity of an explanation or, in this case, a decision rule. The less complex an explanation is, the more understandable it is for humans [25]. Parsimony can be defined as the number of attributes part of an explanation [13, 32]. We extend that definition by considering the number of relational conditions in a decision rule. Relational conditions are defined as conditions where the relationship between two or more attributes is relevant, e.g., `temperature1` $<$ `temperature2` instead of `age` $>$ 40, which adds to the complexity. Weights x and y, with $x+y=1$ and $x, y \in [0,1]$, are optimized using the pre-test results from Sect. 4.

$$Parsimony = \frac{1}{\#Attributes} * x + (1 - \frac{\#RelationalConditions}{\#Conditions}) * y \tag{2}$$

Remine Consistency - RC is the degree to which the decision rule stays consistent when the decision model is re-discovered on the same input data, thereby measuring the "Consistency". In literature, consistency can relate to

[6] Which languages are checked can be changed. Currently, English and German dictionaries are used.

consistency regarding a format, i.e., the proportion of items consistent with a format [10] or consistency regarding the model as used here. For decision mining, the following metric is implemented:

$$Consistency := 1 - \frac{LevenshteinDistance(Rule, ReminedRule)}{Length(Rule)} \tag{3}$$

The extent of change in a decision rule is measured using the Levenshtein distance, which calculates the least changes required to change one string into another. A high remine consistency can indicate that the model cannot accurately represent the underlying logic, therefore describing "Free-of-Error" as well, as the model changes even though the input data stays the same.

In the following sections, the proposed metrics are applied to different data sets and compared to the results of a user study.

4 Experimental Analysis and User Study

To evaluate the *feasibility* of the metrics proposed in Sect. 3 they have been implemented using Python except for accuracy and F1 Score, which are calculated using existing libraries. For the *applicability* evaluation, the metrics are applied to nine decision rules in the experimental analysis. To *validate* the metrics, a user study was conducted, and the results were compared. The source code, datasets, user study questionnaire as well as full results are available online[7].

4.1 Experimental Analysis

We start with a short description of the evaluation data sets.
Use Case I - Logistics is based on the running example (cf. Sect. 1).
Use Case II - Manufacturing is an example from the manufacturing domain, where a workpiece is produced and manually measured. The measurements are compared to the tolerances in the engineering drawing to check if the workpiece is “OK” or “Scrap”. One of the discovered decision rule is ③:
`IF` $measurement1 > 9.5$ `AND` $measurement1 <= 20.0$ `AND` $measurement2 <= 70.5$ `AND` $measurement0 > 19.5$ `AND` $measurement0 <= 80.5$ `AND` $measurement2 > 29.5$ `THEN Put in OK pile`.
Use Case III - Manufacturing contains data from a real-life manufacturing process [7]. Workpieces are produced, and the workpiece’s diameter is subsequently measured using the workpiece silhouette. This takes a couple of seconds but can be inaccurate. Therefore, the workpieces are transferred to a second measuring machine to measure more attributes, e.g., surface quality and flatness, resulting in more precise results. This step takes a couple of minutes. Therefore, the goal is to filter most workpieces using the first measuring step and only continue to the next step with workpieces that are likely to be “OK”. An exemplary

[7] `https://github.com/bscheibel/dm_eval`

discovered decision rule looks like ④:
IF $diameter\_intervall2\_percentchange > 0.16$ THEN Discard Goods

We compare existing decision mining algorithms *BDT*, *EDT-TS*, *EDT*, and *BranchMiner*. *BDT* uses a standard decision mining algorithm without including attribute engineering methods [28]. *EDT-TS* can work with time series data and might lead to more insightful decision rules when time series data is involved [29]. EDT-TS works by applying different attribute engineering methods and can be further divided by which attributes are produced. i.e., if the time series data is split into intervals, calculations are applied on the whole time series, or pattern-based attributes are engineered. *EDT* [28] and *BranchMiner* [21] are decision mining algorithms that can include relational conditions by generating new attributes. For each use case, three different algorithms were applied according to the data, i.e., Use Case 1 contains time series data; therefore, EDT-TS was applied with different attribute engineering methods. As all use cases contain a binary decision, all algorithms result in binary decision rules, i.e., a rule is given for one class, and the other is seen as the default class. The metrics proposed in Sect. 3 are calculated for each result. In addition, a combined metric (**Interpretability&Parsimony–I&P**) is calculated, as these two metrics show a high correlation.

