Al-Farouk journal of sciences
Volume (2), Issue (4) (Spt, 2026)
ISSN: 3135-2359

مجلة الفاروق للعلوم
المجلد (2) العدد 4 سبتمبر 2026
ردمد: 3135-2359

A Leakage-Controlled Prefix-Based Association Rule Mining and Ranking Framework for Predictive Process Monitoring

Emhimmid Elsanosi Algazairi [1*], and Ahmed A Almantsri [2]

[1] Department of Electrical and Electronic Engineering, College of Technical Sciences, Misrata, Libya

[2] Computer Science Dept. Faculty of Arts & Science Kasr Khiar, Elmergib University –Libya

*Corresponding author: Em.Elsanosi@gmail,com

Received: 15/8/2026 Revised: 17/8/2026 Accepted: 05/9/2026 Publication: 04/110/2026

إطار منهجي لتنقيب قواعد الارتباط وترتيبها في المراقبة التنبؤية للعمليات بالاعتماد على بادئات مسارات التنفيذ وضبط تسرّب البيانات

إمحمد السنوسي القزيري[1*]، أحمد عيّاد المنتصري [2]

[1] قسم الهندسة الكهربائية والإلكترونية، كلية العلوم التقنية، مصراتة، ليبيا

[2] قسم الحاسوب، كلية الأداب والعلوم قصر الأخيار، جامعة المرقب، ليبيا

Abstract:

Predictive process monitoring aims to provide useful insights from ongoing cases to support decision-making. In this context, association rules express interpretable relationships between antecedents and consequents. However, mining retrospective event logs may introduce target leakage, while extensive rule collections can conceal patterns that remain stable over time. This study presents a leakage-controlled pipeline. The proposed pipeline first extracts case prefixes and constructs predictors using only information available within each prefix. It then mines target-constrained rules using FP-Growth on a chronological training partition and ranks the resulting rules using a development-only heuristic score.

FP-Growth generated 67 frequent itemset and 19 non-leaking target rules. Of these, 17 rules had a lift higher than one in validation, and 16 of these had a lift higher than one on the untouched test set. Acceptance rules showed the strongest associations. The combination of a medium prefix event volume and a low activity diversity resulted in a test confidence of 0.7047 and a lift of 1.3094 with a support of 0.0289. Cancellation rules showed more moderate associations but remained generally stable; the broadest cancellation rule achieved a test support of 0.2227 and a lift of 1.1028. By contrast, high-rework rules produced only marginal improvements over the 2.19% test prevalence and should therefore be interpreted as weak associations. This study provides a reproducible pipeline for association-rule mining. Its conclusions are limited to observed associations rather than classification accuracy or causal effects.

Keywords: predictive process monitoring; association rule mining; FP-Growth; event logs; target leakage; temporal validation; interpretable analytics; BPI Challenge 2017.

الملخص:

تهدف المراقبة التنبؤية للعمليات إلى استخلاص معلومات من الحالات قيد التنفيذ بما يدعم اتخاذ القرار. وتتيح قواعد الارتباط تمثيل العلاقات بين الشروط والنتائج بصورة قابلة للتفسير. غير أن تنقيب سجلات الأحداث بأثر رجعي قد يؤدي إلى تسرّب معلومات عن المتغير المستهدف، كما أن كثرة القواعد المستخرجة قد تحجب العلاقات التي تحافظ على استقرارها بمرور الزمن. تقدّم هذه الدراسة إطارًا لضبط تسرّب البيانات عند تنقيب قواعد الارتباط وترتيبها، يبدأ باستخراج بادئات مسارات التنفيذ وبناء المتغيرات التنبؤية حصرًا من المعلومات المتاحة ضمن كل بادئة. ثم تُستخدم

خوارزمية **FP-Growth** لتنقيب القواعد المقيّدة بالمتغيرات المستهدفة اعتمادًا على مجموعة تدريب محدّدة زمنيًا، وتُرتَّب القواعد الناتجة وفق مقياس استرشادي يُحتسب باستخدام بيانات التطوير وحدها.

أسفرت عملية التنقيب عن **67 مجموعة عناصر متكررة و19 قاعدة ارتباط مستوفية لضوابط عدم التسرّب**. وحقّقت 17 قاعدة معامل رفع يتجاوز الواحد في مجموعة التحقق، وحافظت 16 منها على ذلك في مجموعة الاختبار التي لم تُستخدم في تطوير الإطار. وأظهرت القواعد المرتبطة بقبول الطلبات أقوى الارتباطات؛ إذ حقّقت قاعدة تجمع بين عدد متوسط من الأحداث في بادئة المسار وانخفاض تنوّع الأنشطة درجة ثقة بلغت **0.7047**، ومعامل رفع **1.3094**، ودعمًا **0.0289** في الاختبار. أما قواعد إلغاء الطلبات، فأظهرت ارتباطات أكثر اعتدالًا مع استقرار عام عبر المجموعات الزمنية؛ وبلغ دعم قاعدة الإلغاء الأوسع تغطيةً **0.2227**، بمعامل رفع **1.1028**. في المقابل، لم تتجاوز قواعد ارتفاع إعادة العمل معدل انتشار هذه النتيجة في مجموعة الاختبار، البالغ **2.19%**، إلا بفارق محدود، مما يستدعي تفسيرها بوصفها ارتباطات ضعيفة.

تقدّم الدراسة إطارًا قابلًا لإعادة الإنتاج لتنقيب قواعد الارتباط وترتيبها مع ضبط تسرّب البيانات والتحقق من استقرارها الزمني. وتقتصر دلالات النتائج على الارتباطات المرصودة، ولا تمثّل دليلًا على دقة التصنيف أو على وجود علاقات سببية.

