DISSERTATION

A new approach for predicting quantitative variables in association rule mining using the example of predicting the length of stay in intensive care units

Ein neuer Ansatz für die Vorhersage quantitativer Variablen im Association Rule Mining am Beispiel der Vorhersage der Intensivverweildauer

zur Erlangung des akademischen Grades
Doctor rerum medicinalium (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

M.Sc. Sedir Mohammed

Erstbetreuung: Prof. Dr. med. Dr. rer. nat. Felix Balzer

Datum der Promotion: 15. September 2025

Contents

List of Tables ..... **ii**

List of Figures ..... **iii**

Zusammenfassung **1**

Abstract **2**

1 Introduction **3**

2 Method **12**

3 Data **15**

4 Results **19**

5 Discussion **23**

6 Conclusion **31**

References **32**

Eidesstattliche Versicherung **36**

Detailed declaration of contribution **37**

Printing copy of the publication **38**

Curriculum Vitae **54**

List of publications **56**

Acknowledgments **57**

List of Tables

Table 1. “Used categorical variables of the MIMIC-IV dataset and the possible states of each variable, including the relative number of occurrences of the states.” [1] (Own table - originally published in [1]). ……16
Table 2. Mapping of admission types from the eICU dataset to corresponding categories in the MIMIC-IV dataset. ……17
Table 3. Used categorical variables of the eICU dataset and the possible states of each variable, including the relative number of occurrences of the states. ……...18
Table 4. Rules for considered minsup values (2.5%, 5%, 7.5%), with highest- and second-highest KLD, for Scenario 1 and Scenario 2. (Based on results presented in multiple tables in [1]). ……20
Table 5. Rules for considered minsup values (2.5%, 5%, 7.5%), with highest- and second-highest KLD, for Scenario 1 and Scenario 2, using eICU. ……..……...…21
Table 6 "Top 10 interesting rules (using minsup 5%) comparison for the MIMIC-IV dataset between generated rules by approach from Aumann et al. and Scenario 2 of our approach." [1] (Own table - originally published in [1]). ……..25
Table 7. Top 10 interesting rules (using minsup 5%) comparison for the eICU dataset between generated rules by approach from Aumann et al. and Scenario 2 of our approach. ……...26

List of Figures

Figure 1. “Example for finding interesting rules based on nonparametric statistics.” [1] (Own visualization - originally published in [1]). 13
Figure 2. ”Distribution of the ICU LOS values of the considered patient treatments.” [1] (Own visualization - originally published in [1] for the MIMIC-IV dataset and adjusted for this thesis). 16
Figure 3. Distribution of the ICU LOS values of the considered patient treatments of the eICU dataset. 18
Figure 4. Distribution of set and counter-set for rule A1 from Scenario 1, using MIMIC-IV (Own visualization - originally published in [1]). 20
Figure 5. Distribution of set and counter-set for rule A2 from Scenario 1, using eICU. . 22
Figure 6. Exemplary distributions with the same means but different SD and high KLD. 26
Figure 7. Statistics about generated and matched rules between MIMIC-IV and eICU, considering the top 40% most interesting rules according to our method. 29
Figure 8. KLD distributions for the matched rules, considering the different minsup values and considering the top 40% most interesting rules according to our method and the full set (100%) of rules. 30

Zusammenfassung

Das Association Rule Mining ist eine bekannte Methode, um häufig auftretende Muster in den Daten in Form von Regeln (Assoziationsregeln) zu identifizieren. Assoziationsregeln haben die Eigenschaft, dass die Regeln menschenlesbar sind und dadurch, im Gegensatz zu vielen anderen Prädiktionsalgorithmen, nachvollziehbare Vorhersagen zulassen. Klassische Algorithmen des Association Rule Minings, wie Apriori oder FP-growth, sind nicht in der Lage numerische Daten (quantitative Variablen) zu prozessieren, was die Anwendbarkeit dieser Algorithmen einschränkt. Aumann et al. haben in der Vergangenheit einen Ansatz eingeführt, welcher die Verarbeitung von quantitativen Variablen auf der rechten Seite einer Regel zulässt. Zur Identifizierung von Regeln nutzt dieser Ansatz jedoch summary statistics (es werden also nur einzelne Momente betrachtet), welche mitunter eine verzerrte Einsicht in die Daten liefern. Diese Arbeit führt auf Basis des Ansatzes von Aumann et al. eine neue Methode zur Verarbeitung quantitativer Variablen auf der rechten Seite einer Regel ein. Unser Ansatz basiert auf der Anwendung der Kullback-Leibler-Divergenz und führt dadurch eine ganzheitliche Betrachtung der Datenverteilung zur Identifizierung von Regeln ein.

Wir demonstrieren unseren Ansatz am Beispiel der Vorhersage der *Verweildauer von Patienten auf Intensivstation* (Intensivverweildauer). Die Intensivverweildauer beschreibt die Anzahl der Tage, die sich ein Patient während eines Krankenhausaufenthaltes auf der Intensivstation befindet. Die Vorhersage der Intensivverweildauer dient insbesondere zur Ressourcenoptimierung im Krankenhaus. So kann durch eine Abschätzung der Intensivverweildauer die Bettenauslastung optimiert werden.

Abstract

Association rule mining is a well-known method for identifying frequently occurring patterns in the data in the form of rules (association rules). Association rules have the property that they are human-readable and, therefore, allow comprehensible predictions, in contrast to many other prediction algorithms. Classical association rule mining algorithms, such as Apriori or FP-growth, cannot process numerical data (quantitative variables), which limits their applicability. In the past, Aumann et al. introduced an approach that allows the processing of quantitative variables on the right side of a rule. However, this approach uses summary statistics to identify rules, which may provide a skewed insight into the data. This work introduces a novel method for processing quantitative variables on the right side of a rule based on the approach of Aumann et al.. Our approach is based on applying the Kullback-Leibler divergence and thereby introduces a holistic view of the data distribution to identify rules.

We demonstrate our approach using the example of predicting the *length of stay of patients in intensive care units* (ICU LOS). The ICU LOS describes the number of days a patient spends in the intensive care unit during a hospital stay. Predicting the ICU LOS optimizes hospital resources, such as bed utilization.

1 Introduction

This work is based on the paper we published in the journal *Information Systems* in 2023: “A statistical method for predicting quantitative variables in Association Rule Mining” [1]. This thesis fundamentally extends the mentioned work by incorporating numerous adjustments, extending beyond a detailed contextualization of association rule mining within the healthcare domain to also include a deeper discussion of LOS prediction in the literature. Additionally, we apply our method to a new dataset for predicting length of stay in intensive care units. The generated rules are thoroughly analyzed using the new dataset. We further compare the rules derived from the new dataset with those from the previously introduced dataset to assess the consistency of length of stay in intensive care unit outcomes and evaluate the generalizability of our approach across datasets.

In the following Sections (1.1-1.3), we introduce the topic of association rule mining and the relevance of the prediction of length of stay in intensive care units. In Section 1.4, we summarize the research objectives of this work before presenting the preliminary work of this paper. In Section 2, we discuss the approach we have developed. Our method distinguishes between two scenarios described in detail in Sections 2.1 and 2.2. In Section 3, we present the datasets that we used in our experiments for this thesis. In Section 4, we show the results produced based on our approach. Section 5 compares the rules generated by our approach with those produced by the method of Aumann et al. [2], highlighting the superiority of our approach. Additionally, we compare the rules and the predicted length of stays between the considered datasets to emphasize the superiority of our approach further. Section 6 closes this thesis with a conclusion and an outlook for future research directions.

1.1 Background on Association Rule Mining

Data mining is an analytic process that involves extracting valuable insights, patterns, and knowledge from large datasets [3]. A key component of data mining is the application of sophisticated techniques capable of sifting through extensive data volumes to pinpoint pertinent patterns and insights [4], [5]. One fundamental technique in data mining is *Association Rule Mining* (ARM), which aims to reveal interesting and frequently occurring patterns or associations within datasets [1], [6], [7], [8]. Originally used in

market basket analysis, ARM identifies sets of items commonly appearing together in transaction records, offering valuable insights into consumer behaviour and item relationships [6]. Its essence lies in generating rules based on these associations, providing comprehensible data insights. Beyond retail applications, ARM finds relevance in various application domains, including healthcare, bioinformatics, and business [7].

The Apriori algorithm, one of the earliest and most renowned ARM methods, has paved the way for various other algorithms and enhancements to improve efficiency, including the FP-growth algorithm, which we used in our study [1] and this manuscript. The underlying principle involves the generation of frequent itemsets, a process known as *frequent itemset mining*, to extract prevalent relationships within a dataset [6], [9]. The discovery of these associations, presented in the form of rules, is especially relevant within the context of transparent prediction of outcomes. The need for transparent predictions becomes specifically critical in the medical context, where comprehending the decision-making process of a model is highly desirable [10]. Modern predictive models often operate as black boxes, rendering their decisions inscrutable to end-users [10], [11], [12].

Given that the well-being and health of patients are at stake, it would be preferable if medical professionals are able to understand why a model arrives at specific predictions [10]. In this regard, association rules provide a significant advantage by generating human-readable rules [13], making the model's predictions both comprehensible and interpretable for users.

Comprehensiveness becomes even more critical under the guidelines of the AI Act [14]. With its stringent focus on ethical and transparent AI applications, particularly in high-risk sectors, this act necessitates deploying transparent AI methods. ARM aligns with this requirement by extracting clear and understandable patterns and associations within complex datasets.

1.2 Key Concepts of Association Rule Mining

The original ARM in form of the Apriori algorithm operates in two distinct phases [6].

In the first phase, the algorithm generates *frequent itemsets*. These groups of items appear together in the data with a frequency exceeding a predefined threshold *minsup* [5], [9, p. 221]. In the second phase, the algorithm derives association rules based on these itemsets. Typically, $X \Rightarrow Y$ represents an association rule, where X and Y are distinct

itemsets. The rule implies that if all items in *X,* also called the antecedent [15]*,* are present (which logically resembles a conjunction), then the items in *Y,* also called the consequence [15]*,* are also likely to be present.

Agrawal et al. [6] introduced two key metrics to select and evaluate these generated rules. The first metric, the S*upport*, relates to the first phase of identifying frequent itemsets. *Support* measures the frequency of an itemset's appearance in the dataset and indicates the rule's relevance and reliability. Mathematically, the support of a rule $X \Rightarrow Y$ is calculated as the proportion of records in the dataset containing both *X* and *Y*. For example, a rule appearing only once in a million records may not be statistically significant.

The second key metric in ARM is C*onfidence*, which is central to the Apriori algorithm's second phase, where the algorithm evaluates the generated association rules. Specifically, *Confidence* assesses the conditional probability that *Y* is present in the data, given that *X* is present. For the rule $X \Rightarrow Y$, the algorithm calculates the *Confidence* as the proportion of records containing *X* that also contain *Y*. This metric is crucial for understanding the strength of the implication of a rule. A high *Confidence* means that *Y* is likely to appear whenever *X* appears, indicating a strong association between *X* and *Y* [1], [6], [8].

1.2 Processing of Quantitative Variables

In their original form, classical ARM algorithms such as Apriori and FP-growth were designed to handle categorical variables, typically discrete and unordered scaled data [2], [6], [16], [17], [18]. These algorithms were not initially intended to process quantitative variables, which include interval or ratio-scaled data, on either side of a rule [2], [7], [16], [18], [19]. Calculating *Support* and *Confidence* is not straightforward for quantitative data. The likelihood of the same set of numerical values occurring frequently enough to meet the *Support* threshold is low, and establishing high *Confidence* levels for specific numeric combinations can be rare. However, as the restriction to categorical variables is not practical or feasible [2], the literature has introduced several methods to incorporate quantitative variables into ARM algorithms. Adhikary et al. [7] classified well-known methods into different categories.

The first class of methods includes partitioning approaches. One of the earliest strategies introduced by Srikant et al. [18] involves partitioning the quantitative variables [7]. These techniques involve segmenting quantitative variables into distinct intervals. Within this framework, the authors identified two key issues: the *minsup* and *minconf* problems. The *minsup* problem describes situations where partitions are too small to reach the minsup threshold. Conversely, the *minconf* problem describes situations when partitions are too large. This situation leads to a loss of information, as the rules only make general statements accordingly.

Clustering approaches represent another category for handling quantitative variables [7]. These methods consider the identified clusters as independent categorical variables within the quantitative variables. However, applying and interpreting the generated association rules can be challenging as these methods may fail to identify separated clusters. Thus, clustering-based approaches may cause the *minsup* or *minconf* problem [1] in a similar way to partitioning approaches, concretely when too many clusters or clusters that are too large are created.

A third common class are statistics-based approaches [7], which include the method proposed by Aumann et al. [2] and the generalization used in our study. When generating association rules, the authors incorporate summary statistics, such as the arithmetic mean, of quantitative variables in their approach. Their method identifies interesting rules where quantitative variables appear on the right side of the rule, if the subset of transactions (considered data) to which the rules apply (these transactions are defined as *set*) shows a statistically significant difference in mean compared to the *counter-set* of transactions (which refers to the transactions to which the respective rules do not apply). Aumann et al. [2] use the Z-test to determine this statistical significance. However, the validity of using summary statistics to identify interesting rules raises the problem that summary statistics correspond to just one *moment*. For example, when using one moment like the mean, the resulting rule set may differ significantly from a rule set that relies on variance [1].

Our research aims to identify interesting rules regardless of the chosen summary statistics. We generalize the approach proposed by Aumann et al. [2] to address scenarios when reducing a quantitative variable to summary statistics is inadequate. With our approach, we can model the entire distribution of a quantitative variable on the right-hand side, representing the part of the rule we aim to predict. This generalisation measures

the discrepancy using the *Kullback-Leibler divergence* (KLD) between the distributions of quantitative variables on the right side of a rule and the distribution of quantitative variables in the counter-set of transactions [20], [21].

We establish our approach in the form of two scenarios. In the first scenario, we use non-parametric histograms to compare the quantitative variables of the set and counter-set, selecting rules where the KLD between these histograms is particularly large. In the second scenario, a two-step process is employed: first, parametric statistics are used to model the distributions of the values in the set and counter-set, focusing on cases where the KLD is particularly small. In the second step, we then select rules where the KLD between the parametric distributions is particularly large. Section 2 provides a detailed explanation of these scenarios.

1.3 Application in Healthcare

While data mining has been widely applied in the healthcare domain [22], this thesis focuses specifically on ARM, which has demonstrated its effectiveness as a powerful tool for uncovering hidden patterns and relationships in large datasets [23]. This capability makes it particularly valuable in healthcare, where data is both abundant and complex. In their survey, Altaf et al. [23] provide a comprehensive overview of ARM applications in healthcare from 2005 to 2014, showcasing its versatility across a broad range of healthcare challenges. We present in the following some of the presented applications mentioned by Altaf et al. [23]:

ARM has frequently been employed to predict health-related issues [23]. For example, it has been used to identify relationships between vital parameters – such as smoking and drinking habits, overweight status, and medical diagnoses like hypertension, diabetes, and cardiovascular diseases. Moreover, ARM has also been applied in predicting brain and breast tumors as well as oral cancer. These insights are instrumental in informing preventative care strategies and guiding targeted interventions [23].

Beyond predictive modelling, ARM has been utilized to analyze medical images, such as CT scans, by identifying associations between regions of interest and specific image characteristics [23]. By uncovering patterns in digitized medical images, ARM supports diagnostic decision-making and enhances the accuracy of image-based disease detection.

Additionally, ARM has been applied to studying healthcare costs, which can be an indicator for future health needs [24], particularly in patients diagnosed with diabetes mellitus [23], [25].

These examples highlight the broad applicability of ARM across various areas inside the healthcare domain. However, it is important to note that the presented examples primarily focus on processing categorical variables. In this manuscript, we go further and demonstrate our approach by applying it to the *MIMIC-IV* [26] and *eiCU* dataset [27] to predict the *length of stay* (LOS) of patients in an *intensive care unit* (ICU), also known as *ICU LOS* [28] – a quantitative variable. LOS and ICU LOS refer to the number of days a patient spends in a hospital during treatment [28]. Due to the day-by-day consideration and its potentially wide range, LOS respectively ICU LOS represents quantitative variables.

The prediction of ICU LOS might be used to optimize resource management in the hospital concerning the usage of beds, staff and medical equipment [29]. In particular, by predicting the ICU LOS, the utilisation of ICU beds can be estimated in advance to increase ICU bed occupation by better distributing unused resources [30], [31], [32]. The optimised resource utilisation allows the hospital personnel to plan elective procedures better and to reduce the waiting time for the provision of the procedures. Another point is the early identification of patients with prolonged ICU LOS. Adapted care actions may be taken for these patients at an early stage to reduce the ICU LOS [29], [30], [31].

Furthermore, family members benefit from the knowledge of the ICU LOS, as they can estimate the costs of the stay [30]. However, cost estimation through the prediction of ICU LOS is relevant not only from the hospital's perspective. The ICU LOS could also be used to estimate reimbursement rates, as hospitals in Germany or in the United States get reimbursed for the provision of medical care according to a payment system [33], [34], [35]. The costs associated with patient care can escalate rapidly due to the need for medications, therapies, and advanced medical equipment required for comprehensive treatment [24]. The healthcare costs can also be used as an indicator for the further care of the patient [24].

Predicting ICU LOS respectively LOS is already subject of previous literature. In their systematic survey Stone et al. list approaches in the literature, by introducing a taxonomy of LOS prediction models [24]. The authors consider publications from 1970 to 2021,

selected through a systematic review focusing on studies relevant to LOS prediction. They differentiate between “operational research-based approaches”, “statistical and arithmetic approaches” and “machine learning and data mining approaches”, which we detail in the following [24]:

"Operational research-based approaches" use mathematical and simulation models to predict LOS, focusing on rather simple models, by considering often the average LOS. Techniques include compartmental modelling (e.g., dividing patients into acute and long-stay compartments), simulation modelling (e.g., discrete event simulation to evaluate development in the LOS), and Markov models (e.g., probabilistic models for patient transitions inside a healthcare system). For instance, a two-compartment model predicts patient flow through acute and long-stay phases, supporting in resource allocation. While these methods provide insights into hospital flows respectively the LOS, they are often overly simplistic and may not capture details in LOS outcomes [24].

"Statistical and arithmetic approaches" rely on regression techniques and other statistical analyses to establish relationships between patient characteristics and LOS. The literature used a range of regression models in this context, spanning from relatively simple linear regression to more sophisticated regression tree models. While regression models are widely used in the literature, they often necessitate formal statistical training and a solid understanding of complex statistical concepts, which requires a high effort to apply effectively in practical settings [24].

Lastly, the "machine learning and data mining approaches" use advanced algorithms, including artificial neural networks, support vector machines, and clustering methods, to identify patterns in the data. For instance, an artificial neural network was employed to predict the LOS of coronary patients. However, these models face challenges in clinical adoption due to their lack of transparency as “black box” models, which has driven growing interest in explainable AI systems, which we already discussed before. In order to make black box models explainable [24]. ARMs have also been used in the past to break down the predictions of these models retrospectively [36]. The idea of using association rule mining in connection with the LOS has also already been used in the literature [37], [38]. However, the literature primarily considered LOS as a categorical variable by splitting it into different intervals.

1.4 Research Objectives

As mentioned before, this work is based on our paper “A statistical method for predicting quantitative variables in association rule mining” [1].

Aumann et al. [2] originally introduced an ARM-based approach that enables processing quantitative variables on the right-hand side of a rule. This processing is made possible by searching for the rules where the set and the counter-set differ significantly based on a concrete summary statistic (see Section 2). In the mentioned paper, the primary objective is to introduce a generalization of the approach of Aumann et al. [2], which does not select rules based on a summary statistic over the set and counter-set. Instead, the approach considers the distribution of set and counter-set holistically and compares the discrepancy between these distributions using the KLD. The idea of using the KLD as a measure to evaluate frequent itemsets or rules comes from my master thesis [39]. However, in the mentioned thesis [39], frequent itemsets were only filtered based on parametric statistics in the form of a simpler one-step approach by searching for the itemsets whose KLD between set and counter-set is particularly small. In this manuscript, we extend this approach by introducing two scenarios (Scenario 1 and Scenario 2). In the first scenario, we find rules based on non-parametric statistics, while in the second scenario, we find rules based on parametric statistics, in the form of a two-step approach (see Section 2 for more details on the scenarios). The introduction of these two scenarios and the distinction between non-parametric and parametric statistics is a fundamental extension of the original idea of the master thesis.

