DOKTORI (PhD) ÉRTEKEZÉS

Csalódi Róbert

Pannon Egyetem
2024

Gépi tanulással támogatott túléléselemzés

Az értekezés doktori (PhD) fokozat elnyerése érdekében készült a Pannon Egyetem Vegyészmérnöki- és Anyagtudományok Doktori Iskolája keretében

Anyagtudományok és technológiák tudományágban

Írta: Csalódi Róbert

DOI:10.18136/PE.2024.893

Témavezetők: Dr. Abonyi János, Dr. Ruppert Tamás

Elfogadásra javaslom (igen / nem)

……. …….
(témavezetők)

Az értekezést bírálóként elfogadásra javaslom:

Bíráló neve: Dr. Kovács Edith Alice igen /nem

…….
(bíráló)

Bíráló neve: Dr. Molontay Roland igen /nem

…….
(bíráló)

A jelölt az értekezés nyilvános vitáján ….....%-ot ért el.

Veszprém, …….
(a Bíráló Bizottság elnöke)

A doktori (PhD) oklevél minősítése….....

Veszprém, …….
(az EDHT elnöke)

Pannon Egyetem

Doktori (PhD) Értekezés

Gépi tanulással támogatott túléléselemzés

Szerző:
Csalódi Róbert

Konzulensek:
Prof. Dr. habil. Abonyi János
Dr. Ruppert Tamás

Értekezés doktori (PhD) fokozat elnyerése érdekében
a ***Pannon Egyetem***

Vegyészmérnöki- és Anyagtudományok
Doktori Iskolájához tartozóan

Folyamatmérnöki Intézeti Tanszék

Pannon Egyetem
2024

University of Pannonia

Doctoral (PhD) Thesis

Machine learning supported survival analysis

Author:
Róbert Csalódi

Supervisors:
Prof. Dr. habil. János Abonyi
Dr. Tamás Ruppert

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Doctoral School in Chemical Engineering and Material Sciences
of University of Pannonia

Department of Process Engineering

University of Pannonia

2024

Az, akinek elég bátorsága és türelme van ahhoz,
hogy egész életében a sötétségbe nézzen,
elsőként fogja meglátni benne a fény felvillanását.

Dmitry Glukhovsky

He who has enough courage and patience
to stare into the darkness for his entire life,
shall be the first to see the flash of light.

Dmitry Glukhovsky

PANNON EGYETEM

Kivonat

Mérnöki Kar
Folyamatmérnöki Intézeti Tanszék

Philosophiæ Doctor

Gépi tanulással támogatott túléléselemzés

írta: Csalódi Róbert

A túléléselemés egy statisztikai módszertan, amit egy bizonyos esemény bekövetkezéséig eltelt idő eloszlásának becslésére használnak. A módszer hatékonyan tárja fel a folyamatok működése során fellépő veszteségeket. Ugyanakkor önmagában gyenge eszköznek bizonyul, amikor a cél feltárni a folyamatok veszteségének mélyebb gyökérokait. Ennek elkerülésére a túléléselemzés gépi tanulási módszerekkel történő integrált alkalmazása javasolt. Ez a disszertáció három integrált algoritmust mutat be, melyek kombinálják a túléléselemzést és a gépi tanulást, ezzel lehetővé téve komplex folyamatok értelmezését. Az első algoritmus egy integrált túléléselemzés és expectation-maximization alapú klaszterezési keretrendszert mutat be, ami a túlélési idők és a magyarázó változók között fellépő hasonlóság alapján klaszterez. A második algoritmus egy integrált túléléselemzés és gyakori elemhalmaz alapú asszociációs szabályokat bányászó módszer, ami az időfüggő kategorikus változókból definiált releváns kiváltó eseményeket identifikál, amelyek a versengő kockázatok következményi eseményeihez vezetnek. A módszer közvetlenül a szabályok támogatottsága és konfidenciái alapján becsli a kummulatív előfordulási függvényt. A harmadik algoritmus egy integrált túléléselemzés és szekvenciális mintázatok bányászatát végző keretrendszer, ami a eseményátmenetek időfüggő konfidenciafüggvényét határozza meg. A disszertáció nem csak az algoritmusok alkalmazhatóságát mutatja be, hanem rávilágít bennük rejlő lehetőségekre, amelyek értékes betekintést nyújtanak a folyamatok működéseibe, valamint javítják az előrejelzések a döntéshozatalok minőségét. Az algoritmusok hatékonyságának bemutatása különböző műszaki, lemorzsolódási és orvosi esettanulmányokon végzett elemzések keretében valósul meg.