الكلمات المفتاحية: المراقبة التنبؤية للعمليات؛ تنقيب قواعد الارتباط؛ FP-Growth؛ سجلات الأحداث؛ تسرّب معلومات المتغير المستهدف؛ التحقق الزمني؛ التحليلات القابلة للتفسير ؛ 2017 BPI Challenge.

I. Introduction

Modern information systems record the execution of business processes for data analysis. These records are called event logs, in which each event is linked to a specific case, an activity, and a timestamp. Predictive process monitoring (PPM) is a part of process mining that uses historical event logs to predict the outcomes of ongoing cases. The value of PPM lies in providing information at an early stage to support timely intervention rather than merely describing a completed process. Recent studies indicate that PPM increasingly relies on complex learning architectures whose lack of transparency may hinder interpretation and practical adoption (Neu et al., 2022; Mehdiyev et al., 2025).

Association rule mining (ARM) provides a complementary and interpretable representation of process behavior. A rule expressed as $X \Rightarrow Y$ allows analysts to examine the antecedent conditions and the associated outcome. ARM was originally introduced for transactional data (Agrawal et al., 1993). FP-Growth improved frequent-pattern discovery by using an FP-tree structure and avoiding the explicit generation of candidate itemset (Han et al., 2004). Rule-based outcome models have also attracted attention in PPM because their global and local reasoning is generally easier to explain than the internal logic of black-box models (Lee et al., 2022; Khan et al., 2022).

However, ARM can introduce a risk of target leakage when applied to event logs. If a transaction includes events or attributes that occur after the prediction point, a rule may appear strong simply because it relies on information that would not have been available when the decision was required. This temporal leakage can undermine the reliability of the resulting rules. A second challenge arises from distributional change, as a relationship identified in earlier cases may weaken or even reverse in later cases. For these reasons, the present study establishes a fixed operational milestone, applies preprocessing based exclusively on the training data, uses chronological development partitions, and reserves an untouched test partition for final evaluation.

This paper does not propose a supervised classifier or a ranking model. The proposed approach is instead a pipeline for prefix-based association-rule mining with a leakage-controlled heuristic ranking mechanism only on development data only. The ranking score is a combination of the training support, training lift and the consistency of the rule confidence between training and validation partitions. The score is computed directly from these measures and does not use any test-set outcome information.

The first contribution of this study is the definition of case prefixes that end at the first completed offer-sent event, with only information available at or before that milestone admitted to rule antecedents. The second contribution is leakage controlled target construction, where final outcomes and post-cutoff rework are only kept as consequents, and preprocessing thresholds are learned only on training cases. Third, rules are checked on the chronological order. Rules are mined from the first 70% of cases, screened on the next 15% and evaluated on the last 15%, keeping the temporal order. The final contribution is transparent ranking and complete reporting: the

heuristic score is explicitly defined, the test set is excluded from ranking, and both successful and failed generalization results are reported.

II. Literature Review

2.1 Association Rule Mining and Interestingness

Association rule mining (ARM) finds combinations of items that occur frequently enough to meet a pre-specified support threshold and then it describes the directional relationships between antecedent and consequent itemsets. Support, confidence and lift are the three pillars of rule evaluation, each capturing a different aspect of the relationship. Support measures the co-occurrence of the antecedent and consequent and confidence measures the conditional probability of finding the consequent given the antecedent. Lift goes a step further by comparing this conditional probability with the overall probability of the consequent.

In the current study, lift is particularly useful because the three target outcomes have substantially different base rates. A frequent outcome may have a high confidence value, but the value may not add much information beyond the frequency of the outcome itself. Conversely, a low confidence value for a rare target might still represent an increase from its baseline. However, no single interestingness measure is sufficient to establish causality or practical usefulness. Furthermore, patterns with frequent occurrence might not be informative if their perceived strength is mainly due to the high marginal prevalence of the individual items (Petitjean et al., 2016).

Apriori repeatedly generates and evaluates candidate itemsets. FP-Growth, in contrast, compresses transactions into an FP-tree and mines conditional pattern bases without explicitly generating candidate sets (Han et al., 2004). This tree-based structure supports efficient frequent-pattern discovery and is used here to mine the compact Boolean transactions. Incremental extensions of FP-Growth aim to maintain frequent patterns efficiently as data evolve. However, these extensions pursue objectives that differ from the fixed historical analysis implemented in this study (Alsaeedi & Alhegami, 2022).

2.2 Predictive Process Monitoring and Event-Log Representation

The event log is a time ordered mapping of the execution of business processes. It is organized in traces, each one containing the events in temporal order belonging to a single process case. Predictive process monitoring converts a partial trace, called trace prefix, into a set of features that can be associated to a future target. The literature provides a wide variety of encoding methods, such as aggregation-based features, index-based sequences, tensors and learned representations. However, reviews of deep learning methods for PPM show considerable differences in preprocessing, prefix generation, target definition, and evaluation procedures. Such variation can blur the basis for direct comparison across studies (Neu et al., 2022). Similarly, research on event-log sampling emphasizes that decisions concerning case selection and data representation can influence the resulting predictive behaviour (Fani Sani et al., 2023).

The present method uses a milestone based prefix rather than using multiple fixed prefix lengths. The first completed event that indicates an offer has been sent will determine the cut-off for all qualifying cases. This milestone provides a uniform business interpretation: the antecedent describes the process state up to and including the first communication of an offer while the consequent describes either a later case outcome or rework after the cut-off. This cutoff distinguishes the information available at the prediction point from future information that must remain outside the rule antecedent.