Similar to the mentioned thesis [39], we demonstrate our approach by predicting the ICU LOS, a relevant value in the medical domain to optimize resource utilization in the ICU (we have already presented, as part of this thesis, an overview of the literature in Section 1.3, both from the perspective of ARM in the healthcare domain and LOS prediction). For this purpose, we use the MIMIC-IV [26] and eICU [27] dataset (the usage of the eICU dataset and all further analysis are specifically added as parts of this thesis and are not part of the mentioned paper [1]), which are well-known in medical research (we present the details of the datasets in Section 3). However, in our work [1], we also use two more datasets from other domains. Also, another fundamental extension to the master thesis mentioned is that we compare the rules generated by our approach with those generated by the approach of Aumann et al. [2] and search for similarities and

differences in the extracted rule sets. We especially do this as part of this thesis also for the eICU dataset. As a further extension of the work by Mohammed et al. [1], this thesis includes a comparative analysis of the rules generated by our approach using the MIMIC-IV and eICU datasets. We analyze these rules in relation to their ICU LOS outcomes, respectively, their distributions, focusing particularly on the most interesting rules ranked by our method. This analysis highlights the effectiveness of our approach in prioritizing meaningful rules and demonstrates its ability to generate generalizable rules across different datasets.

2 Method

As part of our work, we introduced a generalization of the approach of Aumann et al. [2], which does not perform this classification of interestingness based on a specific summary statistic. Instead, probability distributions are modelled holistically and compared based on the *Kullback-Leibler divergence* (KLD) [1]. The KLD, as originally defined, assumes two discrete probability mass functions or two continuous probability density functions [20], [40]:

$$KLD(p \,||\, q) = \sum_{x \in X} p(x) \cdot log \frac{\mathrm{p(x)}}{q(x)} \tag{1}$$

$$KLD(p \,||\, q) = \int_{-\infty}^{\infty} p(x) \cdot log \frac{\mathrm{p(x)}}{q(x)} dx \tag{2}$$

Following the approach of Aumann et al. [2], the *k*-frequent itemsets, which are generated via the categorical variables using an arbitrary ARM algorithm, form the input for our method [1]. In our generalization [1], following the principle of Aumann et al. [2], a rule is considered interesting if the measured KLD exceeds a predefined threshold for each quantitative variable on the right side of a rule. Note that summary statistics can still be obtained post-hoc over the distribution of a set to interpret the rule.

As mentioned before, we introduce our approach by two scenarios. The Scenario 1 finds interesting rules based on nonparametric statistics. The Scenario 2 finds the interesting rules based on parametric statistics. In the following, we discuss these scenarios in detail. In both approaches we assume that the *k*-frequent itemsets are already generated according to a respective *minsup* threshold via an arbitrary ARM algorithm, like Apriori or FP-growth [6], [16]. Additionally, we assume that the respective set and counter-set were extracted for the considered quantitative variables for the right side of the rule. In the following, we always assume one quantitative variable, which results in one pair of set and counter-set per rule. In reality the right side could consist of a set of quantitative variables [2].

2.1 Scenario 1: nonparametric statistics

Based on a set and a counter-set, we obtain non-parametric histograms, which our approach uses to calculate the discrepancy using the KLD (Scenario 1). Since we obtain a binned histogram for the respective set and thus have a discrete distribution, *Eq. 1* calculates the discrepancy between a set and a counter-set. Here, *p* corresponds to the counter-set and *q* to the set. Our approach performs this calculation for each identified rule. The considered rule is classified as interesting if the discrepancy is above a previously defined threshold.

2.1.1 Example for finding interesting rules: nonparametric statistics

In the following, we give an example of finding interesting rules based on nonparametric statistics. This example was introduced first in our original publication, the description of which we adopt word for word in the following [1]:

Figure 2 “shows the example, including the different steps from input data to the final rule set. The input data consists of categorical variables and one quantitative variable. We used the FP-Growth algorithm to obtain frequent itemsets, however, any ARM algorithm can be used. In this example, we set *minsup* to 0.5%. Based on the frequent itemsets, our approach assigns the values of the quantitative variable to set and counter-set. From the filled set and counter-set, our approach calculates the probability functions. Our approach calculates the KLD between the counter-set and set based on the probability functions. For this example, we consider rules interesting if the KLD is at least 0.4; thus, the approach filters the rules *A* and *A⇒C* out.”[1]

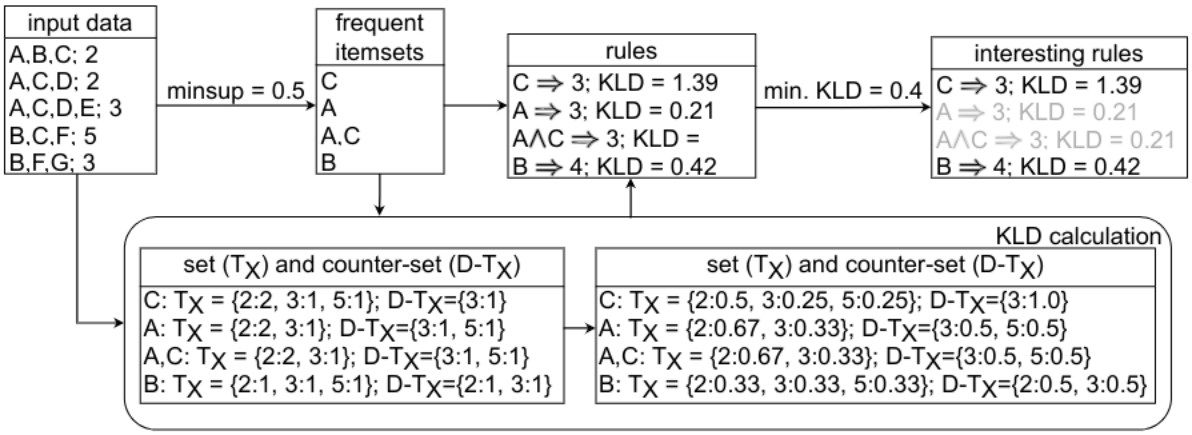


Figure 1. “Example for finding interesting rules based on nonparametric statistics.” [1] (Own visualization - originally published in [1]).

2.2 Scenario 2: parametric statistics

In a second scenario (Scenario 2), our approach can be used to find interesting rules when parametric models represent a set and counter-set [1]. In this scenario, we assume that suitable parametric distributions for the set and the counter-set are known. In this case, we find interesting rules in two steps.

In the first step, our approach searches for rules where the set respectively the counter-set can be modelled particularly well by a parametric distribution. We use the KLD again for this purpose. Unlike in the first scenario, where we always search for particularly large discrepancies, we search for particularly small discrepancies in this step, since the parametric distribution should model the respective binned histogram accurately.
While the theoretical distributions are continuous, while the distribution of set and counter-set is discrete after the binning. The latter requires a transformation of the original KLD equation, which we discuss in detail in our publication [1]. We introduce a threshold for the accepted discrepancy between the respective parametric distribution and the set or counter-set.
Once the rules for which the set or counter-set cannot be modelled sufficiently well by the given parametric distribution have been filtered out in the first step, the parametric distributions of the set and counter-set are compared within the remaining rules in the second step. Following the principle of our first scenario, the approach searches for the rules in which the parametric distributions of set and counter-set are particularly different. Since both distributions are now continuous, we use the KLD according to *Eq. 2* to calculate their discrepancies. Considering a pre-defined threshold for the discrepancies for the second step, the approach filters the rules accordingly.

3 Data

In the following, we present the MIMIC-IV dataset [26] that we used in our experiments [26]. In the original paper, we also demonstrated the approach on other datasets from different domains, to show that our approach is not limited to the medical context [1]. Additionally, we apply our method to the eICU Collaborative Research Database [27] (in the following abbreviated with eICU) as part of this thesis to predict also the ICU LOS. We shall also present this dataset in detail below.

MIMIC-IV. The MIMIC-IV dataset consists of pseudonymized patient data. The data was provided by the Beth Israel Deaconess Medical Center and includes patients admitted to either the emergency department or an ICU between 2008 and 2019. Patients were at least 18 during their first visit to be included in the data [26], [41]. Access is granted to users after registration and successful completion of an online training course (http://www.physionet.org). We consider 50,504 patient treatments in the ICU, for which at least one *International Classification of Diseases* (ICD) code [42] is stored in the respective patient record for their stay. The patient treatments considered correspond to 40,357 individual patients, as a patient may have several ICU stays in their lifetime. We only consider patients who did not die in the ICU, as a shortened LOS characterizes these stays. We consider various attributes (as categorical variables) from the patient record to predict ICU LOS. We list these, including the distribution of conditions to be admitted, in Table 1. As Table 1 shows, we also include the diagnosis and procedures. It is worth noting that for the included patients, the diagnoses and procedures were encoded using either ICD 9 or ICD 10. We apply a day-by-day binning for the ICU LOS. Figure 2 shows the distribution of the ICU LOS of the patient treatments under consideration in the form of a histogram. The ICU LOS ranges from less than one day (0 days) to 110 days, with the frequency decreasing sharply with the number of days. The distribution of the ICU LOS shown in Figure 2 is comparable to an exponential distribution. This also corresponds to how the literature models the distribution of LOS [43].

Table 1. "Used categorical variables of the MIMIC-IV dataset and the possible states of each variable, including the relative number of occurrences of the states." [1] (Own table - originally published in [1]).

categorical variable	possible state
gender	female (43%), male (57%)
insurance type	medicaid (8%), medicare (44), other (49%)
admission type	ambulatory observation (<1%), direct emer. (4%), direct observation (<1%), elective (4%), eu observation (<1%), ew emer. (49%), observation admit (11%), surgical same day admission (13%), urgent (17%)
diagnoses	set of ICD 9 or ICD 10 codes
procedures	

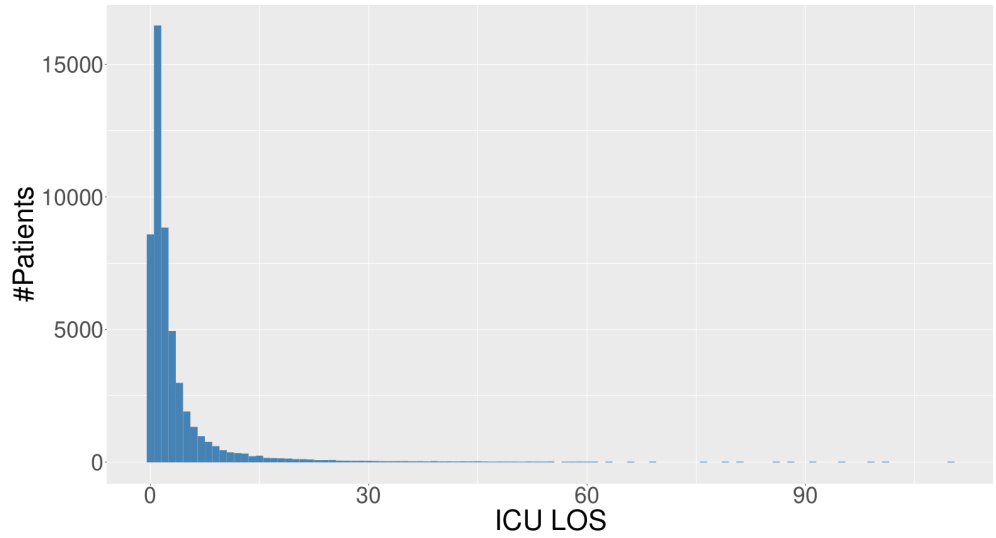


Figure 2. "Distribution of the ICU LOS values of the considered patient treatments." [1] (Own visualization - originally published in [1] for the MIMIC-IV dataset and adjusted for this thesis).

eICU. We extended our analysis with the eICU dataset, provided by Philips Healthcare [27], [44], [45]. This dataset originates from the eICU program, a telehealth ICU service that remotely monitors ICU patients, allowing remote caregivers to support and enhance local patient treatment. Similar to MIMIC-IV, the eICU dataset consists of pseudonymized patient data and is accessible upon registration and successful completion of an online training course. The dataset includes a sample of patient records from discharges occurring between 2014 and 2015. For our experiments, we analyze 100,019 unit stays (we only consider those, where the respective patient did not die and the age of the patient is at least 18), each associated with a specific hospital stay. A single hospital stay may include multiple unit stays, and each hospital stay is uniquely linked to an individual patient. The analyzed unit stays correspond to a total of 79,275 unique patients. Among the categorical variables used in our analysis, we included admission types,

which are essential for understanding patient pathways. However, the admission types available in the eICU dataset differ from those in MIMIC-IV. To enable comparison of the rules generated from eICU with those derived from MIMIC-IV (see Section 5 - Discussion), we mapped the admission types in eICU to corresponding categories in MIMIC-IV, as Table 2 shows. This mapping was determined heuristically, based on general reasoning about the categories, rather than derived from specific clinical or procedural guidelines. Table 3 shows the complete list of categorical variables used in our analysis, including the mapped admission types. In contrast to the MIMIC-IV dataset, the diagnoses here are mostly encoded in ICD 9 and ICD 10, thus there exists a mapping. We will use this property later (see Section 5) to assess the generalizability and consistency of the rules generated by our approach. The distribution of ICU LOS in the eICU dataset, which ranges between 0 (less than one day) and 507 days, also exhibits a pattern following an exponential distribution, as Figure 3 shows (we exclude the ICU LOS values above 110 for better visualization - 9 values are excluded).

Table 2. Mapping of admission types from the eICU dataset to corresponding categories in the MIMIC-IV dataset.

original type	**mapped type**
direct admit	urgent
emergency department	ew. emer.
floor	observation admit
operating room	surgical same day admission
other hospital	urgent
other ICU	urgent
step-down unit (SDU)	observation admit
ICU to SDU	observation admit
recovery room	surgical same day admission
chest pain center	ew. emer.
acute care/floor	urgent
PACU	surgical same day admission
observation	direct observation
ICU	urgent
other	direct observation

Table 3. Used categorical variables of the eICU dataset and the possible states of each variable, including the relative number of occurrences of the states.

categorical variable	**possible state**
gender	female (46%), male (54%)
admission type	direct observation (<1%), ew emer. (55%), observation admit (13%), surgical same day admission (19%), urgent (13%)
diagnoses	set of ICD 9 / ICD 10 codes

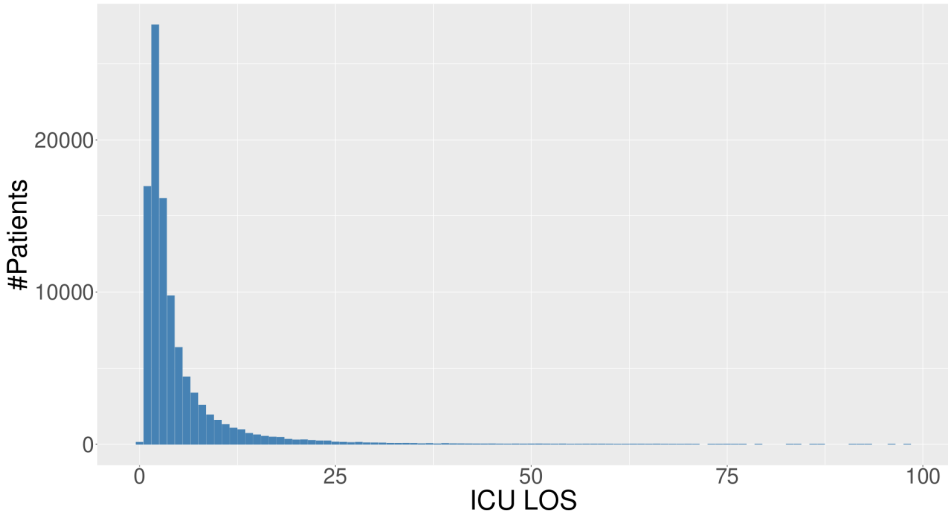


Figure 3. Distribution of the ICU LOS values of the considered patient treatments of the eICU dataset.

4 Results

In this section, we present the rules generated based on our approach using the MIMIC-IV and the eICU dataset to predict ICU LOS. To generate rules, we consider different *minsup* values in our experiments (absolute values in parentheses): 2.5% (MIMIC-IV: 1,263; eICU: 2,501), 5% (MIMIC-IV: 2,526; eICU: 5,001), and 7.5% (MIMIC-IV: 3,788; eICU: 7,502). For example, there must be at least 1,263 entries in the data that contain the left side of a rule, i.e. the *k*-frequent itemset, for *minsup* 2.5%. We use the FP-growth algorithm to generate the *k*-frequent itemsets [16]. All code for dataset preparation and experiment execution is available in the following GitHub repository. However, we may use any ARM algorithm for processing categorical variables.

MIMIC-IV. Table 4 shows the rules for Scenario 1 (non-parametric statistics) and Scenario 2 (parametric statistics) with the two highest KLDs for each *minsup* value considered, using the MIMIC-IV dataset. In addition, Table 2 shows further summary statistics applied to the respective set, including the *mean*, the *standard deviation* (SD) and the *interquartile range* (IQR). To check the robustness of the KLD, Table 2 shows the results of a non-parametric bootstrap with 10,000 samples (rounded to two decimals after the comma). This shows that the error bars of the KLD for the in Table 2 considered rules is at most 0.08, i.e. the calculated KLD is robust to the choice of different samples in the set.

Table 4. Rules for considered minsup values (2.5%, 5%, 7.5%), with highest- and second-highest KLD, for Scenario 1 and Scenario 2. (Based on results presented in multiple tables in [1]).

	rule (left side)	mean	SD	IQR	sup.	KLD	minsup
Scenario 1							
(A1)	5A1955Z	15	11.08	12	1521	1.74 ± 0.07	2.5%
(A2)	EW EMER. $\wedge$ 966 $\wedge$ 9672	13	8.71	9	1391	1.71 ± 0.08	2.5%
(B1)	9672	12	9.37	9	2712	1.65 ± 0.05	5%
(B2)	3E0G76Z	9	9.42	9	3055	0.58 ± 0.02	5%
(C1)	966	8	8.44	9	4461	0.52 ± 0.02	7.5%
(C2)	51881	7	7.62	7	4243	0.33 ± 0.01	7.5%
Scenario 2							
(A1)	02HV33Z $\wedge$ 3E0G76Z	11	10.55	11	1435	0.6427 ± 0.04	2.5%
(A2)	966 $\wedge$ 51881	11	8.86	9	1713	0.6123 ± 0.05	2.5%
(B1)	3E0G76Z	9	9.42	9	3055	0.5199 ± 0.02	5%
(B2)	966	8	8.44	9	4461	0.477 ± 0.02	5%
(C1)	966	8	8.44	9	4461	0.477 ± 0.02	7.5%
(C2)	51881	7	7.62	7	4243	0.301 ± 0.01	7.5%

To illustrate, we examine the details of a specific rule presented in Table 2. Rule (A1) from Scenario 1 illustrates that the average ICU LOS is 15 days when the medical procedure "Respiratory Ventilation, Greater than 96 Consecutive Hours" (5A1955Z) is assigned to a patient who has been admitted to the hospital as an emergency. However, the standard deviation of 11.08 days points to a significant variation in the ICU LOS. This substantial variation is associated by the setting a low *minsup*, in this instance, to 2.5%. To clarify, for events to be considered part of a frequent itemset in the data, they must occur less frequently, making them less unique. For rule (A1), this results in a wide range of ICU LOS among patients, as shown by the highly dispersed distribution depicted in Figure 4. This extensive dispersion, when compared to the counter set, also leads to an increased KLD.

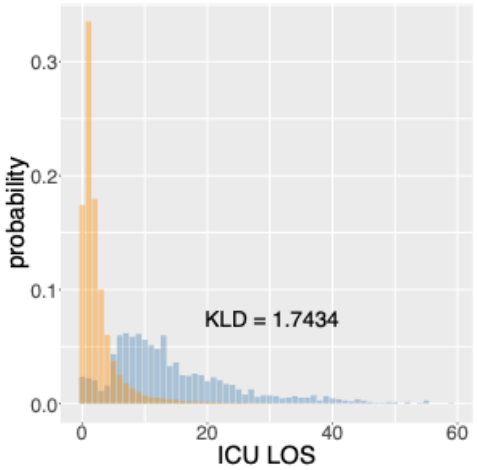


Figure 4. Distribution of set and counter-set for rule A1 from Scenario 1, using MIMIC-IV (Own visualization - originally published in [1]).

eICU. Using the eICU dataset, Table 5 presents the rules with the two highest KLDs for Scenario 1 and Scenario 2 (using ICD 10) for the respective minsup values, following the format used in Table 4. Overall, the results demonstrate trends similar to those observed with MIMIC-IV. The KLDs are generally lower, indicating a smaller discrepancy between the set and counter-set. Additionally, the error bars are smaller compared to MIMIC-IV, with a maximum value of 0.03, but mostly reaching a maximum of only 0.01. This reflects the high robustness of the calculated KLDs.