UNIVERSITY OF PANNONIA

Abstract

Faculty of Engineering
Department of Process Engineering

Doctor of Philosophy

Machine learning supported survival analysis

by Róbert Csalódi

Survival analysis is a statistical methodology used to estimate the probability distribution of time until an event occurs. The method serves as an efficient tool in exploring inefficient operation of processes. However, relying solely on this method can be limiting when aiming to identify the root causes of losses. Therefore, integrating survival analysis with machine learning algorithms is crucial. This thesis introduces three integrated algorithms that combine survival analysis with machine learning models, offering a more profound understanding of complex processes. The first algorithm presents an integrated survival analysis and expectation-maximization-based clustering framework, identifying clusters based on the similarity of survival times and explanatory variables. The second algorithm introduces an integrated survival analysis and frequent itemset-based association rule mining method, that identifies relevant triggering events defined from time-dependent categorical variables, that lead to consequent events of competing risks. The algorithm estimates the cumulative incidence function directly based on the rule supports and confidences. The third algorithm demonstrates an integrated survival analysis and sequential pattern mining framework, determining the time-dependent confidence function of event continuations. Through these algorithms, the thesis not only showcases their adaptability but also highlights their potential to provide valuable insights, improve predictions, and enhance decision-making processes. The effectiveness of these algorithms is demonstrated through diverse case studies across technical, churn, and medical domains, showcasing their broad applicability.

PANNONISCHE UNIVERSITÄT

Auszug

Fakultät für Ingenieurwissenschaften
Abteilung für Verfahrenstechnik

Doktor der Philosophie

Durch maschinelles Lernen unterstützte Überlebensanalyse

von Róbert Csalódi

Die Überlebensanalyse ist eine statistische Methode zur Schätzung der Wahrscheinlichkeitsverteilung der Zeit bis zum Eintreten eines Ereignisses. Die Methode dient als effizientes Instrument zur Untersuchung ineffizienter Prozesse. Die alleinige Verwendung dieser Methode kann jedoch bei der Ermittlung der Ursachen von Verlusten einschränkend sein. Daher ist die Integration der Überlebensanalyse mit Algorithmen des maschinellen Lernens von entscheidender Bedeutung. In dieser Dissertation werden drei integrierte Algorithmen vorgestellt, die die Überlebensanalyse mit Modellen des maschinellen Lernens kombinieren und so ein tieferes Verständnis komplexer Prozesse ermöglichen. Der erste Algorithmus ist eine integrierte Überlebensanalyse und auf Erwartungs-Maximierung basierendes Clustering, der den Cluster anhand der Ähnlichkeit von Überlebenszeiten und erklärenden Variablen identifiziert. Der zweite Algorithmus führt eine integrierte Überlebensanalyse und auf häufigen Itemmengen basierende Assoziationsregel-Mining-Methode ein, die relevante auslösende Ereignisse identifiziert, die aus zeitabhängigen kategorialen Variablen definiert werden und zu konkurrierenden Risiken führen. Der Algorithmus schätzt die kumulative Inzidenzfunktion direkt auf der Grundlage der relativen Häufigkeit und Konfidenzwerte. Der dritte Algorithmus demonstriert einen integrierten Rahmen für Überlebensanalyse und sequentielles Muster Mining, der die zeitabhängige Konfidenzfunktion von Ereignisfortsetzungen bestimmt. Die Effektivität dieser Algorithmen wird anhand verschiedener Fallstudien aus den Bereichen Technik, Studienabbruch und Medizin demonstriert, um ihre breite Anwendbarkeit zu zeigen.