2.3 Explainability, Interpretability, and Trust

Explainable predictive process monitoring can rely on intrinsically interpretable models or on post-hoc explanations of otherwise opaque predictors. This distinction matters because an explanation should be faithful, meaning it reflects the mechanism that produced the prediction, and interpretable in the relevant process context (Stevens & De Smedt, 2024; Mehdiyev et al., 2025). Explainability in outcome-oriented predictive process monitoring has also been quantified using dedicated metrics (Stevens et al., 2022). User evaluations further suggest that the usefulness of an explanation varies with how it is presented and with the technical expertise of its users (Rizzi et al., 2024).

Association rules offer an intuitive, transparent insight into the relationships within the data. They represent global patterns, not local approximations from a separate black-box model. (Lee et al., 2022) showed that rule-based classification can enable interpretable, outcome-oriented PPM on several real-world event logs, but also revealed considerable trade-offs between predictive performance and representation. The present study shares this emphasis on interpretability but differs in its methodology. Instead of training a rule-based classifier, it mines and ranks descriptive predictive associations. The difference is between rules that classify individual cases and rules that describe interpretable patterns at the population level.

Table 1. Positioning of Explainability Approaches in Predictive Process Monitoring

Approach	Typical output	Strength	Limitation for this study
Deep sequence model	Outcome or next-event probability	Models complex temporal patterns	Requires an explanation layer; not implemented
Post-hoc explanation	Feature contribution or local surrogate	Explains an existing black-box predictor	Faithfulness and user interpretation must be assessed
Rule-based classifier	Ordered decision rules	Intrinsically interpretable prediction	Requires classifier training and predictive evaluation
Association-rule pipeline	$X \Rightarrow Y$ with support, confidence, lift	Transparent population-level relations	Associational, potentially redundant, not causal

Table 1 summarizes common approaches to explainability in predictive process monitoring. This study uses the fourth approach, an association-rule pipeline, while the other rows summarize related work discussed in Lee et al. (2022), Stevens and De Smedt (2024), Rizzi et al. (2022), Mehdiyev et al. (2023), and Stevens et al. (2022).

2.4 Ranking, Temporal Reliability, and Research Gap

Finding useful patterns among many rules can be hard without a clear ranking policy for large sets of association rules. Support ranking tends to favor patterns that occur broadly. The confidence ranking may bias toward consequents with a high baseline prevalence. Lift compares confidence with the baseline rate of consequent. However, the lift-based ranking can favor very specific rules that apply only to a small number of cases. A transparent composite score could help balance coverage and relative association strength, provided each score component and each step in rule selection are clearly reported.

Also, a ranking policy for predictive process monitoring should consider the temporal structure of the data. Validation on later cases, instead of randomly mixing early and late cases, provides a more realistic assessment of whether relationships identified in the past continue to be informative over time. Previous research regarding predictive process monitoring has explored interpretability, uncertainty, event-log sampling, and compliance-oriented monitoring (Rinderle-Ma et al., 2023; Mehdiyev et al., 2023).

The pipeline developed in this study includes the following methodological safeguards: a milestone-based cutoff, target-excluded antecedents, training-only preprocessing, chronological validation, and ranking independent of the test set. Together these safeguards work to reduce the risk of information leakage and overly optimistic rule selection. Extensions include diversity-aware re-ranking and supervised learning-to-rank. These are directions for future research and were not implemented as parts of the pipeline evaluated in this study.

III. Methodology

A. Research Design and Scope

In this work, an artifact-centered empirical design is employed based on a reproducible Python pipeline. The pipeline transforms the event log into Boolean transactions at case level, mines association rules with predefined constraints, ranks the obtained rules using the development data and evaluates their behavior on later cases in time. The unit of analysis is an individual loan application. The method is descriptive and evaluated on later cases. It does not estimate causal effects and it does not assign a single predicted class to each case. Figure 1 shows the implemented stages and the respective case and rule counts.

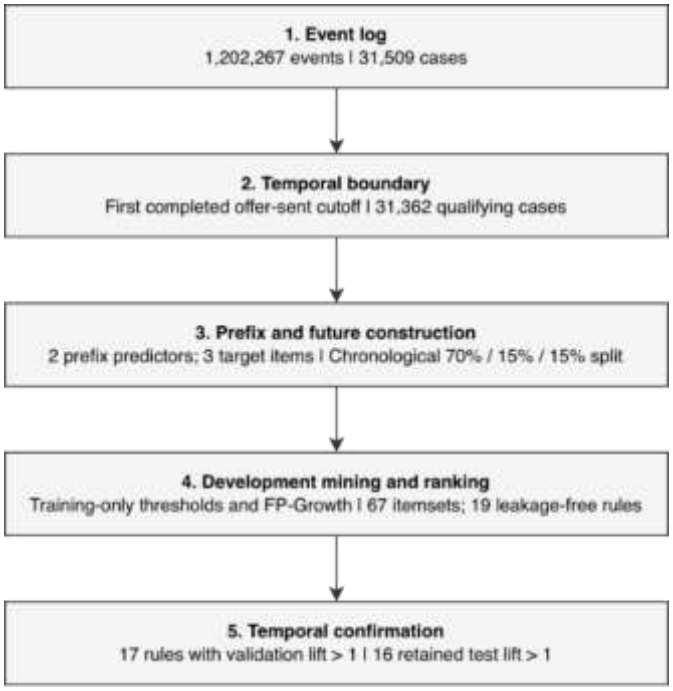


Figure 1. Implemented leakage-controlled prefix association-rule pipeline.

B. Dataset

The public BPI Challenge 2017 event log is used as the empirical basis for this evaluation. It covers loan applications submitted to a Dutch financial institution in 2016 and their processing up to early February 2017 (van Dongen, 2017). The official dataset contains 1,202,267 events, 31,509 loan applications, 42,995 offer records, and 149 event originators. The event log distinguishes between application state changes, offer state changes, and workflow-related events.