Table 2 shows selected metric results calculated by algorithms for Rules ① to ⑤. Rule ⑤ for Use Case I is added for comparison:
IF $temperature\_intervall1\_max > 25.5$ AND $temperature\_intervall2\_max > 25.5$ AND $temperature\_intervall4\_max > 25.5$ THEN Discard Goods.

Table 2. Exemplary results from the experimental analysis.

Rule	**Use Case**	**Algorithm**	**Accuracy**	**I&P**	**Effect. Complexity (EC)**	**Remine Consistency (RC)**
①	I	EDT-TS	1	0.82	1.00	1.00
②	I	EDT-TS	0.99	0.67	0.21	1.00
⑤	I	EDT-TS	0.7	0.75	0.5	0.23
③	II	BDT	1.00	0.8	0.08	1.00
④	III	EDT-TS	0.91	0.84	1.00	1.00

Rule ① has high accuracy,"I&P", "EC" and "RC". Rule ② has almost the same accuracy. However, "I&P" and "EC" are considerably lower. The low "EC" value indicates that the rule is not as concise as possible. The lower value in "I&P" is probably due to complex variable names that hinder an intuitive understanding. Rule ⑤ is lower in accuracy; the "I&P" values lie between rules ① and ②. The rule contains a list of interval features, which are readable but more complex than ① due to multiple conditions, which are not as intuitive. The "EC" is also lower, indicating that not all conditions are essential. The low "RC" hints that the rule does not entirely cover all necessary conditions; therefore, ③ contains redundant conditions, but not all necessary conditions are discovered. Therefore, the first rule is the most suitable for this use case. Similar observations can be made for rule③ and ④. Rule ③ has redundant conditions; the "EC" is very low. The "I&P" value is higher, as the conditions are simple. However, several conditions are part of the rule. Rule ④ has high overall values; "I&P" is at 0.84 as the rule consists of one condition with a complex attribute name.

4.2 User Study

To evaluate the validity of the proposed metrics regarding user perception, a **user study** was conducted. The user study is used to analyze if metrics and user perception correlate, and the metrics enable an assessment of the user's perceived benefits of the textual decision rules. The user study is based on a questionnaire adapted from [17] and contains four sections. First, an introduction to decision mining and general questions are presented. Then, each section covers one use case, including a description of the use case and three decision rule versions, including the rules mentioned in Sect. 1. For each rule presented, the participants had to rate the rules according to the criteria, using a scale from 1(strongly disagree) to 5(strongly agree). In addition, a possibility for comments was provided.

Selection of participants In total, 20 participants filled in the questionnaire. 7 participants were part of the pre-test. According to the pre-test feedback, the phrasing of questions was adapted, and a clear definition of the criteria was added. In the main phase of the user study, 13 participants, consisting of master's students, PhD students, and post-docs, filled out the questionnaire. The selection of participants was guided by several considerations. Firstly, practical feasibility was a key factor, as procuring sufficient participants for a comprehensive user study was challenging. Secondly, the chosen population has similar education and background as data analysts in companies, so they are suitable for evaluation. Thirdly, the user study was intentionally focused on this population rather than spreading resources and efforts across numerous demographic groups. This enables future comparative analyses with other demographic groups. Examples of potentially relevant stakeholder groups for use cases II and III include shop-floor workers and supervisors.

The full results can be seen online[7]. Figure 1 shows the correlation between the calculated metrics and the user study results. Most metrics strongly correlate with multiple aspects of user perception, while the "Completeness" metric does not correlate with any aspect. Looking at Fig. 1, several insights about the validity of metrics regarding user perception can be gained. The performance of the decision rule does not correlate with perceived understandability or interpretability. The metrics parsimony and interpretability correspond to the user perception of understandability and interpretability and other aspects, i.e., if the interpretability value is high, the rule was rated as concise, consistent, complete, relevant, and believable. Most metrics correlate with multiple aspects of user perception. "I&P" has higher correlations than these metrics independently.