Next, we examine a specific rule from Table 5 in detail, namely rule (A2) from Scenario 1. This rule indicates that the average ICU LOS is 11 days when a patient is diagnosed with both “Acute kidney failure, unspecified” and “Acute respiratory failure, unspecified whether with hypoxia or hypercapnia,”. However, the standard deviation of 13.66 suggests substantial variation in ICU LOS for this rule. Although the standard deviation is larger than in the example of MIMIC-IV, Figure 5 shows a smaller discrepancy between the ICU LOS distributions.

Table 5. Rules for considered minsup values (2.5%, 5%, 7.5%), with highest- and second-highest KLD, for Scenario 1 and Scenario 2, using eICU.

	rule (left side)	mean	SD	IQR	sup.	KLD	minsup
Scenario 1							
(A1)	OBSERVATION ADMIT ∧ J96.00	11	10.09	9	2776	0.70 ± 0.03	2.5%
(A2)	N17.9 ∧ J96.00	11	13.66	10	3242	0.50 ± 0.02	2.5%
(B1)	OBSERVATION ADMIT	8	7.29	6	13772	0.27 ± 0.01	5%
(B2)	OBSERVATION ADMIT ∧ male	8	7.60	6	7225	0.25 ± 0.01	5%
(C1)	OBSERVATION ADMIT	8	7.29	6	13772	0.27 ± 0.01	7.5%
(C2)	J96.00	8	9.21	7	16321	0.24 ± 0.01	7.5%
Scenario 2							
(A1)	N17.9 ∧ J96.00	11	13.66	10	3242	0.29 ± 0.02	2.5%
(A2)	OBSERVATION ADMIT ∧ J96.00	11	10.09	9	2776	0.27 ± 0.03	2.3%
(B1)	J96.00	8	9.21	7	16321	0.17 ± 0.01	5%
(B2)	J96.00 ∧ male	8	8.60	6	8654	0.15 ± 0.01	5%
(C1)	J96.00	8	9.21	7	16321	0.17 ± 0.01	7.5%
(C2)	J96.00 ∧ male	8	8.60	7	8654	0.15 ± 0.01	7.5%

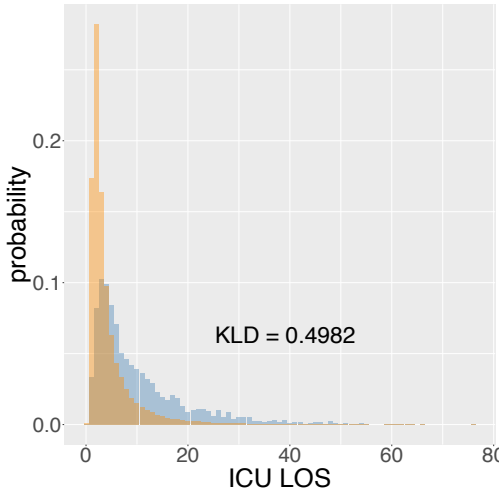


Figure 5. Distribution of set and counter-set for rule A2 from Scenario 1, using eICU.

5 Discussion

Association Rule Mining (ARM) is a well-known method for automatically extracting predictive patterns from datasets. One of the first approaches was the Apriori algorithm, published by Agrawal et al. [1], [6], [7]. However, the ARM in its original form is not designed so that it easily allows the processing of quantitative variables. Aumann et al. [2] have introduced a method that allows processing quantitative variables on the right-hand side of a rule. For this purpose, the approach first searches purely on the categorical variables for *k*-frequent itemsets to assign the quantitative variable's values to the set and the counter-set. A rule is classified as interesting if the means calculated over the set and counter-set are compared to each other particularly large, although other summary statistics can theoretically also be used for this classification. However, the use of summary statistics is problematic, purely depending on the selected summary statistic. In this work, we have introduced a generalization of the approach of Aumann et al. [2], which holistically considers the distributions of set and counter-set to classify the interestingness of a rule. Our approach uses the *Kullback-Leibler divergence* (KLD) to search for rules whose set and counter-set probability distributions are particularly different [1]. We introduce our approach in the form of two scenarios:

In Scenario 1, we assume non-parametric descriptions of the quantitative variable and compute the KLD between the binned histograms of set and counter-set. The rule is classified as interesting if the KLD exceeds a predefined threshold.

In Scenario 2, on the other hand, we assume parametric descriptions of the quantitative variable. We look for parametric distributions that approximate set and counter-set particularly well, i.e. the KLD is particularly small. We then used the parametric distributions to estimate the KLD between set and counter-set.

We demonstrated our approach by predicting the ICU LOS. We used the MIMIC-IV and eICU dataset as a data source and selected various categorical variables we considered in the experiments. For a parametric description, which is relevant for Scenario 2, we assumed an exponential distribution. Additionally, in our corresponding publication, we demonstrated our approach also on two other datasets from different domains [1].

The results using MIMIC-IV (see Table 4) and eICU (see Table 5) show some consistent trends, allowing us to generalize these findings: Rules derived with higher minsup values exhibit lower mean ICU LOS, *standard deviation* (SD), and *interquartile range* (IQR), reflecting larger, more homogeneous patient subgroups with less variability

in their outcomes. Similarly, the KLD is lower for these high-minsup rules, indicating that the ICU LOS distributions of patients who match versus do not match the rules are more similar. This observation suggests that higher-minsup rules describe generalizable patterns applicable to broader populations, capturing common clinical scenarios with standardized care protocols. Conversely, rules with lower minsup and higher KLD are associated with smaller, more distinct subgroups, often characterized by greater variability and higher LOS outcomes. These findings highlight a trade-off between the specificity of low-minsup, high-KLD rules and the generalizability of high-minsup, low-KLD rules, emphasizing the value of KLD as a selection criterion for interesting rules, identifying distinct clinical subgroups and assessing rule applicability. The code for dataset preparation, experiment execution, and results evaluation is available in the following GitHub repository: https://github.com/sedirmohammed/Implementation-A-Statistical-Method-for-Predicting-Quantitative-Variables-in-Association-Rule-Mining.

Comparing our approach to the method provided by Aumann et al. In the following, we compare the rule sets generated by our approach with those produced by the approach of Aumann et al. [2]. In our original publication, we consider the rule sets produced by our minsup value setting (2.5%, 5%, 7.5%) and the different scenarios in combinations with our considered datasets [1]. In this manuscript, we use Table 6 and Table 7 as examples to show the rule comparison for the MIMIC-IV and eICU datasets, considering minsup 5% and Scenario 2. Both tables show the 10 most interesting rules classified according to our approach, respectively, the method from Aumann et al. [2]. The comparison of the extracted rules for MIMIC-IV shows that the rule sets do not overlap at all. There are partial overlaps for other combinations of minsup values, scenarios and datasets, where the rules' positions in the ranking also match in some cases. Table 7 (rules based on eICU, using ICD 10) shows a similar behavior, but with a few overlaps between the rule sets produced by the approach from Aumann et al. and our approach. It is important to note, that both approaches only create 10 rules in this setting. Thus, the probability of producing an overlap is higher. During further evaluation of the rules, we spoke to a domain expert who confirmed that the selection of rules based on the KLD for predicting the ICU LOS, i.e. based on the discrepancy of the quantitative variable on the right-hand side of a rule, is reasonable (which is consistent with our generalization from the results based on Table 4 and Table 5).

Generally, the overlapping of rule sets is relatively rare, which is not unexpected. Our approach takes a holistic view of the distribution of the quantitative variable on the right side of a rule and, in contrast to the approach of Aumann et al. [2], does not focus on just one moment. The problem of relying on just one moment becomes more apparent in the following scenario: given an association rule that divides a domain into set and counter-set. Assuming set and counter-set have the same mean but a different variance, this rule would not be considered interesting according to the approach of Aumann et al. [2], but our approach, on the other hand, does.

It is well known in the literature that two distributions differ even though their moments are the same [46]. Figure 6 shows an example of two hypothetical ICU LOS distributions with identical means but differing SDs. In the set, the SD is 5, while in the counter-set, it is 2.5. As a result, the rule based solely on the mean would not have been considered interesting. However, the differing SDs, as an alternative statistical moment, reveal a distinction between the distributions. Moreover, the KLD, offering a more holistic perspective, indicates a significant discrepancy between the set and counter-set. The literature also shows that using the KLD is preferable to using individual moments [47].

Table 6 "Top 10 interesting rules (using minsup 5%) comparison for the MIMIC-IV dataset between generated rules by approach from Aumann et al. and Scenario 2 of our approach." [1] (Own table - originally published in [1]).

Rules (Aumann et al.)	Rules (Scenario 2)
3E0G76Z ∧ 02HV33Z ⇒ 12	3E0G76Z ⇒ 10
51881 ∧ 966 ⇒ 11	966 ⇒ 9
9604 ∧ 966 ⇒ 11	966 ∧ male ⇒ 9
3324 ⇒ 10	966 ∧ EWEMER. ⇒ 8
966 ∧ 3893 ⇒ 10	Z781 ⇒ 8
3E0G76Z ⇒ 10	51881 ⇒ 7
3E0G76Z ∧ Other ⇒ 10	9604 ⇒ 7
3E0G76Z ∧ male ⇒ 10	51881 ∧ EWEMER. ⇒ 7
966 ⇒ 9	J9601 ⇒ 7
966 ∧ 42731 ⇒ 9	9604 ∧ EWEMER. ⇒ 6

Table 7. Top 10 interesting rules (using minsup 5%) comparison for the eICU dataset between generated rules by approach from Aumann et al. and Scenario 2 of our approach.

Rules (Aumann et al.)	**Rules (Scenario 2)**
J96.00 ⇒ 8	J96.00 ⇒ 8
J96.00 ∧ male ⇒ 8	J96.00 ∧ male ⇒ 8
J96.00 ∧ female ⇒ 8	J96.00 ∧ female ⇒ 8
OBSERVATION ADMIT ∧ male ⇒ 8	OBSERVATION ADMIT ⇒ 8
OBSERVATION ADMIT ⇒ 8	OBSERVATION ADMIT ∧ male ⇒ 8
J189 ∧ male ⇒ 7	J18.9 ∧ male ⇒ 7
OBSERVATION ADMIT ∧ female ⇒ 7	N17.9 ⇒ 7
N17.9 ⇒ 7	OBSERVATION ADMIT ∧ female ⇒ 7
J18.9 ⇒ 7	J18.9 ⇒ 7
N17.9 ∧ male ⇒ 7	N17.9 ∧ male ⇒ 7

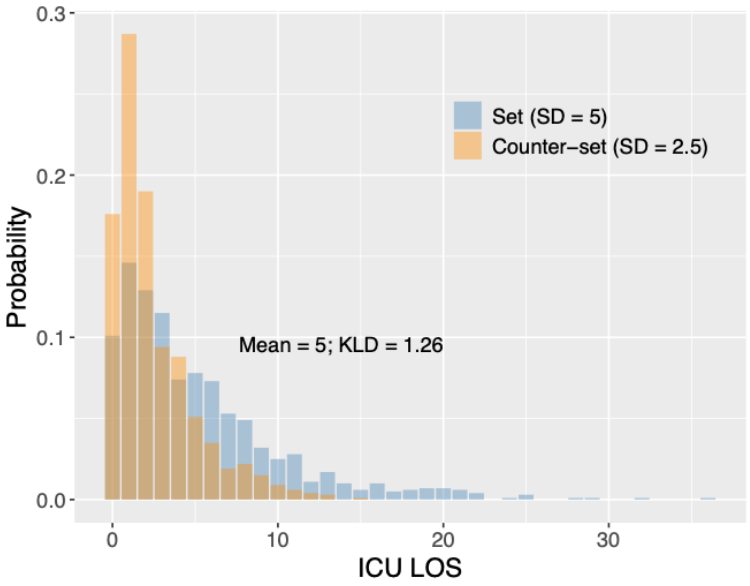


Figure 6. Exemplary distributions with the same means but different SD and high KLD.

Comparing ICU LOS patterns across datasets. In this manuscript, we applied our method to the MIMIC-IV and eICU to predict ICU LOS. To assess the generalizability of our method and the consistency of the discovered rules across datasets, we analyze whether the rules exhibit similar ICU LOS outcomes in both datasets. Specifically, we focus on the top 40% most interesting rules, as ranked by the KLD between set and counter-set. By concentrating on these top-ranked rules, we evaluate whether our method effectively prioritizes rules that are both meaningful within each dataset and consistent in their ICU LOS outcomes across datasets. In addition to focusing on the top-ranked rules, we compare their ICU LOS distributions similarly with those of the full rule set to evaluate how our method enhances consistency in the ICU LOS outcomes (in terms of its distribution) across datasets by introducing a meaningful ranking.

For each rule, we consider its associated ICU LOS values (the set considering our proposed method) and analyze the divergence of these distributions between matched rules from MIMIC-IV and eICU. This analysis not only demonstrates the ability of our method to identify universal patterns but also highlights areas where ICU LOS distributions differ, providing insights into clinical or dataset-specific variations. To compare the ICU LOS distributions, we use again the KLD as a measure of discrepancy, to quantify how well the top-ranked rules align across the two datasets. Since there should be no semantic difference in the context of this analysis, whether we calculate the KLD from the direction of MIMIC-IV to eICU or vice versa, we use the symmetric KLD, which calculates the average KLD from both directions [48].

As our original analysis of the MIMIC-IV dataset included additional categorical variables beyond *gender*, *admission type*, and *diagnoses* (encoded via ICD 9 and ICD 10), a direct comparison of the rules or ICU LOS values with the eICU dataset would not be appropriate. We re-ran our approach on the MIMIC-IV dataset to ensure consistency and comparability, excluding the additional categorical variables mentioned in Table 1. This adjustment allowed us to analyze both datasets using the same set of categorical variables, ensuring a fair and accurate comparison. As mentioned in Section 3, MIMIC-IV includes both ICD-9 and ICD-10 codes, with diagnoses recorded in one of these formats. The eICU dataset typically provides mapped ICD-10 codes alongside the original ICD-9 codes. We leverage this dual coding for matching purposes. If we exclusively use ICD-9 or ICD-10 codes from eICU, the generated rules would be limited to diagnoses in the selected coding system. This would prevent us from matching rules derived from MIMIC-IV that use the alternate ICD version. To address this, we apply our approach to the eICU dataset twice, once using ICD-9 codes and once using ICD-10 codes. We consider both rule sets for the matching with MIMIC-IV.

In the following analysis, we focus on the results generated using Scenario 1 of our approach. Figure 7 illustrates the statistics on the number of rules derived from MIMIC-IV and eICU considering the top 40% interesting rules at various minsup thresholds, as well as the number and percentage of matched rules. Considering the top 40% of interesting rules reflects a balance between maintaining a focus on the most relevant rules and ensuring sufficient matched rules for robust comparison across datasets. This ap-

proach allows us to assess the generalizability of our method while providing meaningful insights into ICU LOS distributions.
Since we generate rules for eICU using both ICD-9 and ICD-10 codes, we always use the larger rule set when calculating the total number of rules and the percentage of matched rules. This approach accounts for the fact that a one-to-one mapping between ICD-9 and ICD-10 codes is not always possible, leading to slight differences in the number of generated rules. When determining the percentage of matched rules between MIMIC-IV and eICU, we use eICU as the reference point, as it has the smaller rule set in all cases (whether based on ICD-9 or ICD-10 rules).
For the full rule set, at a minsup of 2.5%, 642 rules are generated from MIMIC-IV and 117 from eICU, with 89 matches, resulting in 76.1% of eICU rules being matched. At a minsup of 5%, the full set includes 146 rules from MIMIC-IV and 44 from eICU, with 34 matches, corresponding to 77.3% of eICU rules. At 7.5% minsup, the full set contains 62 rules from MIMIC-IV and 23 from eICU, with 19 matches, or 82.6% of eICU rules.

Contrastingly, the top 40% of rules narrow the focus to the most interesting ICU LOS distributions between set and counter-set. At 2.5% minsup, 256 rules are generated from MIMIC-IV and 46 from eICU, with 25 matches, resulting in 54.3% of eICU rules being matched. At 5% minsup, the top 40% includes 58 rules from MIMIC-IV and 17 from eICU, with 12 matches corresponding to 70.6% of eICU rules. At 7.5% minsup, the number of rules decreases to 24 for MIMIC-IV and 9 for eICU, with 5 matches, or 55.6% of eICU rules.

As the minsup value increases, the total number of rules decreases for both datasets, as higher minsup values require itemsets to appear more frequently to meet the threshold, limiting the number of qualifying itemsets. For both the full rule set and the top 40% rules, matching percentages fluctuate slightly, with higher minsup thresholds generally leading to greater alignment. However, stricter thresholds at 7.5% minsup reduce the total overlap, particularly for the top 40% rules.

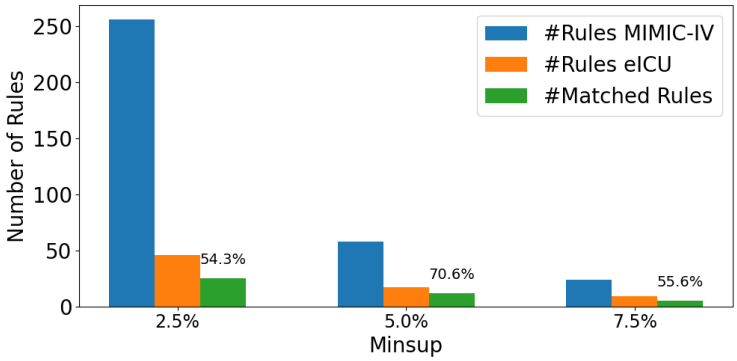


Figure 7. Statistics about generated and matched rules between MIMIC-IV and eICU, considering the top 40% most interesting rules according to our method.

To assess differences in ICU LOS outcomes and validate the effectiveness of our method, we analyze the KLDs of ICU LOS distributions between matching rules from MIMIC-IV and eICU. Figure 8 presents boxplots of the KLD distributions for the top 40% most interesting rules and full rule set across the considered minsup thresholds.

We observe that the top 40% rules exhibit slightly lower median KLD values compared to the full rule set at all minsup thresholds. For instance, at 2.5% minsup, the median KLD decreases from 0.46 for the full set to 0.43 for the top 40% rules. Similarly, at 5% minsup, the median KLD decreases from 0.39 to 0.36, and at 7.5% minsup, it decreases from 0.38 to 0.34. These modest reductions suggest that our method effectively prioritizes rules with slightly better alignment in ICU LOS distributions across datasets.

The boxplots also highlight key differences in variability and outliers. For the full rule set, KLD distributions show a greater number of high outliers, particularly at lower minsup thresholds. For example, at 2.5% minsup, the full set contains multiple outliers, while four exceed 1.0, compared to far fewer and less extreme outliers for the top 40% rules. This indicates that our method filters out noisy, dataset-specific rules, prioritizing rules that represent more consistent trends.

The IQR for the top 40% and full rule set are generally similar, whereby they tend towards a lower KLD. The whiskers for the full set tend to extend slightly lower for minsup 2.5% and 5%, reflecting its larger sample size. For the standard rate of the top 40 %, considering a minsup of 7.5 %, there is no upward whisker, which underlines the low variability of the KLDs, using the ranking from our method. These differences underscore the robustness of our method in curating rules that are both consistent and relevant, even when accounting for sample size differences.

The analysis of KLD distributions demonstrates that our method effectively prioritizes rules that exhibit stronger alignment in ICU LOS outcomes across MIMIC-IV and eICU datasets. The top-ranked rules consistently show lower median KLD values and fewer extreme outliers compared to the full rule set, particularly at lower minsup thresholds, highlighting our method's ability to identify generalizable and robust rules. While the differences in median KLD are modest, the reduction in variability, outliers, and overall shift toward lower KLDs underscores our method's strength in filtering out noisy or dataset-specific rules by searching for interesting rules, based on the KLD between set and counter-set. These findings validate the utility of our method for identifying consistent distributions in ICU LOS outcomes for interesting rules, demonstrating its potential generalizability across diverse datasets.