Acknowledgements

I am sincerely grateful to my supervisors, Prof. Dr. habil. Janos Abonyi, for selflessly providing guidance and support when I faced challenges and could not find the right direction. Dr. Tamás Ruppert, your unwavering support and encouragement throughout the years have been invaluable to me.

I would like to say thank you to Dr. Zsolt Bagyura. Your dedication and collaboration in the medical parts have played a pivotal role in the success of this research.

To my family and friends, I deeply appreciate all the support you have given me. The beautiful moments we have shared together provided the courage and patience I needed to keep moving forward. Your belief in me has been a driving force, and I am profoundly thankful for your love and encouragement.

Ágnes Horváthné Tímár, the entire journey began with you. Thank you for laying excellent foundations and inspiring me to put more and more effort into my practice. Your emphasis on the importance of practice has been a guiding principle throughout the years.

Dedicated to my Family and Friends

Contents

Chapter 1

Introduction to the thesis

1.1 Introduction to the concept of integrated survival analysis and machine learning algorithms and motivations

Survival analysis is a statistical methodology and serves as a pivotal tool for analyzing the time until a specific event occurs. The method was initially applied in biostatistics and medical researches, where the event of interest is typically the time until death [1]. However, this method has spread across various disciplines, as analogies can be drawn between survival times and a multitude of other quantities [2]. For example, investigation into pain endurance have utilized survival analysis, revealing the duration individuals can tolerate discomfort [3]. Similarly, in the field of education, researchers have approached student dropout with discrete time series, seeking to unravel the underlying causes [4]. In the field of degradation analysis, the focus shifts to calculating the remaining useful life of technical systems [5]. Additionally, survival analysis has proven valuable in understanding the dropping habits of individuals seeking assistance in hotline call centers [6].

Survival analysis identifies probability distributions and the related probability functions such as probability density, cumulative distribution, hazard and survival functions. These models can be parametric, semi parametric or nonparametric [7]. In the context of parametric models, typical probability distributions are considered such as normal, exponential, or Weibull. These models serve as templates

based on their distribution types and the associated parameters, exhibiting characteristic features that manifest in the shape properties of the model descriptive functions. This constrained framework imposes certain conditions that the data must satisfy in order to be fitted [8]. However, the advantage lies in the fact that a continuous model can be identified through parameter fitting using the maximum likelihood estimation method [9], allowing for the direct determination of model descriptive functions. On the other hand, nonparametric models determine the survival model empirically, enabling the real representation of the dataset without the need to pre-select the model type. However, the drawback is that without parameters, the model cannot be globally determined as a set of continuous functions. The different types of functions are in this case discrete and empirical, so their analytical integration and derivation becomes noisy. Therefore, individual methods have been developed to determine these functions. Semi-parametric models, a hybrid of parametric and nonparametric approaches, exploit the strengths of both methodologies [10]. They incorporate an empirical baseline component for estimating the fundamental characteristic, while concurrently integrating a parametric component that modifies the function concerning explanatory variables.

Survival analysis has an additional feature called censoring, which proves useful in situations where the complete information about the time to an event of interest is not available for all subjects. This may occur due to various reasons, such as the termination of the study before the events have happened, subjects being lost to follow-up, or the event not occurring within the observation period [11]. In medical studies, censoring plays a pivotal role when evaluating the time until a specific outcome, such as the recurrence of a disease or the occurrence of adverse effects. For example, in cancer studies, patients may be censored if they are still alive at the end of the study or if they withdraw from the study before the event of interest occurs [12]. Censoring is considered independent, meaning when a subject is eliminated, it is assumed that his further survival characteristic remain the same as during the examination [13]. In essence, censoring allows for the incorporation of incomplete yet valuable data, making survival analysis applicable in scenarios where complete information might be inaccessible.