Analyzing the strongest correlation for each aspect of the user study, "I&P" best represents understandability, interpretability, relevance, believability, and consistency, with a correlation between 0.7 and 0.89, indicating a strong correlation. Perceived conciseness is best matched by "EC", exhibiting a correlation of 0.89. Lastly, "RC" best describes completeness with a correlation of 0.57.

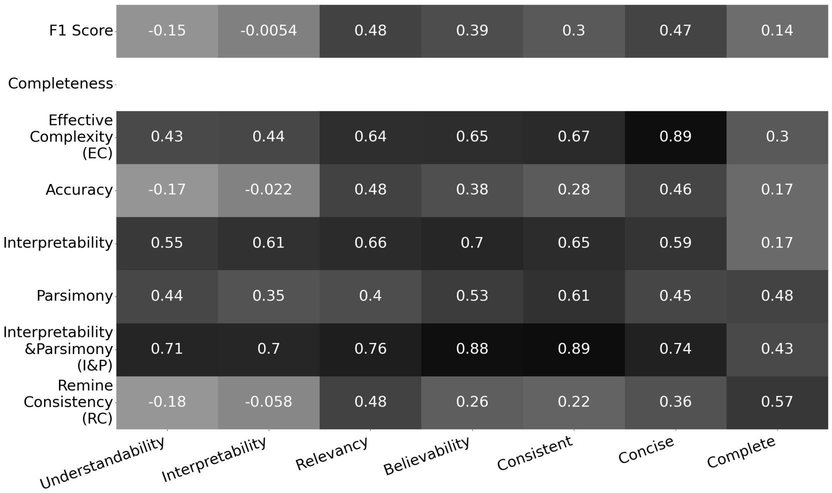

Fig. 1. Correlation between metrics (y-axis) and user perception (x-axis).

Comparing the results with the intended purposes of the metrics defined in Sect. 3, parsimony and interpretability metrics do correlate highly with perceived understandability and interpretability, especially when looking at the combined metric “I&P”. However, “RC” does not correlate strongly with consistency but rather with completeness. “RC” measures how much the textual rule changes when the model is newly discovered on the same input data; high values indicate that only slight changes occur with remining, indicating that the model contains all necessary conditions. The metric “Completeness” is not informative in this case, as only binary decision rules were discovered, and completeness equaled 1 for each case. It might be interesting for future work when more complex decision rules are analyzed. “EC” accurately depicts conciseness. High “EC” values indicate more straightforward rules, i.e., fewer conditions in a rule, making the contained conditions more impactful.

The criteria relevancy and believability had no associated metrics, as we argued that these are subjective. However, these also strongly correlate with “I&P”. The results show that “Interpretability& Parsimony” is the most informative metric, indicating that understandability and interpretability are essential for the user, i.e., if the user does not understand the decision rule, all other criteria cannot be evaluated adequately.

We can, therefore, conclude that four metrics, “Interpretability& Parsimony”, “Effective Complexity”, “Remine Consistency”’, and a performance-based metric, are well suited to draw meaningful conclusions about decision mining results. Therefore, we suggest including these four key metrics when analyzing decision mining results or selecting a suitable algorithm. The analysis shows that the metrics presented in the literature do not cover all aspects sufficiently. Specifically, the aspects covered by the interpretability metric and “RC” are not part of XAI metrics. In addition, “EC” was significantly changed compared to XAI.

In addition to the quantitative results, the comments have been analyzed. The following points were mentioned: A description of used attributes should be added, especially if engineered attributes are used. Users were confused about the measurement units, particularly if multiple values were part of one decision rule. If attributes are split into intervals, an explanation is needed, i.e., how many intervals exist, how many data points are contained in each interval, which intervals are relevant, etc.. Showing intervals as intervals, e.g., $10 <= x <= 20$, and not a combination of conditions, i.e., $x >= 10\ AND\ x <= 20$ is desired. Duplicate attributes feel redundant for users and make them question the correctness and completeness of the decision rule. Special characters have a strong negative impact on interpretability and understandability. Relational conditions might best represent the underlying business logic (i.e., attribute1 <= attribute2), but users mentioned it is hard to understand. However, a mix of relational attributes and constant values was especially hard to understand.