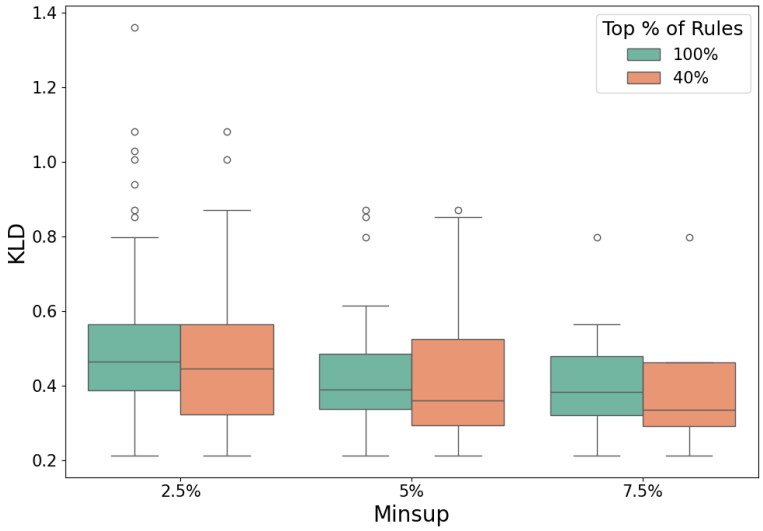


Figure 8. KLD distributions for the matched rules, considering the different minsup values and considering the top 40% most interesting rules according to our method and the full set (100%) of rules.

6 Conclusion

This thesis summarizes our originally published results about processing quantitative variables in *Association Rule Mining* (ARM) [1]. ARM is a data mining method that represents frequently occurring patterns in data unsupervised in the form of human-readable rules (association rules). Rule-based prediction models make predictions comprehensible, which is particularly desirable in domains such as medicine. While the ARM in its original form can only process categorical variables, various approaches have been established in the literature that bypass the restriction. A well-known method in the literature is the approach by Aumann et al. [2], which we have generalized in this manuscript.

We have previously demonstrated our approach in different domains [1], focusing specifically on the prediction of *length of stay of patients in intensive care units* (ICU LOS) in this thesis. ICU LOS describes the number of days a patient spends in the ICU during a hospitalization. In the literature, predicting ICU LOS is an essential parameter for estimating the required hospital resources, which in turn leads to an increase in patient care efficiency. In this thesis, we included two well-known datasets (MIMIC-IV [26] and eICU [27]) for the prediction of ICU LOS.

We propose several directions for future research [1]: Similar to the approach of Aumann et al. [2], our approach is also restricted to the distribution of quantitative variables on the right-hand side of a rule. Thus, extensions that remove this restriction and allow the placement of quantitative variables on the left-hand side of a rule would be interesting. Another interesting direction is comparing the distributions of the quantitative variables of the rules extracted by our approach with distributions extracted by classical regression algorithms. Here, it is particularly interesting to see how much the distributions of the quantitative variables, but also the results of different summary statistics, differ. Another direction would be the integration of further datasets and the prediction of the ICU LOS, including the rule comparison regarding the consistency of outcomes. This would further underline the advantage of the ranking of our approach and its generalizability.

References

[1] S. Mohammed, K. Rubarth, S. K. Piper, F. Schiefenhövel, J.-C. Freytag, F. Balzer, and S. Boie, ‘A statistical method for predicting quantitative variables in association rule mining’, *Inf. Syst.*, vol. 118, p. 102253, Sep. 2023, doi: 10.1016/j.is.2023.102253.
[2] Y. Aumann and Y. Lindell, ‘A Statistical Theory for Quantitative Association Rules’, *J. Intell. Inf. Syst.*, vol. 20, no. 3, pp. 255–283, May 2003, doi: 10.1023/A:1022812808206.
[3] C. Batini, *Data and information quality: dimensions, principles and techniques*. New York, NY: Springer Berlin Heidelberg, 2016.
[4] J. Han and M. Kamber, *Data mining: concepts and techniques*, 3rd ed. Burlington, MA: Elsevier, 2012.
[5] I. H. Witten, E. Frank, and M. A. Hall, *Data mining: practical machine learning tools and techniques*, 3rd ed. in Morgan Kaufmann series in data management systems. Burlington, MA: Morgan Kaufmann, 2011.
[6] R. Agrawal, R. Srikant, H. Road, and S. Jose, ‘Fast Algorithms for Mining Association Rules’.
[7] D. Adhikary and S. Roy, ‘Trends in quantitative association rule mining techniques’, in *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, Kolkata, India: IEEE, Jul. 2015, pp. 126–131. doi: 10.1109/ReTIS.2015.7232865.
[8] Y. Aumann and Y. Lindell, ‘A statistical theory for quantitative association rules’, in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, San Diego California USA: ACM, Aug. 1999, pp. 261–270. doi: 10.1145/312129.312243.
[9] J. M. Luna, P. Fournier-Viger, and S. Ventura, ‘Frequent itemset mining: A 25 years review’, *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 6, p. e1329, 2019, doi: 10.1002/widm.1329.
[10] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, ‘Causability and explainability of artificial intelligence in medicine’, *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 4, p. e1312, 2019, doi: 10.1002/widm.1312.
[11] F. Došilović, M. Brcic, and N. Hlupic, *Explainable Artificial Intelligence: A Survey*. 2018. doi: 10.23919/MIPRO.2018.8400040.
[12] N. Burkart and M. F. Huber, ‘A Survey on the Explainability of Supervised Machine Learning’, *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, Jan. 2021, doi: 10.1613/jair.1.12228.
[13] H. Lakkaraju, S. H. Bach, and J. Leskovec, ‘Interpretable Decision Sets: A Joint Framework for Description and Prediction’, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 1675–1684. doi: 10.1145/2939672.2939874.
[14] ‘EU AI Act: first regulation on artificial intelligence’, Topics | European Parliament. Accessed: Jul. 25, 2024. [Online]. Available: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
[15] X. Yuan, B. P. Buckles, Z. Yuan, and J. Zhang, ‘Mining Negative Association Rules’, 2002.
[16] J. Han, J. Pei, and Y. Yin, ‘Mining frequent patterns without candidate generation’, *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 1–12, Jun. 2000, doi:

10.1145/335191.335372.
[17] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, ‘Algorithms for association rule mining — a general survey and comparison’, *ACM SIGKDD Explor. Newsl.*, vol. 2, no. 1, pp. 58–64, Jun. 2000, doi: 10.1145/360402.360421.
[18] R. Srikant and R. Agrawal, ‘Mining quantitative association rules in large relational tables’, in *Proceedings of the 1996 ACM SIGMOD international conference on Management of data - SIGMOD ’96*, Montreal, Quebec, Canada: ACM Press, 1996, pp. 1–12. doi: 10.1145/233269.233311.
[19] R. J. Miller and Y. Yang, ‘Association rules over interval data’, *ACM SIGMOD Rec.*, vol. 26, no. 2, pp. 452–461, Jun. 1997, doi: 10.1145/253262.253361.
[20] J. R. Hershey and P. A. Olsen, ‘Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models’, in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ’07*, Honolulu, HI: IEEE, Apr. 2007, p. IV-317-IV–320. doi: 10.1109/ICASSP.2007.366913.
[21] C. Röver and T. Friede, ‘Discrete Approximation of a Mixture Distribution via Restricted Divergence’, *J. Comput. Graph. Stat.*, vol. 26, no. 1, pp. 217–222, Jan. 2017, doi: 10.1080/10618600.2016.1276840.
[22] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, ‘Data Mining in Healthcare and Biomedicine: A Survey of the Literature’, *J. Med. Syst.*, vol. 36, no. 4, pp. 2431–2448, Aug. 2012, doi: 10.1007/s10916-011-9710-5.
[23] W. Altaf, M. Shahbaz, and A. Guergachi, ‘Applications of association rule mining in health informatics: a survey’, *Artif. Intell. Rev.*, vol. 47, no. 3, pp. 313–340, Mar. 2017, doi: 10.1007/s10462-016-9483-9.
[24] K. Stone, R. Zwiggelaar, P. Jones, and N. Mac Parthaláin, ‘A systematic review of the prediction of hospital length of stay: Towards a unified framework’, *PLOS Digit. Health*, vol. 1, no. 4, p. e0000017, Apr. 2022, doi: 10.1371/journal.pdig.0000017.
[25] S. Concaro, L. Sacchi, C. Cerra, M. Stefanelli, P. Fratino, and R. Bellazzi, ‘Temporal Data Mining for the Assessment of the Costs Related to Diabetes Mellitus Pharmacological Treatment’.
[26] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L. H. Lehman, L. A. Celi, and R. G. Mark, ‘MIMIC-IV, a freely accessible electronic health record dataset’, *Sci. Data*, vol. 10, no. 1, p. 1, Jan. 2023, doi: 10.1038/s41597-022-01899-x.
[27] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, ‘The eICU Collaborative Research Database, a freely available multi-center database for critical care research’, *Sci. Data*, vol. 5, no. 1, p. 180178, Sep. 2018, doi: 10.1038/sdata.2018.178.
[28] ‘Patient length of stay and mortality prediction: A survey’. Accessed: Dec. 08, 2023. [Online]. Available: https://journals.sagepub.com/doi/epub/10.1177/0951484817696212
[29] A. Azari, V. P. Janeja, and A. Mohseni, ‘Healthcare Data Mining: Predicting Hospital Length of Stay (PHLOS)’, *Int. J. Knowl. Discov. Bioinforma.*, vol. 3, no. 3, pp. 44–66, Jul. 2012, doi: 10.4018/jkdb.2012070103.
[30] M. A. Abd-Elrazek, A. A. Eltahawi, M. H. Abd Elaziz, and M. N. Abd-Elwhab, ‘Predicting length of stay in hospitals intensive care unit using general admission features’, *Ain Shams Eng. J.*, vol. 12, no. 4, pp. 3691–3702, Dec. 2021, doi: 10.1016/j.asej.2021.02.018.
[31] Y. Arabi, ‘A prospective study of prolonged stay in the intensive care unit: predictors and impact on resource utilization’, *Int. J. Qual. Health Care*, vol. 14, no. 5, pp. 403–410, Oct. 2002, doi: 10.1093/intqhc/14.5.403.

[32] A. Suresh, K. V. Harish, and N. Radhika, ‘Particle Swarm Optimization over Back Propagation Neural Network for Length of Stay Prediction’, *Procedia Comput. Sci.*, vol. 46, pp. 268–275, 2015, doi: 10.1016/j.procs.2015.02.020.
[33] ‘Fallpauschalenvereinbarung 2024’, 2024, Accessed: Jan. 24, 2024. [Online]. Available: https://www.g-drg.de/content/download/13325/file/2023_11_06_Vereinbarung%20DRG-System%202024_FPV_final.pdf
[34] ‘§ 17b KHG - Einzelnorm’. Accessed: Jan. 24, 2024. [Online]. Available: https://www.gesetze-im-internet.de/khg/__17b.html
[35] R. R. Sarkar, P. T. Courtney, K. Bachand, P. E. Sheridan, P. J. Riviere, Z. D. Guss, C. R. Lopez, M. G. Brandel, M. P. Banegas, and J. D. Murphy, ‘Quality of care at safety-net hospitals and the impact on pay-for-performance reimbursement’, *Cancer*, vol. 126, no. 20, pp. 4584–4592, 2020, doi: 10.1002/cncr.33137.
[36] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, ‘Interpretable & Explorable Approximations of Black Box Models’, Jul. 04, 2017, *arXiv*: arXiv:1707.01154. doi: 10.48550/arXiv.1707.01154.
[37] I. Nouaouri, A. Samet, and H. Allaoui, ‘Evidential data mining for length of stay (LOS) prediction problem’, in *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, Gothenburg, Sweden: IEEE, Aug. 2015, pp. 1415–1420. doi: 10.1109/CoASE.2015.7294296.
[38] C.-W. Cheng, N. Chanani, J. Venugopalan, K. Maher, and M. D. Wang, ‘icuARM-An ICU Clinical Decision Support System Using Association Rule Mining’, *IEEE J. Transl. Eng. Health Med.*, vol. 1, pp. 4400110–4400110, 2013, doi: 10.1109/JTEHM.2013.2290113.
[39] S. Mohammed, ‘Nachvollziehbare Vorhersage der Verweildauer von Intensivpatienten’, Master thesis, Humboldt University of Berlin, Berlin, 2021.
[40] M. N. Do, ‘Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models’, *IEEE Signal Process. Lett.*, vol. 10, no. 4, pp. 115–118, Apr. 2003, doi: 10.1109/LSP.2003.809034.
[41] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, ‘MIMIC-IV’. PhysioNet. doi: 10.13026/7VCR-E114.
[42] ‘ICD codes overview’. Accessed: Oct. 21, 2024. [Online]. Available: https://gesund.bund.de/en/icd-code-search
[43] A. Awad, M. Bader–El–Den, and J. McNicholas, ‘Patient length of stay and mortality prediction: A survey’, *Health Serv. Manage. Res.*, vol. 30, no. 2, pp. 105–120, May 2017, doi: 10.1177/0951484817696212.
[44] T. J. Pollard, A. E. W. Johnson, J. Raffa, and O. Badawi, ‘The eICU Collaborative Research Database’. physionet.org, 2017. doi: 10.13026/C2WM1R.
[45] T. Pollard, A. Johnson, J. Raffa, L. A. Celi, O. Badawi, and R. Mark, ‘eICU Collaborative Research Database’. PhysioNet. doi: 10.13026/0PZC-DM64.
[46] J. P. Romano and A. F. Siegel, *Counterexamples in probability and statistics*. New York, NY: Chapman & Hall, 1986.
[47] V. Dragalin, V. Fedorov, S. Patterson, and B. Jones, ‘Kullback–Leibler divergence for evaluating bioequivalence’, *Stat. Med.*, vol. 22, no. 6, pp. 913–930, Mar. 2003, doi: 10.1002/sim.1451.
[48] S. Tabibian, A. Akbari, and B. Nasersharif, ‘Speech enhancement using a wavelet thresholding method based on symmetric Kullback–Leibler divergence’, *Signal Process.*, vol. 106, pp. 184–197, Jan. 2015, doi: 10.1016/j.sigpro.2014.06.027.

Eidesstattliche Versicherung

„Ich, Sedir Mohammed, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: (englische Fassung) „A new approach for predicting quantitative variables in association rule mining using the example of predicting the length of stay in intensive care units“ / (deutsche Fassung) „Ein neuer Ansatz für die Vorhersage quantitativer Variablen im Association Rule Mining am Beispiel der Vorhersage der Intensivverweildauer“ selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.
Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren/innen beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) werden von mir verantwortet.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem Erstbetreuer, angegeben sind. Für sämtliche im Rahmen der Dissertation entstandenen Publikationen wurden die Richtlinien des ICMJE (International Committee of Medical Journal Editors; www.icmje.og) zur Autorenschaft eingehalten. Ich erkläre ferner, dass ich mich zur Einhaltung der Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis verpflichte.

Weiterhin versichere ich, dass ich diese Dissertation weder in gleicher noch in ähnlicher Form bereits an einer anderen Fakultät eingereicht habe.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§§156, 161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

Datum Unterschrift

Detailed declaration of contribution

Publication 1: S. Mohammed, K. Rubarth, S. K. Piper, F. Schiefenhövel, J.-C. Freytag, F. Balzer, and S. Boie, 'A statistical method for predicting quantitative variables in association rule mining', Inf. Syst., vol. 118, p. 102253, Sep. 2023, doi: 10.1016/j.is.2023.102253.

Contribution in detail:

As described in Section 1.4, I gained my first insights into the topic of association rule mining as part of my Master thesis. On this basis, I decided to devote further attention to this complex in the context of the publication mentioned above (*Publication 1*) respectively this dissertation and to advance further the area of processing quantitative variables. In the context of my publication, I designed the approach and implemented it programmatically. Kerstin Rubarth and Sophie K. Piper supported me in answering statistical questions. I also conducted the data preparation and data analysis. Sebastian Boie supported me in the selection of the datasets. Using the implementation of the approach and the processed data, I ran various experiments, evaluated the results, and then visualized them. Fridtjof Schiefenhövel mainly supported me in answering medical questions.
Finally, I wrote the first draft of the paper. All co-authors edited the manuscript and provided critical feedback. After several revisions with the journal, the article was published. During the review process, I took responsibility for the correspondence with the reviewers. After the journal announced the acceptance of our manuscript, I produced a final version of the paper, for which I operate as the corresponding author. With this, I took long-term responsibility for written correspondence related to this paper. My supervisors, Felix Balzer, Johann-Christoph Freytag and Sebastian Boie, advised me in all phases of the creation of this publication.

_____

Signature, date, and stamp of first supervising university professor / lecturer

_____

Signature of doctoral candidate

Printing copy of the publication

Information Systems 118 (2023) 102253

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

A statistical method for predicting quantitative variables in association rule mining

Sedir Mohammed [a,*], Kerstin Rubarth [a,d], Sophie K. Piper [a,d], Fridtjof Schiefenhövel [a,b], Johann-Christoph Freytag [c], Felix Balzer [a,1], Sebastian Boie [a,1]

[a] Charité – Universitätsmedizin Berlin, Institute of Medical Informatics, Charitéplatz 1, Berlin, 10117, Germany
[b] Klinikum rechts der Isar, Technical University of Munich, School of Medicine, Department of Anaesthesiology and Intensive Care + Institute of Artificial Intelligence and Informatics in Medicine, Ismaninger Str. 22, Munich, 81675, Germany
[c] Humboldt-Universität zu Berlin, Department of Computer Science, Unter den Linden 6, Berlin, 10099, Germany
[d] Charité – Universitätsmedizin Berlin, Institute of Biometry and Clinical Epidemiology, Charitéplatz 1, Berlin, 10117, Germany

ARTICLE INFO

Article history:
Received 25 January 2023
Received in revised form 4 June 2023
Accepted 3 July 2023
Available online 6 July 2023

Keywords:
Quantitative variables
Data mining
Association rule mining
Frequent itemsets

ABSTRACT

Background: Association rules encode common patterns and structures identified in datasets. They can be derived by association rule mining (ARM) algorithms. The association rules are human-readable and allow comprehensible predictions, unlike many other types of prediction algorithms.
Objective: Classical ARM algorithms, like Apriori or FP-growth, cannot process interval or ratio scaled data (quantitative variables) which limits their applicability.
Results: We address this restriction in classical ARM algorithms, making it possible to process quantitative variables on the right side of a rule. Our approach is based on applying the Kullback–Leibler divergence (KLD) to identify a rule which holistically considers complete data distributions instead of using only summary statistics.

We demonstrate the new approach by using, among others, the example of predicting the length of stay of intensive care patients. The length of stay describes the number of days a patient spends in the intensive care unit. In addition, we further demonstrate our approach by predicting the credit score of bank customers and the contract duration of customers of a fictional telco company based on two publically available datasets.
Conclusion: This paper shows a new approach for predicting quantitative variables in ARM. We demonstrate the new approach using the FP-growth algorithm.
© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

The identification of frequently occurring patterns and structures in data is one of the core tasks of data mining [1]. With the help of association rule mining (ARM) algorithms, it is possible to find rules (association rules) that represent these relationships [2]. One of the first and still well-known ARM methods is the Apriori algorithm [3,4]. Over time, other ARM algorithms or extensions of Apriori have been introduced in the literature, focusing in particular on efficient execution [3,5–7]. One of these algorithms is the FP-growth algorithm, which we use as an example in this work [5]. The basis is generating of frequent itemsets (frequent itemset mining) to extract frequent relationships in the data [3,8]. The discovery of these associations in the form of rules is particularly relevant in the context of the comprehensible prediction of states. Especially in the medical context, it is desirable to understand how a model makes decisions [9]. State-of-the-art predictive models usually represent a black box, where decisions are not explainable to the end-user [9–11]. Since the well-being and health of patients are at risk, medical professionals need to understand why a model makes certain predictions [9]. In contrast, the advantage of association rules is that they generate human-readable rules [12]. Therefore, predictions of the model are comprehensible.