The attributes available at the application level are the amount requested, application type, loan goal, and application identifier. The attributes at the offer level are the amount offered, initial withdrawal, repayment terms, monthly cost, credit score, offer creator, selection status, and acceptance indicators. However, after conversion to a cleaned CSV representation, the pipeline retained only the attributes needed to reconstruct and order the process events. These were the case identifier, activity name, timestamp, lifecycle transition, and stable event identifier. Together, these fields allow each event to be linked to its case and the chronological event sequence of each loan application to be reconstructed.

Table 2. BPI Challenge 2017 Data and Analytic Sample

Element	Value	Use in this study
Original cases	31,509	Chronologically ordered loan applications
Original events	1,202,267	Sorted by case, timestamp, and event ID
Offer records	42,995	Dataset context; not used as a predictor count
Originators	149	Dataset context; resource identity not encoded
Distinct activity names in cleaned file	26	Basis for prefix activity counts
Cases with offer-sent cutoff	31,362 (99.53%)	Final case-level analytic sample
Cases without cutoff	147 (0.47%)	Excluded because the prediction milestone was absent

Source: van Dongen (2017)

C. Event Ordering and Prefix Cutoff

All timestamps have been parsed in UTC. The events were sorted stably by case identifier, time stamp and event identifier to ensure a deterministic and reproducible order of the traces. For each case, the cut-off was defined as the first completed instance of O_Sent (mail and online) or O_Sent (online only). This cutoff effectively divided each case into a prefix and a future segment, where the prefix consisted of all events with timestamps less than or equal to the cutoff time, and the future segment contained only events that occurred strictly after it. Cases without a qualifying cutoff were excluded because they could not be assigned a prefix ending at the specified offer-sent milestone.

This study intentionally picked narrow prefix feature space by only computing two features: (1) `prefix_event_count` is the total count of events observed up to the cutoff, and (2) `prefix_unique_activities` is the

count of distinct activity names in the prefix. These coarse-grained features were chosen because they make rule antecedents concise and easy to interpret. The prefix had a strict information boundary; no final outcome, future activity, post-cutoff timestamp, or future-rework variable was permitted in the antecedent feature space. This constraint is a key part of the leakage-control mechanism of the pipeline.

D. Outcome and Future-Rework Labels

The final case outcome was inferred from activities in the complete qualifying case using an explicit priority order: O_Accepted defined accepted; otherwise, A_Cancelled defined cancelled; otherwise, A_Denied defined denied; remaining cases were labeled other. Accepted and cancelled were used as rule consequents. Denied and other cases remained in the transaction population, preserving the base rates in the qualifying case population, but were not modeled as separate consequents in this implementation.

Future rework was derived from completed workflow activities after cutoff. For each case and each activity with name starting W_ the first completed execution was considered to be the normal occurrence and each additional completed execution added one rework unit. The case score was thus:

$$\textit{Future rework score}(c) = \Sigma_a \max[n\text{complete}(c, a) - 1, 0], \textit{ for post-cutoff W_ activities } a.$$

Any score above zero defined future rework. The high-rework threshold was calculated from positive training scores only, by rounding the 75th percentile of the scores up to the nearest integer with a minimum threshold of one. The procedure yielded a threshold of two; therefore, FUTURE_HIGH_REWORK equaled one when the future rework score was at least two.

E. Chronological Split and Training-Only Discretization

The cases were chronologically sorted by start time and partitioned into 70% training, 15% validation and 15% test sets. Only the training cases were used to estimate the quartile cut-offs. Prefix event count was coded as LOW for up to 10, MEDIUM for 11-14, HIGH for above 14. Prefix unique activities was coded as LOW for values up to 7, MEDIUM for 8–9, and HIGH for values above 9. The fixed thresholds were then used unchanged for the validation and test cases. Figure 2 shows the differences in target prevalence in these partitions.

Table 3. Chronological Splits and Target Prevalence

Split	Cases	Accepted	Cancelled	High rework	Any rework
Training	21,953	55.25%	33.27%	3.56% (782)	13.46% (2,954)
Validation	4,704	54.57%	33.01%	1.72% (81)	7.93% (373)
Test	4,705	53.82%	31.12%	2.19% (103)	9.12% (429)

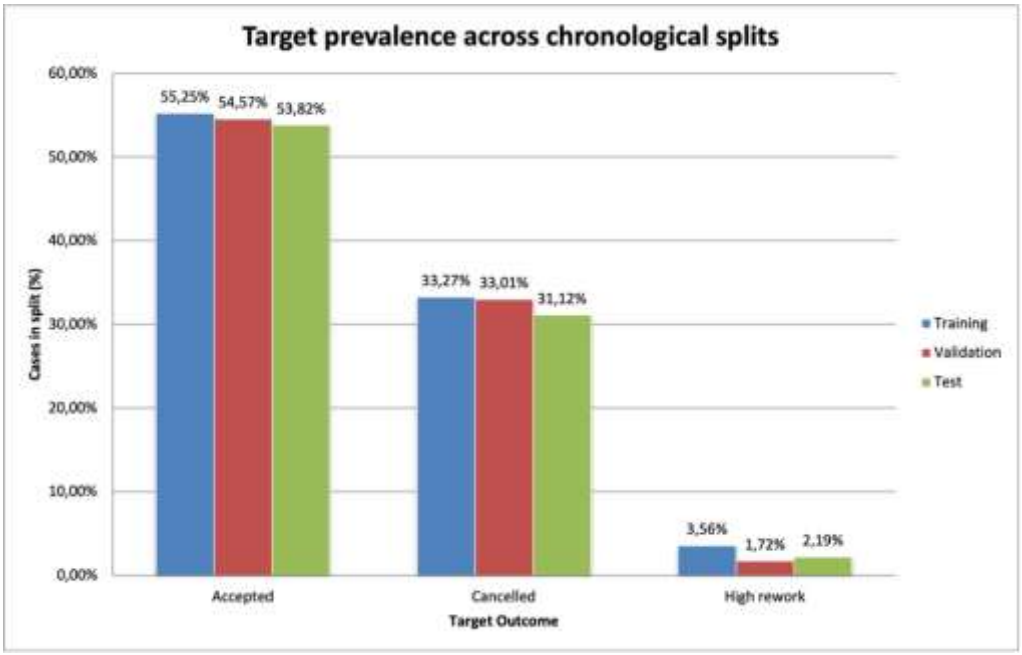


Figure 2. Target prevalence across the chronological training, validation, and test splits.