Guidelines: Based on the study insights, the following guidelines for decision rule discovery and representation are proposed: (I) Additional information about the used attributes should be included as part of the decision rule, e.g., explaining the attribute name or usage of intervals. (II) Engineered attribute names should be kept as simple as possible and should be explained. (III) Measurement units should be given (e.g., centimeters, minutes). (IV) Relational decision rules are complex to read and understand. A combination of relational rules and constant values should be avoided without explaining the attributes in depth. (V) Parsimony and interpretability can be the first indicators to check decision rules, as these are the essential preconditions for the user to benefit from the rule. (VI) Effective complexity can be used to check the decision rules concerning redundant conditions. (VII) Remine consistency allows for an additional “sanity check” for the decision model as it allows for an assessment of completeness.

5 Conclusion

This work analyzes metrics to evaluate decision mining results comprehensively. One of the main findings is that performance-based metrics do not automatically relate to valuable decision rules. Another main finding is that understandability and interpretability are essential for all other criteria and can be seen as a first indicator. Furthermore, additional information about the variable names and used units, especially when using engineering attributes or intervals, is essential for understandability. In general, four metrics, “Interpretability&Parsimony”, “Effective Complexity”, “Remine Consistency” and one performance-based metrics, can comprehensively evaluate decision mining results.

Limitations and Threats to Validity: So far, binary decision rules in an “IF-THEN” format have been studied. However, decision rules can include more than two classes and be visualized in tree or table form. A more extensive, quantitative evaluation should be part of future work, focusing on more aspects, e.g., the impact of the different formats. Furthermore, the user study only contained limited participants with similar backgrounds. Therefore, future work will include comparing different stakeholder groups to ensure the generalizability of the results. Moreover, quality metrics for process event logs are considered out-of-scope for this paper. However, the guidelines for log creation [1, 2] can be additionally followed to achieve valuable results.

In future work, we plan to expand the evaluation metrics to runtime decision mining and address challenges, such as data storage and outdated data.

References

1. van der Aalst, W.M.P.: Extracting Event Data from Databases to Unleash Process Mining. In: BPM - Driving Innovation in a Digital World, pp. 105–128 (2015)
2. Andrews, R., van Dun, C.G.J., Wynn, M.T., Kratsch, W., Röglinger, M.K.E., ter Hofstede, A.H.M.: Quality-informed semi-automated event log generation for process mining. Decision Support Systems **132** (2020)
3. Bazhenova, E., Haarmann, S., Ihde, S., Solti, A., Weske, M.: Discovery of Fuzzy DMN Decision Models from Event Logs. In: CAiSE. pp. 629–647 (2017)
4. Dean, P., Famili, A.: Comparative Performance of Rule Quality Measures in an Induction System. Applied Intelligence **7**(2), 113–124 (Apr 1997)
5. Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning (Mar 2017)
6. Dunkl, R., Rinderle-Ma, S., Grossmann, W., Fröschl, K.A.: A Method for Analyzing Time Series Data in Process Mining: Application and Extension of Decision Point Analysis. In: CAiSE Forum (Selected Extended Papers). pp. 68–84 (2014)
7. Ehrendorfer, M., Mangler, J., Rinderle-Ma, S.: Assessing the impact of context data on process outcomes during runtime. In: ICSOC. pp. 3–18 (2021)
8. Ehrlinger, L., Wöß, W.: A Survey of Data Quality Measurement and Monitoring Tools. Frontiers in Big Data **5** (2022)
9. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys **51**(5), 93:1–93:42 (Aug 2018)
10. Hassenstein, M.J., Vanella, P.: Data Quality—Concepts and Problems. Encyclopedia **2**(1), 498–510 (Mar 2022)
11. Heravizadeh, M., Mendling, J., Rosemann, M.: Dimensions of Business Processes Quality (QoBP). In: Business Process Management Workshops, pp. 80–91 (2009)
12. Hosseini, M., Shahri, A., Phalp, K., Ali, R.: Four reference models for transparency requirements in information systems. Requirements Eng. **23**(2), 251–275 (2018)
13. Islam, S.R., Eberle, W., Ghafoor, S.K.: Towards Quantification of Explainability in Explainable Artificial Intelligence Methods (Nov 2019)
14. Kahn, B., Strong, D., Wang, R.: Information Quality Benchmarks: Product and Service Performance. Commun. ACM **45**, 184–192 (Apr 2002)
15. Knight, S.A.: Developing a Framework for Assessing Information Quality on the World Wide Web. Informing Science: The International Journal of an Emerging Transdiscipline **8** (2005)
16. Koorn, J.J., Beerepoot, I., Dani, V.S., Lu, X., Weerd, I.v.d., Leopold, H., Reijers, H.A.: Bringing Rigor to the Qualitative Evaluation of Process Mining Findings: An Analysis and a Proposal. In: Process Mining. pp. 120–127 (2021)