In its original form, however, classical ARM algorithms, like Apriori or FP-growth, can only deal with categorical variables (discrete and typically unordered scaled data) [3–5,13,14]. Processing of quantitative variables (interval or ratio scaled data) either on the left or right side of a rule is not within the scope of most existing approaches [2,4,5,14,15]. Since a restriction to categorical variables is not always possible [4], various methods have been presented in the literature to process quantitative variables with ARM algorithms, which we discuss in detail in Section 2.

* Corresponding author.
E-mail address: sedir.mohammed@charite.de (S. Mohammed).
[1] Contributed equally.

https://doi.org/10.1016/j.is.2023.102253
0306-4379/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

An extension to ARM algorithms to include quantitative variables is proposed by Aumann et al. [4]. When generating association rules, the authors include summary statistics (e.g., arithmetic mean) of quantitative variables. The approach finds an interesting rule where quantitative variables are on the right side if the subset of transactions (considered data) to which the rules apply has a statistically significantly different mean than the counter set of transactions. Aumann et al. use the Z-test to establish statistical significance [4]. However, the validity of using summary statistics to identify interesting rules is questionable. The final rule set depends significantly on the statistical measure chosen. When the mean is applied, the rule set may differ from a rule set that uses the variance. The goal of our work is to identify interesting rules regardless of the chosen summary statistics. We generalize the approach of Aumann et al. to problems where a reduction of a quantitative random variable to summary statistics is insufficient (more details in Section 3). With our approach, we can model the entire distribution of a quantitative variable on the right-hand side, i.e., the part of the rule that we want to predict. The generalization measures the discrepancy using the Kullback–Leibler divergence (KLD) between the distributions of quantitative variables on the right-hand side and the distribution of the quantitative variables of the counter-set of transactions [16, 17]. The counter-set of the transactions on the right-hand side describes the transaction to which the rule does not apply.

We demonstrate our approach using several datasets. First, we apply our method to the MIMIC-IV dataset to predict the length of stay (LOS) of patients in an intensive care unit (ICU), which is also known as intensive care unit length of stay (ICU LOS). LOS respectively ICU LOS describes the number of days a patient is in a hospital during a course of treatment [18]. Due to the day-by-day consideration, LOS and thus also ICU LOS represent a quantitative variable. In general, predicting the LOS of patients is crucial for hospitals to plan needed resources for patient care [18,19].

Furthermore, we demonstrate our approach using the fictional customer data from the Telco Customer Churn dataset to predict the number of months a person has been a customer (tenure) [20]. We also demonstrate our approach based on the Churn for Bank Customers dataset and predict the credit score of bank customers based on it [21].

Our contribution is structured as follows: Section 2 first presents related work, including the approach of Aumann et al. [4]. In Section 3, we generalize the approach from Aumann et al. [4]. The result is a new, more flexible approach to predicting quantitative variables on the right side of a rule. We define two scenarios of our new approach, which we present in detail. We apply the new approach to patient data from the MIMIC-IV to predict ICU LOS, to fictional customer data from Telco Customer Churn to predict tenure, and to bank customer data from Churn for Bank Customers datasets to predict the credit score [20–22] in Section 4. Based on the used datasets, we compare the rules generated with our approach with those generated by the approach of Aumann et al.. We end with a discussion of open questions in Section 4 and summarize our results in Section 5.

2. Related work

This section discusses the different classes of extensions to the ARM algorithms to treat quantitative variables.

2.1. Association rule mining processing of quantitative variables

Since classical ARM algorithms were not designed to process quantitative variables, several modifications have been presented by Adhikary et al. [2]. The authors classify the approaches in different categories; among the most widely used categories are [2]:

Statistical methods, like those of Aumann et al. and our extension, form an independent class for processing quantitative variables [2,4]. The approach of Aumann et al. applies summary statistics over a quantitative variable to predict this variable.

Methods based on data partitioning form another category [2]. The partitioning approaches divide the quantitative variables into intervals. In this context, Srikant et al. define the minsup and the minconf problem, derived from the minimum support and the minimum confidence similar to the Apriori algorithm [15]. A rule has minimum support s and is called a frequent itemset if the ratio of transactions in D that contain both X (left side of a rule) and Y (right side of a rule) are above a predefined threshold [3]. A rule has minimum confidence if the percentage of the transactions that contain X also contain Y is above a predefined threshold [3]. The minsup problem states that individual itemsets do not occur frequently enough to be classified as frequent itemsets if the chosen intervals are too small. The minconf problem states that using too large intervals results in a loss of information. Thus, only general statements are possible about the variables when using these approaches [14].

Clustering approaches form another category for handling quantitative variables [2]. They treat the identified clusters as independent categorical variables within the quantitative variables. However, if they cannot identify separated clusters applying and interpreting the generated association rules is challenging. A clustering approach may also lead to the minsup or minconf problem if the parameters for determining the clusters are chosen inappropriately.

2.2. Interesting rules

Aumann et al. develop extensions to the ARM algorithms for problems (i) with only categorical variables on the left-hand side and quantitative variables on the right-hand side, and (ii) for problems with a single quantitative variable on the left and a single quantitative variable on the right [4]. We consider case (i) and generalize this approach to distributions instead of summary statistics. To keep this article self-contained, we describe the notation and reproduce the relevant definitions from [4] below.

The left-hand side of a rule consists of categorical variables $X \subseteq E_C \times C$, where E_C is the set of categorical variables and C is the complete set of categorical variables values. X is thereby referred to as the profile. A transaction t now has profile X if $X \subseteq t$ holds. T_X denotes a set of transaction with a profile X.

The right-hand side of a rule gives the expected value of each quantitative variable. However, the calculation uses only the values of the transactions in T_X. For a set of quantitative variables $J \subseteq E_Q$, where E_Q represents the set of all quantitative variables, and a set of transactions $T \subseteq D$, where D is the set of all transactions, the $Mean_J(T)$ represents the vector of attributes in J for the given set of transactions T. Mean here means that the approach calculates the mean for each attribute in J for T. Thus, the right-hand side of a rule corresponding to the profile has the form $Mean_J(T_X)$. Due to the structure of a rule, the measurement of the confidence in the form as presented by Srikant et al. is not possible since the right side of a rule is not part of the previously found frequent itemset [3]. Thus, an alternative measure for the evaluation of a rule is needed.

For this reason, we now focus on rules with a so-called extraordinary behavior on the right-hand side. According to Aumann et al. extraordinary means that the mean value of the attributes J in the transactions T_X differs statistically significantly from the mean value of the rest of the data $(D - T_X)$ of the respective attributes. To test whether the mean values are statistically significantly different, the approach uses the Z-test. Only

if the Z-test finds a significant difference is the rule evaluated as exceptional, that is, $Mean_J(T_X) \not\approx Mean_J(D - T_X)$. Rules that show exceptional behavior on the right side are then declared as interesting.

In addition to the significance measurement, Aumann et al. introduce an optional minimum difference of means (mindif) to filter out rules irrelevant to the user [4]. Thus, for any $e \in J$, it must hold that $\|Mean_e(T_x) - Mean_e(D - T_X)\| > mindif$. The adequate choice of mindif requires expert knowledge.

To use our approach, we must compute frequent itemsets first. For this purpose, any ARM algorithm may generate them that assumes only categorical variables. In the context of this work, we use the FP-growth algorithm. However, when calculating the frequent itemsets, only the categorical variables, which are the variables on the left side of the rules, are considered. A second step, which incorporates the quantitative variables on the right, selects those rules that are *interesting*. In Section 3, we define our notion of *interesting*.

3. Results

A search for interesting rules based on summary statistics is generally dependent on the choice of the summary statistic. For example, a rule may be interesting when considering the mean but may not be classified as interesting when using the variance or vice versa. It is not a priori clear which summary statistic to use, and it might require meticulous considerations. We present a generalization of the approach of Aumann et al. which does not use summary statistics but models the entire distribution.

Following the approach of Aumann et al. k-frequent itemsets are still the starting point of our new approach [4]. Instead of evaluating summary statistics over J for the sets T_X and $D - T_X$, we measure the discrepancy between the distributions by the Kullback–Leibler divergence (KLD) [16,17]. The KLD is widely used in statistics to measure the distance between two probability distributions, also known as relative entropy [16,23–25].

The KLD for two discrete probability mass functions or two continuous probability density functions is defined as [16,25]:

$$KLD(p \parallel q) = \sum_{x \in X} p(x) \cdot \log \frac{p(x)}{q(x)}, \tag{1}$$

$$KLD(p \parallel q) = \int_{-\infty}^{\infty} p(x) \cdot \log \frac{p(x)}{q(x)} dx \tag{2}$$

In this generalization, following the principle of Aumann et al. a rule is considered interesting if the measured KLD exceeds a predefined threshold. In these rules, the two sets T_X and $D - T_X$ exhibit different distributions of the quantitative variables J. Note that summary statistics can still be obtained post-hoc via the empirical distribution.

3.1. Generating association rules

We demonstrate our method with quantitative variables on the right-hand side of a rule within the FP-growth algorithm on three different datasets. We emphasize that the frequent itemsets generated by FP-growth serve as input for our approach. However, we used the FP-growth algorithm as an example; any classical ARM algorithm for generating frequent itemsets can be used.

In Section 3.1.1, we first present the datasets we used, as they partially influence the use of our approach. Considering scalability, we implemented both, our approach and the approach of Aumann et al. based on the provided FP-growth implementation in Apache Spark [26,27]. Our implementation is freely available [28]. We run FP-growth with different minsup values. The frequent itemsets, and thus the rules are classified as interesting, depend generally on the choice of minsup values. We demonstrate this relationship on the MIMIC-IV dataset by choosing 2.5%, 5%, and 7.5% for minsup for the experiments. For the experiments on the other datasets (Telco Customer Churn and Churn for Bank Customers), we restrict ourselves to minsup 5%.

Table 1
Used categorical variables of the MIMIC-IV dataset and the possible states of each variable, including the relative number of occurrences of the states. We show the translations of the ICD Codes that are involved in a rule in this paper in Table G.19.

Categorical variable	Possible state
Gender	female (43%), male (57%)
Insurance type	medicaid (8%), medicare (44), other (49%)
Admission type	ambulatory observation (<1%), direct emer. (4%), direct observation (<1%), elective (4%), eu observation (<1%), ew emer. (49%), observation admit (11%), surgical same day admission (13%), urgent (17%)
Diagnoses Procedures	set of ICD 9 or ICD 10 codes

Table 2
Overview of the set minsup values and the number of frequent itemsets after applying the FP-growth algorithm on the MIMIC-IV dataset.

minsup	#frequent itemsets
2.5% (1263)	983
5% (2526)	212
7.5% (3788)	87

3.1.1. Datasets

MIMIC-IV. We used patient data from the MIMIC-IV dataset to predict the length of stay (LOS) of patients in an intensive care unit (ICU), which is also known as intensive care unit length of stay (ICU LOS) [22]. Predicting ICU LOS is a case of prediction of general LOS, which is treated separately in the literature [29–33]. LOS describes the number of days a patient is hospitalized during treatment [18]. On the other hand, ICU LOS includes a patient's LOS during treatment in the ICU. Due to the day-by-day consideration, LOS and, thus also, ICU LOS represent a quantitative variable. In general, predicting the LOS of patients is crucial for hospitals to plan needed resources for patient care [18,19]. From the MIMIC-IV dataset, we use 50504 patient treatments, with each patient treated at the ICU [22] and has at least one International Classification of Diseases (ICD) code for a diagnosis and procedure coded in their patient record. A patient may have multiple stays in the ICU throughout his or her lifetime. Thus, the considered patient treatments correspond to 40357 unique patients. We only consider hospital admissions of patients who did not die during the stay, as death is always accompanied by a shorter LOS. We use several attributes of the patient record as features to predict ICU LOS. Table 1 shows the different categorical variables and the possible states of each variable. Fig. 1 shows the histogram of ICU LOS values of the considered patient treatments. ICU LOS values range from 0 to 110, with a sharp decrease in frequency as ICU LOS increases. In addition, Table 2 lists the absolute values for minsup and the number of frequent itemsets generated by the FP-growth algorithm. For example, the algorithm requires for minsup 5% that a k-itemset exists at least 2526 times. In total, the FP-growth algorithm generated 212 frequent itemsets for minsup 5%. The MIMIC-IV dataset is a pseudonymized data set of real ICU patients. Since this is pseudonymized patient data, access is only provided to registered users (www.physionet.org) after credentialing and a online training course [22].

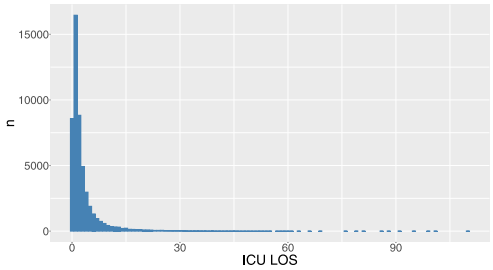


Fig. 1. Distribution of the ICU LOS values of the considered patient treatments.

Table 3
Used categorical variables of the Telco Customer Churn dataset and the possible states of each variable, including the relative number of occurrences of the states.

Categorical variable	Possible state
Gender	female (50%), male (50%)
Senior citizen	yes (25%), no (75%)
Partner	yes (36%), no (64%)
Dependents	yes (17%), no (83%)
Phone service	yes (91%), no (9%)
Paperless billing	yes (75%), no (25%)
Multiple lines	yes (45%), no (45%), no phone service (9%)
Internet service	fiber optic (69%), DSL (25%), no (6%)
Online security	yes (16%), no (78%), no internet service (6%)
Online backup	yes (28%), no (66%), no internet service (6%)
Device protection	yes (65%), no (29%), no internet service (6%)
Tech support	yes (77%), no (17%), no internet service (6%)
Streaming TV	yes (50%), no (44%), no internet service (6%)
Streaming movies	yes (50%), no (44%), no internet service (6%)
Contract	month-to-month (89%), one year (9%), two year (3%)
Payment method	electronic check (57%), mailed check (16%), bank transfer (automatic) (14%), credit card (automatic) (12%)

Telco customer churn. The Telco Customer Churn dataset includes 1869 customer records of a fictional telco company after pre-processing [20]. We use this dataset to predict tenure, the number of months a person has been a customer of this company. Thus, tenure represents a quantitative variable. From this dataset, we consider only those individuals who have already terminated their customer relationship with the company to take the actual number of months per user into account. In Table 3, we have listed the categorical variables we used, including the possible states of these. Fig. 2 shows the histogram of the quantitative variable tenure of the considered customers. The values for the variable tenure range from 1 to 72. Like ICU LOS, the frequency of values here decreases as the tenure value increases. Applying the FP-growth algorithm, we obtain 147,844 frequent itemsets for minsup 5% (94 as absolute value). The Telco Customer Churn dataset is freely available after registration [20].

Churn for bank customers. As a third example, we demonstrate on the Churn for Bank Customers dataset [21]. After pre-processing, this dataset contains data on 7963 bank customers, who have not terminated their relationship with the bank. We predict the customers' credit score, which is a quantitative variable. Table 4 shows which categorical variables we used, including the possible states and the relative number of occurrences of these. Fig. 3 shows the histogram of the considered quantitative variable credit score. The values for the credit score range from 350 to 850. We assume that the frequency distribution can be modeled with a bell curve. The exception is credit score 850, which occurs much more frequent in the data. For this dataset, the algorithm produces for minsup 5% (399 as absolute value) 176 frequent itemsets after applying the FP-growth algorithm. The Churn for Bank Customers dataset is freely available after registration [21].

Table 4
Used categorical variables of the Churn for Bank Customers dataset and the possible states of each variable, including the relative number of occurrences of the states.

Categorical variable	Possible state
Geography	France (53%), Spain (26%), Germany (21%)
Gender	female (43%), male (57%)
#products	1 (46%), 2 (53%), 3 (1%)
Credit card	yes (71%), no (29%)
Is active member	yes (55%), no (45%)

3.2. Finding interesting rules: nonparametric statistics

This section demonstrates our approach by obtaining nonparametric histograms of the considered quantitative variables (ICU LOS, tenure, credit score) from the used datasets (Scenario 1). We show the corresponding pseudocode in Algorithm 1. We consider the k-frequent itemsets mentioned as above for the different minsup values. Based on the generated k-frequent itemsets and the resulting rules for the individual minsup values, we evaluate the KLD for each pair of T_X and $D - T_X$ (generation of set and counter-set in Algorithm 1; Lines 6–12). We calculate the discrepancy between the distribution of the quantitative variable on the right-hand side of a rule (T_X) and its counter-set for each rule ($D - T_X$), according to Eq. (1) (Algorithm 1; Line 14). Here, the distribution over $D - T_X$ corresponds to p, and the distribution over T_X corresponds to q. Based on a predefined KLD threshold, we identify rules which are above the threshold as interesting (Algorithm 1; Lines 15–17).

3.2.1. Example for finding interesting rules: nonparametric statistics

We provide an example for finding interesting rules based on nonparametric statistics. Fig. 4 shows the example, including the different steps from input data to the final rule set. The input data consists of categorical variables and one quantitative variable. We used the FP-growth algorithm to obtain frequent itemsets, however, any ARM algorithm can be used. In this example, we set minsup to 0.5%. Based on the frequent itemsets, our approach assigns the values of the quantitative variable to set and counter-set. From the filled set and counter-set, our approach calculates the probability functions. Our approach calculates the KLD between the counter-set and set based on the probability functions. For this example, we consider rules interesting if the KLD is at least 0.4; thus, the approach filters the rules A and A $\wedge$ C out.

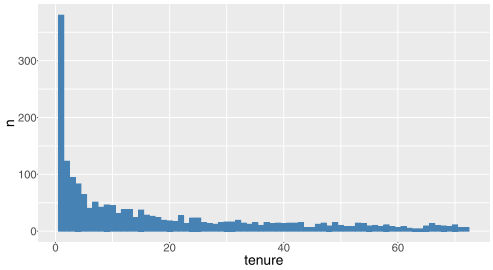


Fig. 2. Distribution of the quantitative variable tenure of the considered customers.

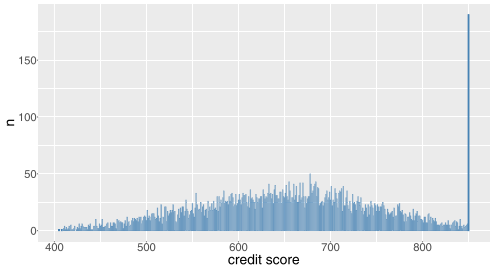


Fig. 3. Distribution of the considered quantitative variable credit score.

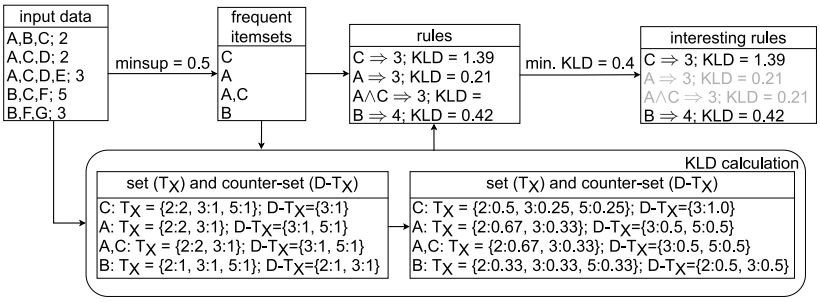


Fig. 4. Example for finding interesting rules based on nonparametric statistics.

3.2.2. Nonparametric statistics: Worst-case runtime complexity analysis

In the following, we analyze the worst-case runtime complexity of our approach based on nonparametric statistics.