F. Boolean Transaction Encoding and Rule Mining

Each case was converted into a Boolean transaction with six possible predictor items (three mutually exclusive bins for prefix event count and three mutually exclusive bins for unique activity count) and three possible target items: OUTCOME_ACCEPTED, OUTCOME_CANCELLED and FUTURE_HIGH_REWORK. There was one predictor item per transaction in each bin set. We only applied FP-Growth to the training transactions with minimum support of 0.001 and a maximum itemset length of four. Association rules were generated from the

frequent itemsets and only kept them if the consequent was exactly one target item, the antecedent was non-empty and did not contain any target item and the training lift was greater than one.

$$support(X \Rightarrow Y) = |X \cap Y| / N$$
$$confidence(X \Rightarrow Y) = support(X \cap Y) / support(X)$$
$$lift(X \Rightarrow Y) = confidence(X \Rightarrow Y) / support(Y)$$

Support, confidence and lift were recalculated independently for the training, validation and test partitions. A lift greater than one was used as a flag for a positive association within each partition, meaning that the rate of the consequent among cases satisfying the antecedent is greater than the baseline rate for that consequent in the partition. This flag is not a test for statistical significance.

G. Development-Only Ranking

Only training and validation data were used for ranking. The confidence stability for each rule was computed by taking the NumPy population standard deviation of the training and validation confidence values. The score implemented was:

$$Stability(r) = max\{0, 1 - SD[Conftrain(r), Confvalidation(r)]\}$$
$$IntelligentScore(r) = Supporttrain(r) \times Lifttrain(r) \times max\{Stability(r), 0.10\}$$

Rules were ranked separately for each consequent in decreasing order of heuristic score with validation lift only used as a tie-breaker. Confidence and lift of test were not included in the score or ranking. Hence, the test partition was only used for final evaluation on temporally later cases.

Table 4. Implemented Pipeline Configuration

Stage	Implemented setting	Leakage control
Prediction point	First completed offer-sent event	Future defined strictly after cutoff
Predictors	Prefix event count; prefix unique activities	Computed at or before cutoff
Discretization	Training Q25/Q75	Validation/test do not set thresholds
Rework threshold	ceil(Q75 of positive training scores) = 2	No validation/test target information
Frequent patterns	FP-Growth; min support 0.001; max length 4	Training transactions only
Rule constraints	Singleton target consequent; clean antecedent; lifttrain > 1	Targets never occur on the left
Ranking	Supporttrain × lifttrain × stability	Test excluded from score and rank
Confirmation	liftvalidation > 1, then lifttest > 1	Temporal partitions evaluated independently

IV. Results and Data Analysis

A. Data Preparation and Outcome Distribution

The cut-off procedure retained 31,362 out of 31,509 cases. Among retained cases 17228 (54.93%) were accepted, 10321 (32.91%) were canceled, 3715 (11.85%) were denied and 98 (0.31%) were other. This final-outcome composition is mutually exclusive and shown in Figure 3. The prevalence of high-rework was not consistent across temporal partitions (3.56% training, 1.72% validation, and 2.19% test). Small changes in confidence are important to interpreting these baseline differences.

Table 5. Outcome and Future-Rework Summary for 31,362 Qualifying Cases

Measure	Training	Validation	Test	Overall/definition
Accepted outcome	55.25%	54.57%	53.82%	17,228 (54.93%)
Cancelled outcome	33.27%	33.01%	31.12%	10,321 (32.91%)
Denied outcome	—	—	—	3,715 (11.85%); not a consequent
Other outcome	—	—	—	98 (0.31%); not a consequent
Mean rework score	0.1839	0.1010	0.1214	Repeated completed W_ events
Maximum rework score	8	5	5	Partition maximum
High rework	782 (3.56%)	81 (1.72%)	103 (2.19%)	Score ≥ 2

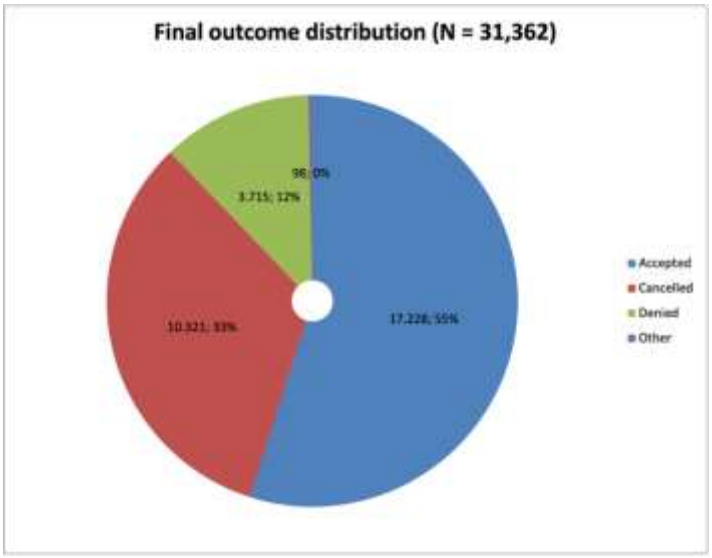


Figure 3. Final outcome distribution among qualifying cases.