17. Lee, Y.W., Strong, D.M., Kahn, B.K., Wang, R.Y.: AIMQ: a methodology for information quality assessment. Information & Management **40**(2), 133–146 (2002)
18. Leewis, S., Berkhout, M., Smit, K.: Future Challenges in Decision Mining at Governmental Institutions. In: AMCIS 2020 Proceedings. vol. 6 (2020)
19. Leite, J.C.S.d.P., Cappelli, C.: Software Transparency. Business & Information Systems Engineering **2**(3), 127–139 (Jun 2010)
20. de Leoni, M., van der Aalst, W.M.P.: Data-aware process mining: discovering decisions in processes using alignments. In: Symp. on Applied Comp. p. 1454 (2013)
21. de Leoni, M., Dumas, M., García-Bañuelos, L.: Discovering Branching Conditions from Business Process Execution Logs. In: Fundamental Approaches to Software Engineering. pp. 114–129 (2013)
22. de Leoni, M., Mannhardt, F.: Decision Discovery in Business Processes. In: Encyclopedia of Big Data Technologies, pp. 1–12 (2018)
23. Mannhardt, F., de Leoni, M., Reijers, H.A., van der Aalst, W.M.P.: Decision Mining Revisited - Discovering Overlapping Rules. In: Advanced Information Systems Engineering. pp. 377–392. Springer, Cham (Jun 2016)
24. Markus, A.F., Kors, J.A., Rijnbeek, P.R.: The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J. Biomed. Informatics **113**, 103655 (2021)
25. Miller, G.A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review **63**, 81–97 (1956)
26. Nguyen, A.p., Martínez, M.R.: On quantitative aspects of model interpretability (2020)
27. Rozinat, A.: Towards an Evaluation Framework for Process Mining Algorithms. In: BPM reports, vol. 0706. BPMcenter. org, Eindhoven (2007)
28. Scheibel, B., Rinderle-Ma, S.: Comparing decision mining approaches with regard to the meaningfulness of their results. arXiv:2109.07335 [cs] (Sep 2021)
29. Scheibel, B., Rinderle-Ma, S.: Decision Mining with Time Series Data Based on Automatic Feature Generation. In: Advanced Information Systems Engineering. pp. 3–18 (2022)
30. Scheibel, B., Rinderle-Ma, S.: Online Decision Mining and Monitoring in Process-Aware Information Systems. In: Conceptual Modeling. pp. 271–280 (2022)
31. Smedt, J.D., Hasić, F., vanden Broucke, S.K.L.M., Vanthienen, J.: Towards a Holistic Discovery of Decisions in Process-Aware Information Systems. In: Business Process Management. pp. 183–199 (2017)
32. Stevens, A., De Smedt, J.: Explainability in Process Outcome Prediction: Guidelines to Obtain Interpretable and Faithful Models (Dec 2022)
33. Ting, K.M.: Precision and Recall. In: Encyclopedia of Machine Learning, p. 781. Springer (2010)
34. Vanderfeesten, I., Cardoso, J., Mendling, J., Reijers, H., van der Aalst, W.M.P.: Quality Metrics for Business Process Models. IEEE Transactions on Software Engineering (2007)
35. Wang, R.Y., Strong, D.M.: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems **12**(4), 5–33 (1996)
36. Weller, A.: Transparency: Motivations and Challenges. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 23–40 (2019)
37. Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. Electronics **10**(5), 593 (Jan 2021)