We assume in the complexity analysis that the frequent itemsets have already been generated and are used as input for our algorithm. First, we assign all values of the quantitative variable from the input data to the set (T_X) and the counter-set ($D-T_X$). To perform this step, our approach checks for each frequent itemset for each transaction in the input data, whether the itemset is a subset of the categorical variables of the transaction currently under consideration (Algorithm 1; Lines 6–12). The comparison of all frequent itemsets with all transactions from the input data D results in a runtime complexity of

$$\mathcal{O}(|\textit{frequent itemsets}| \cdot |D|) \tag{3}$$

For the subset comparison between a frequent itemset and the categorical variables of a transaction, we consider the *maxmimum length of the frequent itemsets* (α) and the *maxmimum length of the categorical variables of all transactions from the input data* (β). The subset comparison has a complexity of

$$\mathcal{O}(\alpha + \beta) \tag{4}$$

when using a hashmap. This complexity is achieved by the algorithm indexing either the current frequent itemset or the categorical variables of the current transaction in a hashmap. For indexing all items of the respective frequent itemset, the algorithm iterates over the respective list resulting in a complexity of $\mathcal{O}(1)$ based on using a hashmap. Afterward, the algorithm steps through the other list, resulting in linear complexity. We assume a complexity of $\mathcal{O}(1)$ for adding values in T_X or $D - T_X$. Thus, finding and filling the quantitative values in set and counter-set results in a worst-case complexity of

$$\mathcal{O}(|\textit{frequent itemsets}| \cdot |D| \cdot (\alpha + \beta)) \tag{5}$$

To determine the probability functions for computing the KLD between T_X and $D - T_X$, our approach iterates over these lists

Algorithm 1: rule generation with nonparametric statistics (*frequent_itemsets*, *D*, *th*)

Input : Let *frequent_itemsets* be a set of lists, where the inner lists are considered as frequent and consist of categorical variables, *D* the input data, where each entry consists of a tuple where the first element is a list of categorical variables and the second is a quantitative value and *th* minimum KLD threshold for a rule to be considered as interesting

Output: *rule_set* is a set of tuples, where the first element of each tuple is a list of categorical variables and second element is a list of quantitative values

```
1  rule_set ← [ ];
2  set ← [ ];
3  c_set ← [ ];
4  foreach fi ∈ frequent_itemsets do
5      foreach entry ∈ D do
6          categorical_variables ← entry.categorical_variables;
7          quantitative_variable ← entry.quantitative_variable;
8          if ∀fi.entries ⊆ categorical_variables then
9              set.append(quantitative_variable);
10         else
11             c_set.append(quantitative_variable);
12         end
13     end
14     KLD ← calculate_KLD(set, c_set);
15     if KLD ≥ th then
16         rule_set.append((fi, set));
17     end
18 end
19 return rule_set;
```

at least once in each case. We consider the maximum length of T_X ($max|T_X|$) and $D - T_X$ ($max|D - T_X|$) for complexity analysis. We assume a complexity of $\mathcal{O}(1)$ for the pure KLD calculation (Algorithm 1; Line 14). Since the approach performs the KLD calculation according to the number of frequent itemsets, the total complexity is:

$$\mathcal{O}(|frequent\ itemsets| \cdot (max|T_X| + max|D - T_X|)) \tag{6}$$

For checking the threshold check and potentially appending the results to the final rule set, we also assume a complexity of $\mathcal{O}(1)$ (Algorithm 1; Lines 15–17). Combining the determined complexities and applying the distributive law yields a worst-case runtime complexity of:

$$\mathcal{O}(|frequent\ itemsets| \cdot (|D| \cdot (\alpha + \beta) + (max|T_X| + max|D - T_X|))) \tag{7}$$

3.2.3. Nonparametric statistics: sample results

We apply our approach to the presented datasets to predict ICU LOS, tenure, and credit score, respectively. For the ICU LOS values, we apply a day-by-day binning; therefore, we look at the ICU LOS values by day. According to the same principle, we apply a month-by-month binning for the tenure variable and a score-by-score binning for the credit score variable. Based on KLD, the classification of the interestingness of a rule is then performed. In Figs. 5 and 6, we show for the considered minsup values the distributions of the respective quantitative variable within T_X and $D - T_X$ with the highest KLD values. We show the distributions of the respective quantitative variable with the second-highest KLD in Appendix D.

Correspondingly, Table 5 presents the rules associated with the distributions in Figs. 5 and 6. We show the rules with the second-highest KLD in Appendix D. Besides the Rule, we show a set of summary statistics (mean (rounded up to integer), standard deviation (SD), interquartile range (IQR)) applied to T_X and the support (sup.). In addition, we plot the error bars of the KLD values generated by a non-parametric bootstrap with 10,000 samples (rounded to two decimals after the comma). For example, Rule (e) shows that customers that have purchased one product from the bank and are located in France have an average credit score of 650 with a standard deviation of 96.48.

3.3. Finding interesting rules: parametric statistics

Alternatively, we show that the approach can be used when parametric models for the sets T_X and $D - T_X$ are suitable (Scenario 2). We present the corresponding pseudocode in Algorithm 2. We obtain interesting rules in a two-step approach.

In what follows, we assume that a suitable parametric distribution is known for T_X and $D - T_X$. Similarly, we assume that the values for T_X and $D - T_X$ have already been determined (assignment of the values to T_X and $D - T_X$ in Algorithm 2; Lines 6–12). The first step aims to find the rules for which the sets and the chosen parametric descriptions show a good fit (Algorithm 2; Lines 14–16). We again use the KLD to evaluate the goodness of fit. The sets T_X and $D - T_X$ correspond in each calculation to p, while the theoretical distribution corresponds to q. The theoretical distributions are continuous, while the binned distributions of the sets T_X and $D - T_X$ are discrete. We introduce a maximum threshold for KLD between the binned distribution for T_X and $D - T_X$ and their respective parametric descriptions. After the first step, rules removed whose distribution of the sets T_X and $D - T_X$ differs from the parametric distribution too much according to the threshold (Algorithm 2; Line 16).

The second step measures the discrepancy between the two parametric distributions for T_X and $D - T_X$ (Algorithm 2; Line 17). Rules are classified as interesting if the calculated KLD between the parametric descriptions is above a predefined threshold. Based on the predefined KLD threshold between parametric distributions, we filter out these rules where the KLD is lower than the threshold (Algorithm 2; Lines 18–20). We consider the remaining rules as interesting (Algorithm 2; Line 23).

Since only the evaluation of the KLD changes, we do not provide a separate example for the approach based on parametric statistics. Other than the KLD evaluation it is analogous to Section 3.2.1.

Typically, ICU LOS is modeled by an exponential distribution [18]. The histogram in Fig. 1 also shows that this is the case. Fig. 2 shows that tenure can also be approximated well by an exponential distribution. The bell curve of the credit score, on the other hand, suggests a good approximation by a normal distribution (see Fig. 3).

In the following, we show the calculation of the KLD in the first step between the discrete values in T_X and $D - T_X$ and an exponential (case for ICU LOS and tenure) or normal distribution (case for credit score). In the second step, we show the calculation of the KLD between two exponential or normal distributions.

Exponential distribution as parametric statistic. For the exponential distribution we consider the density function $q(x)$ [34]

$$q(x) = \lambda e^{-\lambda x} \tag{8}$$

and the empirical distributions of ICU LOS and tenure values are again obtained by binning. The parameter λ is derived from the expected value of the respective set.

The following equation results after substituting the density function (8) into Eq. (2) and rearranging terms (detailed transformation is shown in Appendix A):

$$KLD(p \parallel q) = \sum_{i=0}^{\infty} p(i) \cdot \left(\log p(i) - \log \lambda + \frac{\lambda}{2}(2i + 1) \right) \tag{9}$$

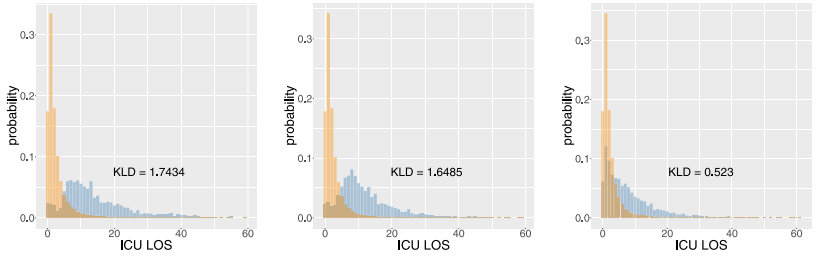


Fig. 5. Scenario 1: Distribution of values from MIMIC-IV in T_X and $D - T_X$ for the rules with the highest KLD according to Eq. (1). Blue corresponds to the distribution of values in T_X, and orange corresponds to values in $D - T_X$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

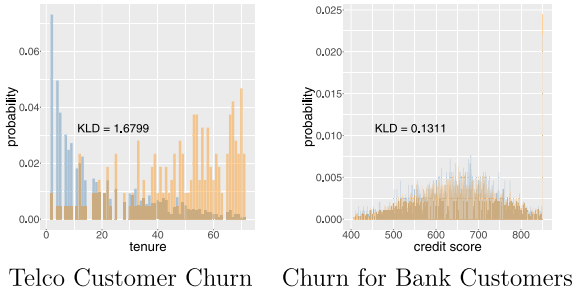


Fig. 6. Scenario 1: Distribution of values from Telco Customer Churn and Churn for Bank Customers dataset in T_X and $D - T_X$ for the rules with the highest KLD according to Eq. (1). Blue corresponds to the distribution of values in T_X, and orange corresponds to values in $D - T_X$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Scenario 1: The rules for each considered dataset and minsup value with the second-highest KLD between T_X and $D - T_X$ associated with the distributions in Figs. 5 and 6.

	Rule (left side)	Mean	SD	IQR	sup.	KLD	minsup
MIMIC-IV							
(a)	5A1955Z	15	11.08	12	1521	1.74 ± 0.07	2.5%
(b)	9672	12	9.37	9	2712	1.65 ± 0.05	5%
(c)	966	8	8.44	9	4461	0.52 ± 0.02	7.5%
Telco Customer Churn							
(d)	Month-to-month	15	15.83	19	1655	1.68 ± 0.19	5%
Churn for Bank Customers							
(e)	1 (#products) ∧ France	650	96.48	134.75	1950	0.13 ± 0.02	5%

As part of the first step, Eq. (9) calculates the KLD between T_X or $D - T_X$ and the optimal parametric exponential distribution.

We define the threshold for the KLD in the first step as exemplary at 0.15 for the MIMIC-IV and 0.3 for the Telco Customer Churn dataset (as the threshold of 0.15 is too strict for this dataset).

For the second step, we calculate the KLD between the two exponential distributions for T_X and $D - T_X$. Combining Eqs. (2) and (8) yields (detailed transformation is shown in Appendix B)

$$KLD(p \parallel q) = \frac{\lambda_q}{\lambda_p} - \log\left(\frac{\lambda_p}{\lambda_q}\right) - 1, \tag{10}$$

where λ_p is the expected value of the respective quantitative variable on $D - T_X$ and λ_q the expected value of the respective quantitative variable on T_X.

Normal distribution as parametric statistic. For the normal distribution we consider the density function $q(x)$ [35]:

$$q(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2} \tag{11}$$

and the empirical distribution of credit score is again obtained by score-by-score binning. The parameters σ and μ are derived from the values of the respective set.

The calculation of the KLD between a discrete set and the normal distribution yields by substituting the density function (11) in Eq. (2) in the following (detailed transformation is shown in Appendix C):

$$KLD(p \parallel q) = \sum_{i=350}^{\infty} p(i) \cdot \left(\log p(i) - \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2}\left(\frac{(-\mu + i + 1)^3 - (i - \mu)^3}{3\sigma^2}\right)\right) \tag{12}$$

We consider 350 as the starting point for i, as it is the minimum credit score in the dataset. For the Churn for Bank Customers dataset we define a KLD threshold of 0.3 in the first step.

Algorithm 2: rule generation with parametric statistics (*frequent_itemsets*, *D*, *th*1, *th*2, *p_function*)

Input : Let *frequent_itemsets* be a set of lists, where the inner lists are considered as frequent and consist of categorical variables, *D* the input data, where each entry consists of a tuple where the first element is a list of categorical variables and the second is a quantitative value, *th*1 the maximum KLD between nonparametric and parametric distribution, *th*2 the minimum KLD between two nonparametric distributions and *p_function* a parametric probability density function

Output: *rule_set* is a set of tuples, where the first element of each tuple is a list of categorical variables and second element is a list of quantitative values

```
rule_set ← [ ];
set ← [ ];
c_set ← [ ];
foreach fi ∈ frequent_itemsets do
    foreach entry ∈ D do
        categorical_variables ← entry.categorical_variables;
        quantitative_variable ← entry.quantitative_variable;
        if ∀fi.entries ⊆ categorical_variables then
            set.append(quantitative_variable);
        else
            c_set.append(quantitative_variable);
        end
    end
    KLD_np_p_set ← calculate_KLD(set, p_function);
    KLD_np_p_c_set ← calculate_KLD(c_set, p_function);
    if KLD_np_p_set ≤ th1 & KLD_np_p_c_set ≤ th1 then
        KLD_p_p ← calculate_KLD(set, c_set, p_function);
        if KLD_p_p ≥ th2 then
            rule_set.append((fi, set));
        end
    end
end
return rule_set;
```

After combining Eqs. (2) and (11) yields in (transformation showed by Belov et al. [35]):

$$KLD(p \parallel q) = \frac{1}{2}\left(\log\frac{\sigma_q^2}{\sigma_p^2} + \frac{\sigma_p^2}{\sigma_q^2} + \frac{(\mu_p - \mu_q)^2}{\sigma_q^2} - 1\right), \tag{13}$$

where p again refers to the quantitative variable on $D - T_X$ and q to the respective quantitative variable on T_X, respectively.

3.3.1. Parametric statistics: worst-case runtime complexity analysis

The parametric-statistics-based approach contains identical steps as the nonparametric statistics-based approach despite additional operations that we can summarize with a complexity of $\mathcal{O}(1)$. Given these complexities, this approach's worst-case runtime complexity is identical to that of the nonparametric statistics-based approach presented in Eq. (7).

3.3.2. Parametric statistics: sample results

Figs. 7 and 8 show the exponential or the normal distributions for the considered datasets with the highest KLD values approximating the sets T_X and $D - T_X$, including the discrete distributions of the sets. Again, we show the distributions of the respective quantitative variable with the second-highest KLD in Appendix E.

According to the set threshold for KLD in the first step, the exponential distributions (darker lines) represent a good approximation for the respective set.

Table 6 shows the rules corresponding to the distributions in Figs. 7 and 8. We show the rules with the second-highest KLD in Appendix E. In addition to the rules, we show summary statistics applied to T_X, the support, and the error bars for the KLD. For example, Rule (a) shows that a patient will spend an average of 11 days at the ICU if assigned the ICD codes 02HV33Z and 3E0G76Z.

4. Discussion

ARM is a commonly used approach for finding predictive patterns in very large datasets [8,36]. However, ARM algorithms in their original form are not suitable for quantitative variables [3,4,14]. Aumann et al. present an extension, which allows the processing of quantitative variables within ARM algorithms by using summary statistics [4]. This extension classifies rules as interesting if the mean of the set T_X is statistically significantly different from the counter-set $D - T_X$. We emphasize that other summary statistics can also be used [4]. However, the set of interesting rules depends on the choice of the summary statistics. We developed a generalization of the approach that considers the entire distribution of the quantitative variable on the right side of a rule. We demonstrated our approach using the ICU LOS, tenure and credit score as quantitative variables in two scenarios. We measure the discrepancy between two probability distributions using the KLD [16].

In the first scenario (Scenario 1), we do not assume a parametric description of the quantitative variable and calculate the KLD values from binned histograms between the set where the rules apply and the counter-set. A high KLD identifies interesting rules.

In the second scenario (Scenario 2), we assume parametric descriptions for the quantitative variable on the right. A minimal KLD between a parametric distribution and the data selects the parameters for an optimal fit for set and counter-set. Subsequently, we use the parametric distributions to estimate the KLD between set and counter-set and identify interesting rules by a large value.

We demonstrated our approach by predicting ICU LOS, tenure and credit score. In this context, we used several categorical variables, which we presented in Tables 1, 3 and 4. While Scenario 1 explicitly searches for distributions for which the KLD is exceptionally high, Scenario 2 first searches for rules where an exponential distribution (for MIMIC-IV and Telco Customer Churn) or normal distribution (for Churn for Bank Customers) approximates the quantitative variables well.

In the following, we compare the rule sets produced by our method with those generated by the approach from Aumann et al.. We compare the rule sets between Scenario 1 and Scenario 2 of our approach and the rule sets when using the approach of Aumann et al. for each considered dataset, using the different minsup values. In the rule sets, we consider the top 10 most interesting rules. For our approach, this selection includes the rules that have the highest KLD. For the approach of Aumann et al. we consider the rules with the highest difference in the expected values. For the right-hand side of a rule, we calculate the expected value of set in each case. We implemented the approach of Aumann et al. following the same model as our approach based on the FP-growth algorithm in Apache Spark. Unlike Aumann et al. introduced their approach, we do not test for the significance of the difference in expected values between set and counter-set in these experiments. This filter could interfere with the results as it filters out rules that our approach considers.

Table 7 shows the generated rules for the Telco Customer Churn dataset with minsup 5% based on the approach of Aumann

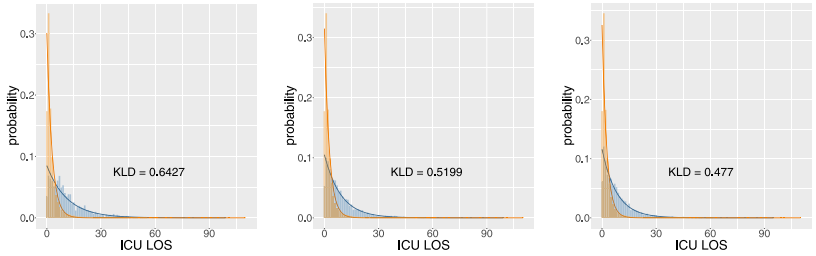


Fig. 7. Scenario 2: Distributions from the MIMIC-IV dataset of the sets T_X and $D - T_X$ including their approximating exponential distributions for the rules with the highest KLD according to Eq. (10) between the exponential distributions. Blue corresponds distributions in context of T_X, and orange corresponds to distributions in context of $D - T_X$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

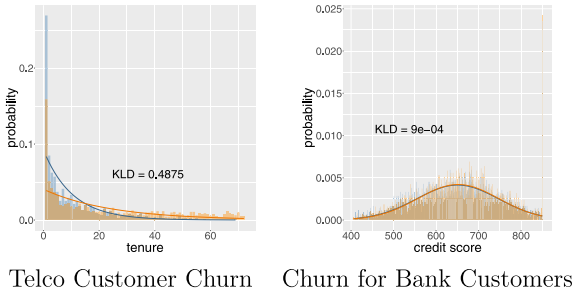


Fig. 8. Scenario 2: Distributions of values from the Telco Customer Churn and Churn for Bank Customers dataset of the sets T_X and $D - T_X$ for the respective rule with the second-highest KLD, including their approximating exponential or normal distribution for the rule with the highest KLD according to Eq. (10) or Eq. (13) between the exponential or normal distribution. Blue corresponds to distributions in the context of T_X, and orange corresponds to distributions in the context of $D - T_X$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
Scenario 2: The rules for each minsup value with the highest KLD between the approximating exponential distributions of the sets T_X and $D - T_X$ associated with the distributions in Figs. 7 and 8.

	Rule (left side)	Mean	SD	IQR	sup.	KLD	minsup
MIMIC-IV							
(a)	02HV33Z ∧ 3E0G76Z	11	10.55	11	1435	0.6427 ± 0.04	2.5%
(b)	3E0G76Z	9	9.42	9	3055	0.5199 ± 0.02	5%
(c)	966	8	8.44	9	4461	0.477 ± 0.02	7.5%
Telco Customer Churn							
(d)	no (online backup) ∧ no (senior citizen) ∧ Month-to-Month	11	12.26	14	837	0.488 ± 0.05	5%
Churn for Bank Customers							
(e)	France	652	96.34	135	4204	0.0009 ± 0.01	5%

et al. and compares them with the generated rules based on Scenario 1 (Appendix F shows the tables for the other considered combinations between datasets, scenarios and minsup values).

The tables show that the top 10 interesting rules for the Telco Customer Churn dataset and the Churn for Bank Customers dataset differ in every point, regardless of the order. For the MIMIC-IV dataset, the tables show partial overlap among the top 10 interesting rules, and in some cases, even the position in the list is identical. A domain expert has confirmed that rule selection based on the KLD, i.e., the discrepancy in the distribution of set and counter-set, is reasonable in the context of ICU LOS prediction.

Overall, these overlaps appear relatively rare among the considered rules. However, this is not unexpected since our approach considers the entire distribution of set and counter-set when selecting the rules of interest and not just the first moment. It is instructive to consider the differences between selecting interesting rules based on the difference of means vs. KLD when a normal distribution can model distributions. One can imagine a scenario where an association rule divides a domain into set and counter set, where the quantitative variable has the same mean but different variance. It is clear from Eq. (13) that such rules would not be considered interesting according to Aumann et al. but would appear in the set of interesting rules according to our approach.