B. Mining Yield and Temporal Screening

From the 21,953 training cases, FP-Growth generated 67 frequent itemsets. After the generation of association rules and the application of the target, antecedent and training-lift constraints, 19 target rules were remaining who satisfied these constraints: seven for acceptance, seven for Cancellation and five for future high rework. Seventeen rules for which lift in validation was > 1. Sixteen of these rules also maintained lift above one in the test partition.

Table 6. Rule Counts through the Evaluation Pipeline

Consequent	Training-Qualified	Validation lift > 1	Test-confirmed among validation-Qualified
OUTCOME_ACCEPTED	7	6	6
OUTCOME_CANCELLED	7	7	7
FUTURE_HIGH_REWORK	5	4	3
Total	19	17	16

Note. A rule is called test-confirmed only in the directional sense that lift remained above one; confidence intervals and significance tests were not computed.

The validation screen did not pass two rules:

1. PREFIX_EVENTS_HIGH ⇒ FUTURE_HIGH_REWORK PREFIX_EVENTS_MED
2. PREFIX_UNIQUE_ACTIVITIES_HIGH ⇒ OUTCOME_ACCEPTED.

While A third rule, PREFIX_EVENTS_HIGH AND PREFIX_UNIQUE_ACTIVITIES_MED ⇒ FUTURE_HIGH_REWORK, passed validation with lift 1.0491 but had test lift below one, 0.8413. This reversal emphasizes the significance of evaluating ranked rules on an independent test partition.

C. Accepted-Outcome Rules

Acceptance rules showed the highest lift values among the target groups. Low unique-activity diversity alone had a test confidence of 63.21%, compared with a 53.82% acceptance baseline, and a test lift of 1.1747. The conjunction of medium event count and low unique-activity diversity had the highest test lift, 1.3094, and confidence, 70.47%, but a test support of only 2.89%. This contrast illustrates the trade-off between coverage and relative association strength reflected in the ranking score.

Table 7. Development-Valid Rules for OUTCOME_ACCEPTED

Antecedent	Score	Supp.test	Conf.test	Liftval	Lifttest	Status
Unique activities LOW	0.266760	0.2140	0.6321	1.1735	1.1747	Confirmed
Events LOW	0.246245	0.1940	0.6219	1.1333	1.1557	Confirmed
Events LOW + unique activities LOW	0.212179	0.1762	0.6247	1.1573	1.1609	Confirmed
Events MED + unique activities LOW	0.043890	0.0289	0.7047	1.2183	1.3094	Confirmed
Events LOW + unique activities MED	0.034386	0.0179	0.5957	1.0152	1.1070	Confirmed
Events HIGH + unique activities LOW	0.009872	0.0089	0.5753	1.4522	1.0691	Confirmed

Note. Rules are ordered by the development-only score. MED denotes 11–14 prefix events or 8–9 unique activities, depending on the item.

D. Cancelled-Outcome Rules

Cancellation associations were more moderate, but all seven rules retained from training maintained lift above one in both later partitions. Medium unique-activity diversity had the highest test support: it occurred jointly with cancellation in 22.27% of test cases, with confidence of 34.32% and lift of 1.1028 against a 31.12% cancellation baseline. High event volume combined with medium activity diversity reached confidence of 35.32% and lift of 1.1351 at 6.93% joint support. The two high-diversity rules had higher test lift but very low support, indicating limited coverage.

Table 8. Development-Valid Rules for OUTCOME_CANCELLED

Antecedent	Score	Supp.test	Conf.test	Liftval	Lifttest	Status
Unique activities MED	0.266722	0.2227	0.3432	1.1169	1.1028	Confirmed
Events MED	0.162112	0.1571	0.3371	1.0742	1.0835	Confirmed
Events MED + unique activities MED	0.158334	0.1460	0.3452	1.1194	1.1095	Confirmed
Events HIGH	0.099469	0.0765	0.3445	1.1229	1.1071	Confirmed
Events HIGH + unique activities MED	0.093541	0.0693	0.3532	1.1686	1.1351	Confirmed
Unique activities HIGH	0.004934	0.0045	0.3621	1.2910	1.1636	Confirmed
Events HIGH + unique activities HIGH	0.004521	0.0038	0.3673	1.2511	1.1806	Confirmed

Note. Support is the joint prevalence of antecedent and cancellation; confidence is the cancellation rate conditional on the antecedent.

E. Future High-Rework Rules

Future high rework was rare, and its rules showed weak associations. Among rules with test lift above one, the highest test confidence was 2.38% for low event count, only 0.19 percentage points above the test baseline of 2.19%, with a lift of 1.0891. All high-rework rules that passed validation had test support below 0.8%. Figure 4 compares these rules with the acceptance and cancellation rules, showing the trade-off between joint support and lift. These results represent weak population-level signals and do not warrant describing the rules as strong predictors of rework.

Table 9. Development-Valid Rules for FUTURE_HIGH_REWORK

Antecedent	Score	Supp.test	Conf.test	Liftval	Lifttest	Status
Unique activities LOW	0.015745	0.0077	0.0226	1.2326	1.0323	Confirmed
Events LOW	0.015156	0.0074	0.0238	1.1949	1.0891	Confirmed
Events LOW + unique activities LOW	0.014255	0.0066	0.0234	1.1911	1.0671	Confirmed
Events HIGH + unique activities MED	0.008173	0.0036	0.0184	1.0491	0.8413	Failed test

Note. The failed-test rule was retained in the table for scientific transparency. High rework denotes at least two repeated completed post-cutoff workflow executions.