In general, it is well-known that two distributions may be different even though their moments are the same [37]. In other domains it has been recognized that using the KLD over

Table 7
Top-10 rules comparison for the Telco Customer Churn dataset between generated rules by approach from Aumann et al. and Scenario 1 of our approach.

Rules (Aumann et al.)	Rules (Scenario 1)
one year ∧ yes (streaming movies) ∧ fiber optic ⇒ 52	Month-to-Month ⇒ 15
one year ∧ yes (streaming movies) ∧ fiber optic ∧ yes (phone service) ⇒ 52	yes (online backup) ∧ yes (streaming TV) ∧ yes (streaming movies) ⇒ 41
one year ∧ fiber optic ⇒ 52	yes (online backup) ∧ yes (streaming TV) ∧ yes (streaming movies) ∧ yes (phone service) ⇒ 41
one year ∧ fiber optic ∧ yes (phone service) ⇒ 52	yes (online backup) ∧ yes (streaming TV) ∧ yes (streaming movies) ∧ yes (phone service) ∧ fiber optic ⇒ 41
one year ∧ yes (streaming movies) ∧ no (dependents) ⇒ 51	yes (online backup) ∧ yes (streaming TV) ∧ yes (streaming movies) ∧ fiber optic ∧ yes (phone service) ⇒ 41
one year ∧ yes (streaming movies) ∧ yes (paperless billing) ⇒ 50	yes (device protection) ∧ yes (partner) ∧ yes (streaming TV) ∧ yes (phone service) ⇒ 40
one year ∧ yes (streaming movies) ∧ yes (phone service) ⇒ 49	yes (online backup) ∧ yes (device protection) ∧ yes (phone service) ⇒ 39
one year ∧ no (dependents) ∧ yes (phone service) ⇒ 49	yes (partner) ∧ yes (streaming TV) ∧ yes (multiple lens) ∧ no (senior citizen) ∧ yes (phone service) ⇒ 40
one year ∧ yes (streaming movies) ⇒ 49	yes (partner) ∧ yes (streaming TV) ∧ yes (multiple lens) ∧ no (senior citizen) ⇒ 40
one year ∧ yes (streaming TV) ∧ yes (streaming movies) ∧ yes (phone service) ⇒ 49	yes (online backup) ∧ yes (streaming TV) ∧ yes (paperless billing) ⇒ 38

comparing individual moments is preferable [38]. In this sense, our approach generalizes the approach by Aumann et al..

5. Conclusion

In this paper, we developed a generalization of the approach from Aumann et al. for predicting quantitative variables in ARM. Classical ARM algorithms have the restriction that they cannot process quantitative variables [2–4,14]. ARM is a method known from data mining for generating human-readable rules that reveal frequently occurring relationships and correlations in the data (association rules) [3,39]. The generation of interpretable rules explains the decisions of the predictive model. In contrast, there are classes of predictive models with decisions of the model that are not explainable [9]. However, in domains such as medicine, it is desirable to have knowledge about the reasons for a model's predictions [9].

We demonstrated our approach by predicting the ICU LOS, tenure, and credit score. Tenure refers to the contract term of customers. The credit score represents the creditworthiness of bank customers. ICU LOS refers to the number of days a patient is in an intensive care unit. The ICU LOS is especially an interesting parameter in the context of patient treatment. Estimating the ICU LOS serves to optimize resource utilization in the hospital and thus being able to provide more efficient healthcare. It is also possible to identify patterns that lead to high ICU LOS [29].

However, similar to the approach of Aumann et al. our method cannot handle quantitative variables on the left-hand side of a rule but only categorical features. Extensions that permit quantitative variables within a rule on the left side are an interesting avenue for further research.

Another valuable direction is to compare the distributions of the quantitative variables, in the interesting rules obtained by our approach, with distributions derived by applying classical regression algorithms. In this context, it would be interesting to determine how much the distributions and various summary statistics differ.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Johann-Christoph Freytag collaborated with Gottfried Vossen (https://dl.gi.de/bitstream/handle/20.500.12116/20491/lni-p-144-komplett.pdf?sequence=1) [2009]; Johann-Christoph Freytag collaborated with Matthias Weidlich (https://doi.org/10.1145/3242153.3242155) [2018]; Johann-Christoph Freytag collaborated with Holger Pirk (https://doi.org/10.14778/3007328.3007332) [2016]; Johann-Christoph Freytag collaborated with Theo Härder (https://doi.org/10.1007/s13222-016-0228-3) [2016]; Johann-Christoph Freytag collaborated with Volker Markl (https://doi.org/10.1007/s00778-014-0357-y [2014], https://doi.org/10.1109/ICDE.2013.6544927 [2013]); Johann-Christoph Freytag collaborated with Themis Palpanas (https://doi.org/10.1145/2536669.2536673) [2013]

Data availability

Among others, we used patient data from the MIMIC-IV database. We published our code for generating association rules on GitHub. We shared the links in the manuscript.

Appendix A. Transformation of the Kullback–Leibler divergence where p is discrete and q is an exponential density function

As already mentioned, we consider the density function of the exponential distribution $q(x)$ [34]:

$$q(x) = \lambda e^{-\lambda x} \tag{A.1}$$

and the distribution of ICU LOS values $p(x)$ inside a frequent itemset.

First, the lower interval limit of the original KLD Eq. (2) can be constrained since the ICU LOS takes at least the value 0:

$$KLD(p \parallel q) := \int_0^\infty p(x) \cdot \log \frac{p(x)}{q(x)} dx \tag{A.2}$$

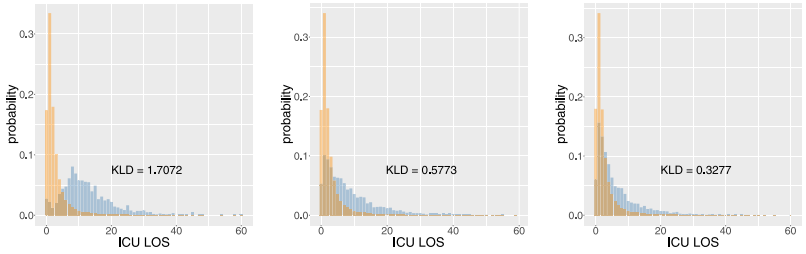


Fig. D.9. Distribution of values from MIMIC-IV in T_X and $D - T_X$ for the rules with the second-highest KLD according to Eq. (1). Blue corresponds to the distribution of values in T_X, and orange corresponds to values in $D - T_X$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

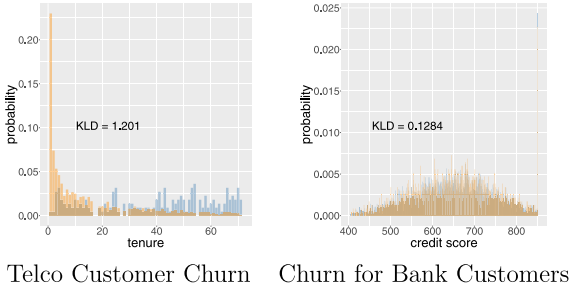


Fig. D.10. Scenario 1: Distribution of values from Telco Customer Churn and Churn for Bank Customers dataset in T_X and $D - T_X$ for the rules with the second-highest KLD according to Eq. (1). Blue corresponds to the distribution of values in T_X, and orange corresponds to values in $D - T_X$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table D.8
Scenario 1: The rules for each considered dataset and minsup value with the second-highest KLD between T_X and $D - T_X$ associated with the distributions in Figs. D.9 and D.10.

	Rule (left side)	Mean	SD	IQR	sup.	KLD	minsup
MIMIC-IV							
(a)	EW EMER. ∧ 966 ∧ 9672	13	8.71	9	1391	1.71 ± 0.08	2.5%
(b)	3E0G76Z	9	9.42	9	3055	0.58 ± 0.02	5%
(c)	51881	7	7.62	7	4243	0.33 ± 0.01	7.5%
Telco Customer Churn							
(d)	yes (online backup) ∧ yes (streaming TV) ∧ yes (streaming movies)	41	21.89	35	221	1.2 ± 0.42	5%
Churn for Bank Customers							
(e)	yes (credit card)	652	95.71	131	5631	0.13 ± 0.02	5%

Based on equation

$$\int_0^\infty f(x)dx = \sum_{i=0}^{\infty} \int_i^{i+1} f(x)dx \tag{A.3}$$

we obtain the following rewriting of Eq. (2):

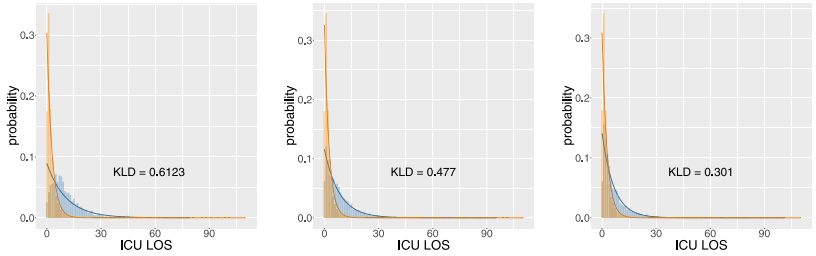
$$\begin{aligned} KLD(p \parallel q) &:= \int_0^\infty p(x) \cdot \log \frac{p(x)}{q(x)} dx \\ &= \sum_{i=0}^{\infty} p(i) \cdot \left(\log p(i) - \int_i^{i+1} \log q(x)dx \right) \end{aligned} \tag{A.4}$$

Now we have to solve the integral of $\int_i^{i+1} \log q(x)dx$. For this $q(x)$ is substituted:

$$\begin{aligned} \int_i^{i+1} \log q(x)dx &= \int \log \lambda e^{-\lambda x} dx \\ &= \int \log \lambda dx + \int -\lambda x dx \\ &= \log \lambda - \frac{\lambda}{2}(2i+1) \end{aligned} \tag{A.5}$$

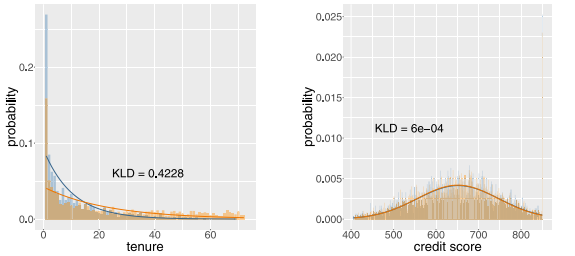
The substitution of the results in (A.4) provide

$$\begin{aligned} KLD(p \parallel q) &= \sum_{i=0}^{\infty} p(i) \cdot \left(\log p(i) - \int_i^{i+1} \log q(x)dx \right) \\ &= \sum_{i=0}^{\infty} p(i) \cdot \left(\log p(i) - \log \lambda + \frac{\lambda}{2}(2i+1) \right) \end{aligned} \tag{A.6}$$



(a) minsup=2.5% (b) minsup=5% (c) minsup=7.5%

Fig. E.11. Scenario 2: Distributions from the MIMIC-IV dataset of the sets T_X and $D - T_X$ including their approximating exponential distributions for the rules with the second-highest KLD according to Eq. (10) between the exponential distributions. Blue corresponds distributions in context of T_X, and orange corresponds to distributions in context of $D - T_X$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



Telco Customer Churn Churn for Bank Customers

Fig. E.12. Scenario 2: Distributions of values from the Telco Customer Churn and Churn for Bank Customers dataset of the sets T_X and $D - T_X$ for the respective rule with the second-highest KLD, including their approximating exponential or normal distribution for the rule with the highest KLD according to Eq. (10) or Eq. (13) between the exponential or normal distribution. Blue corresponds to distributions in the context of T_X, and orange corresponds to distributions in the context of $D - T_X$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Appendix B. Transformation of the Kullback–Leibler divergence based on two exponential density function

We consider the density function of the exponential distribution $p(x)$ and $q(x)$ [34]:

$$p(x) = \lambda_p e^{-\lambda_p x}, \tag{B.1}$$

$$q(x) = \lambda_q e^{-\lambda_q x} \tag{B.2}$$

First, the lower interval limit of the original KLD Eq. (2) can be constrained since the ICU LOS or the tenure takes at least the value 0 or 1:

$$KLD(p \parallel q) := \int_0^\infty p(x) \cdot \log \frac{p(x)}{q(x)} dx \tag{B.3}$$

Based on Eq. (B.1) and (B.2) we obtain the following equation of the KLD based on Eq. (B.3):

$$\begin{aligned} KLD(p \parallel q) &:= \int_0^\infty \lambda_p e^{-\lambda_p x} \cdot \log \frac{\lambda_p e^{-\lambda_p x}}{\lambda_q e^{-\lambda_q x}} dx \\ &= \int_0^\infty \lambda_p e^{-\lambda_p x} \cdot (\log \lambda_p e^{-\lambda_p x} - \log \lambda_q e^{-\lambda_q x}) dx \\ &= (\log \lambda_p - \log \lambda_q) \cdot \lambda_p \int_0^\infty e^{-\lambda_p x} dx \\ &\quad + (\lambda_q - \lambda_p) \cdot \lambda_p \int_0^\infty x e^{-\lambda_p x} dx \end{aligned} \tag{B.4}$$

The integral $\int_0^\infty$ in Eq. (B.4) over the density function of the exponential distribution results in 1, therefore

$$KLD(p \parallel q) = (\log \lambda_p - \log \lambda_q) + \lambda_p(\lambda_q - \lambda_p) \int_0^\infty x e^{-\lambda_p x} dx \tag{B.5}$$

By considering the integration by parts rule:

$$\int_0^\infty uv' = \Big[uv\Big]_0^\infty - \int_0^\infty u'v \tag{B.6}$$

we obtain the following rewriting of Eq. (B.5):

$$\begin{aligned} KLD(p \parallel q) &= (\log \lambda_p - \log \lambda_q) + \lambda_p(\lambda_q - \lambda_p)\Big\{\Big[x \cdot \Big(-\frac{1}{\lambda_p}\Big) e^{-\lambda_p x}\Big]_0^\infty \\ &\quad - \int_0^\infty \Big(-\frac{1}{\lambda_p}\Big) e^{-\lambda_p x} dx\Big\} \\ &= (\log \lambda_p - \log \lambda_q) \\ &\quad + \lambda_p(\lambda_q - \lambda_p)\Big\{0 + \frac{1}{\lambda_p} \int_0^\infty e^{-\lambda_p x} dx\Big\} \end{aligned} \tag{B.7}$$

By calculating the primitive function over the integral $\int_0^\infty$ in Eq. (B.7), we get the following Eq. for the KLD:

$$\begin{aligned} KLD(p \parallel q) &= (\log \lambda_p - \log \lambda_q) + \lambda_p(\lambda_q - \lambda_p) \cdot \Big(\frac{1}{\lambda_p}\Big)\Big\{0 - e^{-\lambda_p 0}\Big\} \\ &= (\log \lambda_p - \log \lambda_q) + \frac{\lambda_q - \lambda_p}{\lambda_p} \\ &= \log \frac{\lambda_p}{\lambda_q} + \frac{\lambda_q}{\lambda_p} - 1 \end{aligned} \tag{B.8}$$

Table E.9
Scenario 2: The rules for each considered dataset and minsup value with the second-highest KLD between T_X and $D - T_X$ associated with the distributions in Fig. E.11 and E.12.

	Rule (left side)	Mean	SD	IQR	sup.	KLD	minsup
MIMIC-IV							
(a)	966 ∧ 51881	11	8.86	9	1713	0.6123 ± 0.05	2.5%
(b)	966	8	8.44	9	4461	0.477 ± 0.02	5%
(c)	51881	7	7.62	7	4243	0.301 ± 0.01	7.5%
Telco Customer Churn							
(d)	no (online backup) ∧ no (senior citizen) ∧ Month-to-Month ∧ yes (phone service)	11	12.24	14	755	0.423 ± 0.04	5%
Churn for Bank Customers							
(e)	2 (#products) ∧ yes (credit card)	653	96.38	132	3010	0.0006 ± 0.01	5%

Table F.10
Top 10 interesting rules (using minsup 5%) comparison for the Telco Customer Churn dataset between generated rules by approach from Aumann et al. and Scenario 2 of our approach.

Rules (Aumann et al.)	Rules (Scenario 2)
one year ∧ yes (streaming movies) ∧ fiber optic ⇒ 52	no (online backup) ∧ no (senior citizen) ∧ Month-to-Month ⇒ 11
one year ∧ yes (streaming movies) ∧ fiber optic ∧ yes (phone service) ⇒ 52	no (online backup) ∧ no (senior citizen) ∧ Month-to-Month ∧ yes (phone service) ⇒ 11
one year ∧ fiber optic ⇒ 52	yes (streaming TV) ∧ no (multiple lines) ∧ fiber optic ∧ yes (paperless billing) ∧ no (tech support) ∧ no (dependents) ∧ Month-to-Month ⇒ 9
one year ∧ fiber optic ∧ yes (phone service)⇒ 52	yes (streaming TV) ∧ no (multiple lines) ∧ fiber optic ∧ yes (paperless billing) ∧ no (tech support) ∧ no (dependents) ∧ Month-to-Month ∧ yes (phone service) ⇒ 9
one year ∧ yes (streaming movies) ∧ no (dependents) ⇒ 51	yes (streaming TV) ∧ no (multiple lines) ∧ yes (paperless billing) ∧ no (tech support) ∧ no (dependents) ∧ Month-to-Month ⇒ 9
one year ∧ yes (streaming movies) ∧ yes (paperless billing) ⇒ 50	yes (streaming TV) ∧ no (multiple lines)∧ yes (paperless billing) ∧ no (tech support) ∧ no (dependents) ∧ Month-to-Month ∧ yes (phone service) ⇒ 9
one year ∧ yes (streaming movies) ∧ yes (phone service) ⇒ 49	no (partner) ∧ no (online security) ∧ no (dependents) ∧ Month-to-Month ∧ yes (phone service) ⇒ 11
one year ∧ no (dependents) ∧ yes (phone service) ⇒ 49	yes (streaming TV) ∧ no (multiple lines) ∧ yes (paperless billing)∧ Month-to-Month ⇒ 9
one year ∧ yes (streaming movies) ⇒ 49	yes (streaming TV) ∧ no (multiple lines) ∧ yes (paperless billing) ∧ Month-to-Month ∧ yes (phone service) ⇒ 9
one year ∧ yes (streaming TV) ∧ yes (streaming movies) ∧ yes (phone service) ⇒ 49	yes (streaming TV) ∧ no (multiple lines) ∧ no (online backup) ∧ yes (paperless billing) ∧ Month-to-Month ⇒ 9

Appendix C. Transformation of the Kullback–Leibler divergence where *p* is discrete and *q* is an normal density function

We consider the density function of the normal distribution $q(x)$ [35]:

$$q(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2} \tag{C.1}$$

and the distribution of credit score values $p(x)$ inside a frequent itemset.

First, the lower interval limit of the original KLD Eq. (2) can be constrained since the credit score takes at least the value 350:

$$KLD(p \parallel q) := \int_{350}^{\infty} p(x) \cdot \log \frac{p(x)}{q(x)} dx \tag{C.2}$$

Based on equation

$$\int_0^{\infty} f(x)dx = \sum_{i=0}^{\infty} \int_i^{i+1} f(x)dx \tag{C.3}$$

We obtain Eq. (A.4) after rewriting Eq. (2).

Now we have to solve the integral of $\int_i^{i+1} \log q(x)dx$. For this $q(x)$ is substituted:

$$\begin{aligned}\int_i^{i+1} \log q(x)dx &= \int \log \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2} dx \\ &= \int \log \frac{1}{\sigma\sqrt{2\pi}} dx + \int -\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2 dx \\ &= \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2}\left(\frac{(-\mu+i+1)^3 - (i-\mu)^3}{3\sigma^2}\right)\end{aligned} \tag{C.4}$$

Table F.11
Top 10 interesting rules (using minsup 5%) comparison for the Churn for Bank Customers dataset between generated rules by approach from Aumann et al. and Scenario 1 of our approach.