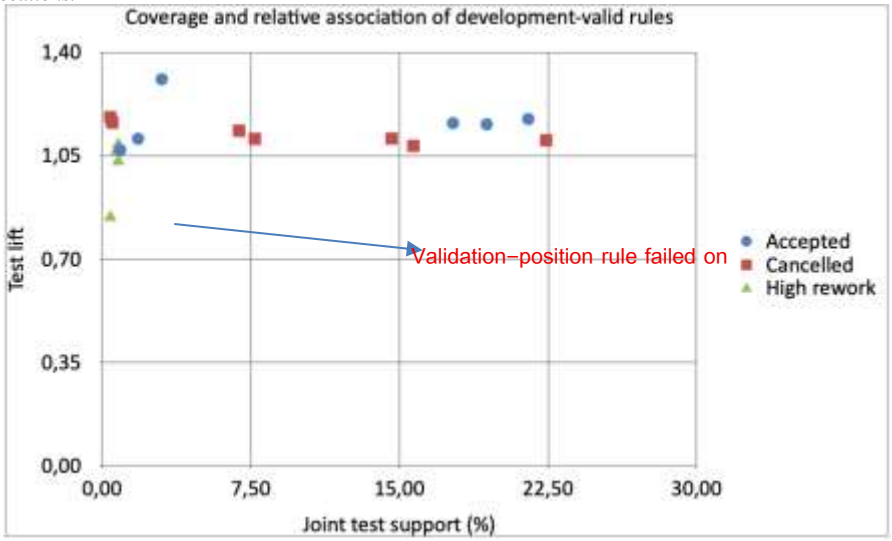


Figure 4. Shows Test-set coverage and lift of the 17 development-valid rules.

Note. Rule identifiers A1–A6, C1–C7, and R1–R4 follow Tables 7–9. The dashed line marks lift = 1. R4 is the validation-positive high-rework rule that failed test confirmation. Support is expressed as a percentage of all test cases.

F. Interpretation of Support, Confidence, and Lift

For PREFIX_UNIQUE_ACTIVITIES_LOW ⇒ OUTCOME_ACCEPTED, test support of 0.2140 means that 21.40% of all test cases satisfied both the antecedent and acceptance. It does not mean that the conditional probability of acceptance was 21.40%; the conditional probability is measured by confidence, which was 63.21%.

The lift of 1.1747 means that the conditional acceptance probability was approximately 17.47% higher in relative terms than the 53.82% test baseline.

For PREFIX_EVENTS_MED AND PREFIX_UNIQUE_ACTIVITIES_LOW ⇒ OUTCOME_ACCEPTED, test confidence was 70.47%. Relative to the test baseline, the absolute difference was 16.65 percentage points and the relative increase was 30.94%. Joint support was 2.89%, corresponding to approximately 136 of 4,705 test cases. This rule is therefore stronger in relative association but narrower in coverage than the first-ranked acceptance rule.

V. Discussion

Three major conclusions are supported by the empirical results. First, simple features of prefix-complexity are clearly linked to final acceptance and Cancellation. Medium activity diversity and higher event volume were more often associated with Cancellation while acceptance was associated with low activity diversity. Second, most of the development-valid outcome rules were in the same direction of association in the later test period: all six validation-valid acceptance rules and all seven Cancellation rules had test lift above one. Third, the high-rework associations were rare, unstable and not suitable for strong operational claims.

The main contribution of this study is methodological integration instead of a new frequent-itemset algorithm. The pipeline incorporates an operationally meaningful prefix, training-only target thresholds and discretization, constrained rule direction, chronological validation, and an explicitly defined heuristic ranking formula. It does so by instantiating these methodological choices in an executable ARM workflow, building on prior work on interpretability and the use of process context (Stevens & De Smedt, 2024; Lee et al., 2022).

It's important to distinguish between an association-rule pipeline and a classifier. The confidence values provided are conditional frequencies for subsets. There is no prediction per case, no confusion matrix, no AUC or F1 score. Similarly, the heuristic IntelligentScore is not trained on human relevance labels and should not be referred to as learning-to-rank. The goal is to rank rules in a transparent manner based on training support, relative association and confidence consistency between training and validation.

Acceptance and Cancellation rules may be exploratory monitoring cues. For example, a case with a medium number of events and a small set of distinct activities by the first offer-sent milestone is associated with a subgroup with a higher observed acceptance rate. Conversely, a subgroup with a moderately increased Cancellation rate is identified by high event volume combined with medium activity diversity. These descriptive associations may inform future investigation, but do not warrant automated intervention without richer features and prospective validation.

The high-rework results are not operationally compelling. Lift is greater than one but going from 2.19% baseline to 2.38% conditional probability is not much. Signal strength may be limited by class imbalance, coarse representation of antecedents, and differences in temporal prevalence. Before deployment, A useful extension would be uncertainty-aware monitoring (Mehdiyev et al., 2023).

First, the antecedents rely on only two prefix counts. These features do not specify particular activities, resources, durations, amounts, or process paths, so operational specificity is limited. In addition, denied and other cases were included in the population denominators, but not modeled as target consequents. Therefore, the study does not provide rules for every final status.

In addition, rework is operationalized as repeated completed W_ activities after the cutoff. Other forms of rework or correction may not be captured.

The lift-above-one flags indicate direction of association only and are the basis for rule evaluation. Confidence intervals were not estimated, multiplicity corrections were not performed and significance tests were not performed, so low-support rules should be interpreted with particular caution.

The IntelligentScore is heuristic and has not been validated against expert relevance judgments. Its stability term only uses confidence on training and validation, not multiple rolling windows or statistical uncertainty. The assessment includes one business milestone and financial process. Generalization to other organizations and domains, and other cutoff definitions, is unvalidated. Association rules do not establish that the outcome of acceptance, Cancellation, or rework would change if the event counts or the activity diversity were changed.

VI. Future Work

Future work will investigate more expressive prefix encodings including elapsed time, activity identities, offer and application attributes, resources and stage-aware features that respect the cutoff boundary. Rolling-origin evaluation, bootstrap confidence intervals and multiple-comparison controls will also be used to distinguish stable associations from sampling variation. We will also consider rare-target modeling for evaluating imbalance-aware rule criteria and alternative definitions of rework, estimating thresholds from training data only.

In future work we will also validate the ranking by obtaining expert relevance judgments or task-based utility labels and then compare heuristic ranking with learning-to-rank using evaluation metrics like NDCG or MAP. Instead of assuming that ranking is sufficient to remove overlap, we will consider closed and maximal itemsets, subsumption pruning, and MMR for diversity and redundancy control. Finally, Future works will assess external

validity via the pipeline on more public logs and prospective validity by checking whether selected rules remain informative for future cases.

VII. Conclusion

In this paper, we introduce a leakage-controlled prefix-based pipeline for mining and ranking association rules for predictive process monitoring. The pipeline was applied to 31,362 BPI Challenge 2017 loan applications, each cut at the first offer-sent event. It generated 19 candidate rules from the training partition; 17 of these had a lift above one in validation, and 16 of those maintained it on the untouched test partition. Because the rules describe associations rather than causes, they suit exploratory monitoring rather than automated intervention.

The strength of evidence differed according to the target. The acceptance rules were the clearest. A medium prefix event count with low activity diversity achieved a test confidence of 70.47% and a lift of 1.3094, but only 2.89% of test cases satisfied it. Cancellation rules were weaker but broader, with medium activity diversity covering 22.27% of the test cases at lift 1.1028. The high-rework rules were close to the 2.19% test baseline, with one rule reversing direction, going from a validation lift of 1.0491 to a test lift of 0.8413.

Overall, two coarse counts available at the offer-sent milestone were sufficient to carry a directional signal for acceptance and Cancellation, which survived a chronological split, but this did not hold for rework.

References

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (pp. 207–216). ACM. https://doi.org/10.1145/170035.170072

Alsaeedi, H. A., & Alhegami, A. S. (2022). An incremental interesting maximal frequent itemset mining based on FP-Growth algorithm. Complexity, 2022, 1942517. https://doi.org/10.1155/2022/1942517

Fani Sani, M., Vazifehdoostirani, M., Park, G., Pegoraro, M., van Zelst, S. J., & van der Aalst, W. M. P. (2023). Performance-preserving event log sampling for predictive monitoring. Journal of Intelligent Information Systems, 61, 53–82. https://doi.org/10.1007/s10844-022-00775-9

Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery, 8(1), 53–87. https://doi.org/10.1023/B:DAMI.0000005258.31418.83

Khan, N., Tariq, Z., Ali, A., McClean, S., Taylor, P., & Nauck, D. (2022). Early prediction of complex business processes using association rule based mining. In K. C. Santosh, R. Hegadi, & U. Pal (Eds.), *Recent trends in image processing and pattern recognition* (pp. 183–194). Springer. https://doi.org/10.1007/978-3-031-07005-1_17

Lee, S., Comuzzi, M., & Kwon, N. (2022). Exploring the suitability of rule-based classification to provide interpretability in outcome-based process predictive monitoring. Algorithms, 15(6), 187. https://doi.org/10.3390/a15060187

Mehdiyev, N., Majlatow, M., & Fettke, P. (2023). Quantifying and explaining machine learning uncertainty in predictive process monitoring: An operations research perspective. arXiv. https://doi.org/10.48550/arXiv.2304.06412

Mehdiyev, N., Majlatow, M., & Fettke, P. (2025). Interpretable and explainable machine learning methods for predictive process monitoring: A systematic literature review. Artificial Intelligence Review, 58, 378. https://doi.org/10.1007/s10462-025-11399-0

Neu, D. A., Lahann, J., & Fettke, P. (2022). A systematic literature review on state-of-the-art deep learning methods for process prediction. Artificial Intelligence Review, 55, 801–827. https://doi.org/10.1007/s10462-021-09960-8

Petitjean, F., Li, T., Tatti, N., & Webb, G. I. (2016). SkOPUS: Mining top-k sequential patterns under leverage. Data Mining and Knowledge Discovery, 30, 1086–1111. https://doi.org/10.1007/s10618-016-0467-9

Rinderle-Ma, S., Winter, K., & Benzin, J.-V. (2023). Predictive compliance monitoring in process-aware information systems: State of the art, functionalities, research directions. arXiv. https://doi.org/10.48550/arXiv.2205.05446

Rizzi, W., Comuzzi, M., Di Francescomarino, C., Ghidini, C., Lee, S., Maggi, F. M., & Nolte, A. (2024). Explainable predictive process monitoring: A user evaluation. Process Science, 1, Article 3. https://doi.org/10.1007/s44311-024-00003-3

Stevens, A., & De Smedt, J. (2024). Explainability in process outcome prediction: Guidelines to obtain interpretable and faithful models. European Journal of Operational Research, 317(2), 317–329. https://doi.org/10.1016/j.ejor.2023.09.010Stevens, A., De Smedt, J., & Peeperkorn, J. (2022). Quantifying explainability in outcome-oriented predictive process monitoring. In Process Mining Workshops (LNBIP 433, pp. 194–206). Springer. https://doi.org/10.1007/978-3-030-98581-3_15

van Dongen, B. F. (2017). BPI Challenge 2017 [Data set]. 4TU.ResearchData. https://doi.org/10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b