Rules (Aumann et al.)	Rules (Scenario 1)
Germany ∧ no (credit card) ⇒ 659	1 (#products) ∧ France ⇒ 650
no (credit card) ∧ female ∧ 2 (#products) ⇒ 658	yes (credit card) ⇒ 652
no (active member) ∧ France ∧ 2 (#products) ∧ male ∧ yes (credit card) ⇒ 658	2 (#products) ∧ male ∧ yes (credit card) ⇒ 654
no (credit card)∧ 2 (#products) ∧ yes (active member) ⇒ 658	1 (#products) ∧ France ∧ male ⇒ 651
yes (active member) ⇒ 655	1 (#products) ∧ male ⇒ 651
no (credit card) ∧ yes (active member) ∧ male ⇒ 657	no (credit card) ∧ France ⇒ 650
Germany ∧ 2 (#products) ∧ yes (active member) ⇒ 657	female ∧ yes (active member) ⇒ 654
Spain ∧ 2 (#products) ∧ yes (active member) ⇒ 657	France ⇒ 652
2 (#products) ∧ yes (active member) ⇒ 656	1 (#products) ∧ yes (active member) ⇒ 654
female ∧ 2 (#products) ∧ yes (active member) ⇒ 656	yes (active member) ⇒ 655

Table F.12
Top 10 interesting rules (using minsup 5%) comparison for the Churn for Bank Customers dataset between generated rules by approach from Aumann et al. and Scenario 2 of our approach.

Rules (Aumann et al.)	Rules (Scenario 2)
Germany ∧ no (credit card) ⇒ 659	France ⇒ 652
no (credit card) ∧ female ∧ 2 (#products) ⇒ 658	2 (#products) ∧ yes (credit card) ⇒ 653
no (active member) ∧ France ∧ 2 (#products) ∧ male ∧ yes (credit card) ⇒ 658	yes (active member) ⇒ 655
no (credit card) ∧ 2 (#products) ∧ yes (active member) ⇒ 658	no (active member) ⇒ 649
yes (active member) ⇒ 655	no (active member) ∧ yes (credit card) ⇒ 649
no (credit card) ∧ yes (active member) ∧ male ⇒ 657	France ∧ yes (credit card) ⇒ 652
Germany ∧ 2 (#products) ∧ yes (active member) ⇒ 657	yes (active member) ∧ yes (credit card) ⇒ 655
Spain ∧ 2 (#products) ∧ yes (active member) ⇒ 657	2 (#products) ⇒ 653
2 (#products) ∧ yes (active member) ⇒ 656	1 (#products) ∧ yes (credit card) ⇒ 652
female ∧ 2 (#products) ∧ yes (active member) ⇒ 656	France ∧ male ⇒ 652

The substitution of the results in (A.4) provide

$$\begin{aligned} KLD(p \parallel q) &= \sum_{i=350}^{\infty} p(i) \cdot \left(\log p(i) - \int_{i}^{i+1} \log q(x) dx \right) \\ &= \sum_{i=350}^{\infty} p(i) \cdot \left(\log p(i) - \log \frac{1}{\sigma\sqrt{2\pi}} \right. \\ &\quad \left. - \frac{1}{2} \left(\frac{(-\mu + i + 1)^3 - (i - \mu)^3}{3\sigma^2} \right) \right) \end{aligned} \tag{C.5}$$

Appendix D. Scenario 1: Distributions and rules with the second-highest KLD

See Figs. D.9 and D.10 and Table D.8.

Appendix E. Scenario 2: Distributions and rules with the second-highest KLD

See Figs. E.11 and E.12 and Table E.9.

Appendix F. Rule comparison

See Tables F.10–F.18.

Appendix G. Translation of the ICD codes for procedures and diagnoses

See Table G.19.

Table F.13
Top 10 interesting rules (using minsup 2.5%) comparison for the MIMIC-IV dataset between generated rules by approach from Aumann et al. and Scenario 1 of our approach.

Rules (Aumann et al.)	Rules (Scenario 1)
5A1955Z ⇒ 15	5A1955Z ⇒ 15
9672 ∧ 966 ⇒ 13	9672 ∧ 966 ∧ EW EMER. ⇒ 13
9672 ∧ male ⇒ 13	9672 ∧ 966 ⇒ 14
9672 ∧ 966 ∧ EW EMER. ⇒ 13	9672 ∧ EW EMER. ⇒ 13
9672 ⇒ 13	9672 ⇒ 13
9672 ∧ 51881 ⇒ 13	9672 ∧ 51881 ⇒ 13
9672 ∧ 9604 ⇒ 13	9672 ∧ male ⇒ 13
9672 ∧ EW EMER. ⇒ 13	9672 ∧ 9604 ⇒ 13
3E0G76Z ∧ 02HV33Z ⇒ 12	51881 ∧ 966 ⇒ 11
51881 ∧ 966 ⇒ 11	9604 ∧ 966 ⇒ 11

Table F.14
Top 10 interesting rules (using minsup 2.5%) comparison for the MIMIC-IV dataset between generated rules by approach from Aumann et al. and Scenario 2 of our approach.

Rules (Aumann et al.)	Rules (Scenario 2)
5A1955Z ⇒ 15	3E0G76Z ∧ 02HV33Z ⇒ 12
9672 ∧ 966 ⇒ 13	51881 ∧ 966 ⇒ 11
9672 ∧ male ⇒ 13	9604 ∧ 966 ⇒ 11
9672 ∧ 966 ∧ EW EMER. ⇒ 13	3324 ⇒ 10
9672 ⇒ 13	966 ∧ 3893 ⇒ 10
9672 ∧ 51881 ⇒ 13	3E0G76Z ⇒ 10
9672 ∧ 9604 ⇒ 13	3E0G76Z ∧ Other ⇒ 10
9672 ∧ EW EMER. ⇒ 13	3E0G76Z ∧ male ⇒ 10
3E0G76Z ∧ 02HV33Z ⇒ 12	966 ⇒ 9
51881 ∧ 966 ⇒ 11	966 ∧ 42731 ⇒ 9

Table F.15
Top 10 interesting rules (using minsup 5%) comparison for the MIMIC-IV dataset between generated rules by approach from Aumann et al. and Scenario 1 of our approach.

Rules (Aumann et al.)	Rules (Scenario 1)
3E0G76Z ∧ 02HV33Z ⇒ 12	9672 ⇒ 13
51881 ∧ 966 ⇒ 11	3E0G76Z ⇒ 10
9604 ∧ 966 ⇒ 11	966 ⇒ 9
3324 ⇒ 10	966 ∧ EWEMER. ⇒ 8
966 ∧ 3893 ⇒ 10	966 ∧ male ⇒ 9
3E0G76Z ⇒ 10	Z781 ⇒ 8
3E0G76Z ∧ Other ⇒ 10	51881 ⇒ 7
3E0G76Z ∧ male ⇒ 10	51881 ∧ EWEMER. ⇒ 7
966 ⇒ 9	9604 ⇒ 7
966 ∧ 42731 ⇒ 9	9604 ∧ EWEMER. ⇒ 6

Table F.16
Top 10 interesting rules (using minsup 5%) comparison for the MIMIC-IV dataset between generated rules by approach from Aumann et al. and Scenario 2 of our approach.

Rules (Aumann et al.)	Rules (Scenario 2)
3E0G76Z ∧ 02HV33Z ⇒ 12	3E0G76Z ⇒ 10
51881 ∧ 966 ⇒ 11	966 ⇒ 9
9604 ∧ 966 ⇒ 11	966 ∧ male ⇒ 9
3324 ⇒ 10	966 ∧ EWEMER. ⇒ 8
966 ∧ 3893 ⇒ 10	Z781 ⇒ 8
3E0G76Z ⇒ 10	51881 ⇒ 7
3E0G76Z ∧ Other ⇒ 10	9604 ⇒ 7
3E0G76Z ∧ male ⇒ 10	51881 ∧ EWEMER. ⇒ 7
966 ⇒ 9	J9601 ⇒ 7
966 ∧ 42731 ⇒ 9	9604 ∧ EWEMER. ⇒ 6

Table F.17
Top 10 interesting rules (using minsup 7.5%) comparison for the MIMIC-IV dataset between generated rules by approach from Aumann et al. and Scenario 1 of our approach.

Rules (Aumann et al.)	Rules (Scenario 1)
966 ⇒ 9	966 ⇒ 9
51881 ⇒ 7	51881 ⇒ 7
9604 ⇒ 7	9604 ⇒ 7
02HV33Z ⇒ 6	02HV33Z ⇒ 6
3893 ⇒ 5	3961 ⇒ 3
3897 ⇒ 5	3897 ⇒ 5
3893 ∧ EWEMER. ⇒ 5	3893 ⇒ 5
5990 ⇒ 4	9671 ⇒ 4
N179 ⇒ 4	SURGICALSAMEDAYADMISSION ∧ male ⇒ 3
D62 ⇒ 4	SURGICALSAMEDAYADMISSION ∧ Other ⇒ 3

Table F.18
Top 10 interesting rules (using minsup 7.5%) comparison for the MIMIC-IV dataset between generated rules by approach from Aumann et al. and Scenario 2 of our approach.

Rules (Aumann et al.)	Rules (Scenario 1)
966 ⇒ 9	966 ⇒ 9
51881 ⇒ 7	51881 ⇒ 7
9604 ⇒ 7	9604 ⇒ 7
02HV33Z ⇒ 6	02HV33Z ⇒ 6
3893 ⇒ 5	3893 ⇒ 5
3897 ⇒ 5	3897 ⇒ 5
3893 ∧ EWEMER. ⇒ 5	3893 ∧ EWEMER. ⇒ 5
5990 ⇒ 4	5990 ⇒ 4
N179 ⇒ 4	N179 ⇒ 4
D62 ⇒ 4	53081 ⇒ 3

Table G.19
All ICD codes for procedures involved in a rule from Tables. 5, 6, D.8, E.9, F.13, F.14, F.15, F.16, F.17 and F.18.

ICD code	Description
9672	Continuous invasive mechanical ventilation for 96 consecutive hours or more
966	Enteral infusion of concentrated nutritional substances
3E0G76Z	Introduction of Nutritional Substance into Upper GI, Via Natural or Artificial Opening
5A1955Z	Respiratory Ventilation, Greater than 96 Consecutive Hours
51881	Acute respiratory failure
02HV33Z	Insertion of Infusion Device into Superior Vena Cava, Percutaneous Approach
9604	Insertion of endotracheal tube
3324	Closed [endoscopic] biopsy of bronchus
3893	Venous catheterization, not elsewhere classified
42731	Atrial fibrillation
Z781	Physical restraint status
J9601	Acute respiratory failure with hypoxia
3897	Central venous catheter placement with guidance
3961	Extracorporeal circulation auxiliary to open heart surgery
9671	Continuous invasive mechanical ventilation for less than 96 consecutive hours
5990	Urinary tract infection, site not specified
N179	Acute kidney failure, unspecified
D62	Acute posthemorrhagic anemia
53081	Esophageal reflux

References

[1] P. Adriaans, D. Zantinge, Data Mining, Addison-Wesley Longman Publishing Co., Inc., United States, 1997.

[2] D. Adhikary, S. Roy, Trends in quantitative association rule mining techniques, in: 2015 IEEE 2nd International Conference on Recent Trends in Information Systems, ReTIS, IEEE, 2015, pp. 126–131, http://dx.doi.org/10.1109/ReTIS.2015.7232865.

[3] R. Agrawal, R. Srikant, et al., Fast algorithms for mining association rules, in: Proc. 20th Int. Conf. Very Large Data Bases, Vol. 1215, VLDB, Citeseer, 1994, pp. 487–499.

[4] Y. Aumann, Y. Lindell, A statistical theory for quantitative association rules, J. Intell. Inf. Syst. 20 (3) (2003) 255–283, http://dx.doi.org/10.1145/312129.312243.

[5] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, ACM Sigmod Rec. 29 (2) (2000) 1–12.

[6] M. Al-Maolegi, B. Arkok, An improved apriori algorithm for association rules, 2014, http://dx.doi.org/10.5121/ijnlc.2014.3103, arXiv preprint arXiv:1403.3948.

[7] S. Raj, D. Ramesh, M. Sreenu, K.K. Sethi, EAFIM: Efficient apriori-based frequent itemset mining algorithm on spark for big transactional data, Knowl. Inf. Syst. 62 (9) (2020) 3565–3583, http://dx.doi.org/10.1007/s10115-020-01464-1.

[8] J.M. Luna, P. Fournier-Viger, S. Ventura, Frequent itemset mining: A 25 years review, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9 (6) (2019) e1329, http://dx.doi.org/10.1002/widm.1329.

[9] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9 (4) (2019) e1312, http://dx.doi.org/10.1002/widm.1312.

[10] F.K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: A survey, in: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO, IEEE, 2018, pp. 0210–0215, http://dx.doi.org/10.23919/MIPRO.2018.8400040.

[11] N. Burkart, M.F. Huber, A survey on the explainability of supervised machine learning, J. Artificial Intelligence Res. 70 (2021) 245–317, http://dx.doi.org/10.1613/jair.1.12228.

[12] H. Lakkaraju, S.H. Bach, J. Leskovec, Interpretable decision sets: A joint framework for description and prediction, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1675–1684, http://dx.doi.org/10.1145/2939672.2939874.

[13] J. Hipp, U. Güntzer, G. Nakhaeizadeh, Algorithms for association rule mining—a general survey and comparison, ACM Sigkdd Explor. Newsl. 2 (1) (2000) 58–64, http://dx.doi.org/10.1145/360402.360421.

[14] R.J. Miller, Y. Yang, Association rules over interval data, ACM SIGMOD Rec. 26 (2) (1997) 452–461, http://dx.doi.org/10.1145/253262.253361.

[15] R. Srikant, R. Agrawal, Mining quantitative association rules in large relational tables, in: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, 1996, pp. 1–12, http://dx.doi.org/10.1145/233269.233311.

[16] J.R. Hershey, P.A. Olsen, Approximating the Kullback Leibler divergence between Gaussian mixture models, in: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 4, ICASSP'07, IEEE, 2007, pp. IV–317, http://dx.doi.org/10.1109/ICASSP.2007.366913.

[17] C. Röver, T. Friede, Discrete approximation of a mixture distribution via restricted divergence, J. Comput. Graph. Statist. 26 (1) (2017) 217–222, http://dx.doi.org/10.1080/10618600.2016.1276840.

[18] A. Awad, M. Bader-El-Den, J. McNicholas, Patient length of stay and mortality prediction: a survey, Health Serv. Manag. Res. 30 (2) (2017) 105–120, http://dx.doi.org/10.1177/0951484817696212.

[19] A. Suresh, K. Harish, N. Radhika, Particle swarm optimization over back propagation neural network for length of stay prediction, Procedia Comput. Sci. 46 (2015) 268–275, http://dx.doi.org/10.1016/j.procs.2015.02.020.

[20] Telco Customer Churn, URL https://www.kaggle.com/datasets/blastchar/telco-customer-churn. (Last accessed: 2023-06-04).

[21] Churn for Bank Customers, URL https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers. (Last accessed: 2023-06-04).

[22] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L.A. Celi, R. Mark, MIMIC-IV, PhysioNet, 2022, http://dx.doi.org/10.13026/7vcr-e114, URL https://physionet.org/content/mimiciv/2.0/.

[23] T. Van Erven, P. Harremos, Rényi divergence and Kullback-Leibler divergence, IEEE Trans. Inform. Theory 60 (7) (2014) 3797–3820.

[24] F. Pérez-Cruz, Kullback-Leibler divergence estimation of continuous distributions, in: 2008 IEEE International Symposium on Information Theory, IEEE, 2008, pp. 1666–1670.

[25] M.N. Do, Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models, IEEE Signal Process. Lett. 10 (4) (2003) 115–118.

[26] M. Zaharia, M. Chowdhury, M.J. Franklin, S. Shenker, I. Stoica, et al., Spark: Cluster computing with working sets, HotCloud 10 (10–10) (2010) 95.

[27] Apache spark™ - Unified analytics engine for big data, 2022, https://spark.apache.org/. (Last accessed: 2022-12-04).

[28] S. Mohammed, sedirmohammed/Implementation-A-Statistical-Method-for-Predicting-Quantitative-Variables-in-Association-Rule-Mining, 2023-06-04, URL https://github.com/sedirmohammed/Implementation-A-Statistical-Method-for-Predicting-Quantitative-Variables-in-Association-Rule-Mining. (Last accessed: 2023-06-04).

[29] L. Turgeman, J.H. May, R. Sciulli, Insights from a machine learning model for predicting the hospital length of stay (LOS) at the time of admission, Expert Syst. Appl. 78 (2017) 376–385, http://dx.doi.org/10.1016/j.eswa.2017.02.023.

[30] I.W.M. Verburg, A. Atashi, S. Eslami, R. Holman, A. Abu-Hanna, E. de Jonge, N. Peek, N.F. de Keizer, Which models can I use to predict adult ICU length of stay? A systematic review, Crit. Care Med. 45 (2) (2017) e222–e231, http://dx.doi.org/10.1097/CCM.0000000000002054.

[31] X. Ma, Y. Si, Z. Wang, Y. Wang, Length of stay prediction for ICU patients using individualized single classification algorithm, Comput. Methods Programs Biomed. 186 (2020) http://dx.doi.org/10.1016/j.cmpb.2019.105224.

[32] J.V. Tu, M.R. Guerriere, Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery, Comput. Biomed. Res. 26 (3) (1993) 220–229, http://dx.doi.org/10.1006/cbmr.1993.1015.

[33] C. Gholipour, F. Rahim, A. Fakhree, B. Ziapour, Using an artificial neural networks (ANNs) model for prediction of intensive care unit (ICU) outcome and length of stay at hospital in traumatic patients, J. Clin. Diagn. Res. JCDR 9 (4) (2015) OC19, http://dx.doi.org/10.7860/JCDR/2015/9467.5828.

[34] E. Behrends, Elementare Stochastik: Ein Lernbuch-Von Studierenden Mitentwickelt, Springer-Verlag, 2012.

[35] D.I. Belov, R.D. Armstrong, Distributions of the Kullback–Leibler divergence with applications, Br. J. Math. Stat. Psychol. 64 (2) (2011) 291–309.

[36] M. Shahin, S. Arakkal Peious, R. Sharma, M. Kaushik, S. Ben Yahia, S.A. Shah, D. Draheim, Big data analytics in association rule mining: a systematic literature review, in: 2021 the 3rd International Conference on Big Data Engineering and Technology, BDET, 2021, pp. 40–49.

[37] A.F. Siegel, Counterexamples in Probability and Statistics, Routledge, 2017.

[38] V. Dragalin, V. Fedorov, S. Patterson, B. Jones, Kullback–Leibler divergence for evaluating bioequivalence, Stat. Med. 22 (6) (2003) 913–930.

[39] S. Kotsiantis, D. Kanellopoulos, Association rules mining: A recent overview, GESTS Int. Trans. Comput. Sci Eng. 32 (1) (2006) 71–82.

Curriculum Vitae

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

List of publications

1. S. Mohammed, K. Rubarth, S. K. Piper, F. Schiefenhövel, J.-C. Freytag, F. Balzer, and S. Boie, ‘A statistical method for predicting quantitative variables in association rule mining’, *Inf. Syst.*, vol. 118, p. 102253, Sep. 2023, doi: 10.1016/j.is.2023.102253.

Acknowledgments

Many people, including family, friends, and colleagues, have accompanied and supported me on my journey to completing this doctorate. The list would be too long to mention everyone by name and express my thanks individually. Therefore, I want to thank all those who have supported me directly or indirectly in completing this doctorate.

First, I am deeply grateful to my supervisors. I owe special thanks to my first supervisor, Prof. Dr. med. Dr. rer. nat. Felix Balzer. Under his guidance, I had an exciting time at Charité and was always supported by him with technical and organizational questions.

My sincere appreciation goes to my second supervisor, Prof. Johann-Christoph Freytag, PhD. He gave me my first insights into scientific work during my studies. He was also the one who motivated me to pursue a doctorate. He is now not only a professional supervisor but also a mentor for me.

I would also like to thank my third supervisor, Sebastian Boie, PhD. Sebastian has supported me during this doctorate with many discussions and valuable tips. His critical view of problems and regular exchanges have ensured that this doctorate has progressed steadily.

I am also grateful to my close friend Jörg P. Schäfer. He always had an open ear for me and gave me a lot of moral support. I remember many sessions where we discussed our work and ended the evening with good food and video games.

I would also like to thank my long-time friend, Dennis. He has always been there for me. I am glad that I can still count him among my closest friends after such a long time.

Last, I would like to express my special thanks to Maike. For years, she has supported me in all my undertakings, ensuring I can focus on my work while also reminding me to take necessary breaks.