Machine learning algorithms are computational techniques that enable systems to learn and make predictions or decisions without explicit programming. These algorithms fall into two broad categories: supervised and unsupervised learning [14]. In supervised learning, algorithms are trained on labeled datasets, where input

features are paired with their corresponding output labels. On the other hand, unsupervised learning algorithms deal with unlabeled data and aim to discover inherent patterns or structures. Clustering [15] and frequent pattern mining [16] are typical tasks in unsupervised learning. The choice of machine learning algorithm depends on the research question, and each problem may require a tailored approach. Machine learning algorithms exhibit diverse properties, each with its own set of advantages and disadvantages. Interpretability is the ease of understanding and explaining how a model reaches a decision [17]. Linear regression and decision tree, for instance, offer high interpretability, allowing researchers to comprehend the inner connections of the model. However, this interpretability may come at the cost of accuracy in capturing intricate relationships within the data. On the other hand, more complex models like neural networks often excel in accuracy but may lack interpretability due to their complex structures [18].

Survival analysis serves as an efficient tool in exploring the losses of processes [19]. However, for a more comprehensive analysis that explores the root causes of inefficient operations, relying solely on this method proves to be a limited approach. Therefore, enhancing survival analysis with machine learning algorithms is essential to explore the underlying factors in the necessary depth [20]. Recent applications have emphasized the integration of survival analysis with various machine learning techniques, showcasing the versatility of this combined approach. Ensemble methods have been employed to improve predictive accuracy [21], clustering algorithms have been utilized to uncover latent patterns within the data [22], and even neural networks have been leveraged for their capacity to capture complex relationships [23]. The essence of identifying root causes lies in the ability of exploring the factors that influence the model. Therefore, the creation of interpretable models emerges as crucial in revealing the sources of inefficient operations.

This thesis aims to present approaches that integrate survival analysis with machine learning models, contributing to a more comprehensive understanding of processes and their underlying dynamics. Within this framework, several machine learning methods are considered, including Gaussian mixture models, frequent itemset mining, association rule mining, frequent sequence mining and sequential rule mining. The effectiveness of the proposed algorithms is demonstrated through diverse case studies, showcasing their applicability across various domains. The methods are applied to estimate:

1. Remaining useful life of Li-ion batteries based on their capacity, internal resistance and charging conditions
2. Survival times of patients with prostate cancer based on their age, serum hemoglobin level and treatment
3. Mortality rate per 100K population of countries related to COVID-19 pandemic based on demographical and economical data.
4. Dropout rate of university students based on uncompleted subjects patterns
5. Occurrence chances of disorders based on their already existing ones

Through a diverse range of case studies, this thesis aims to showcase not only the adaptability of the proposed algorithms, but also their potential to offer valuable insights and enhance prediction and decision-making processes. These studies span across complex and dynamic scenarios in various domains, emphasizing the ability of algorithms to contribute to improved outcomes. The next section introduces the main research questions of this thesis.

1.2 Research questions and thesis outline

The thesis provides a collection of integrated survival analysis and machine learning algorithms that can be applied for general purposes. The main research questions are the followings:

- *How can local models be identified within a heterogeneous population based on explanatory variables?*

  The main idea is that the explanatory variables influence the survival characteristics of the population. Consequently, there is a need for the development of a clustering algorithm that creates local models by assessing the similarity of survival times and explanatory variables. In Chapter 3, I present an integrated survival analysis and expectation-maximization-based clustering algorithm. This algorithm not only aids in identifying local models but also represents cluster membership using Takagi-Sugeno fuzzy rules. This representation provides a comprehensive framework for determining the operating domain of continuous variables.

- *How can competing risks be modeled alternatively by considering time - dependent categorical variables?*

  The time-dependent categorical variables are conceptualized as triggering events, acting as precursors, initiating consequent events that signify the competing outcomes of the survival process. In Chapter 4, I present an integrated survival analysis and frequent itemset-based association rule mining algorithm that identifies relevant triggering events, that lead to a competing risk. Moreover, the algorithm estimates the cumulative incidence function directly based on the rule supports and confidences.

- *How can the confidence of sequential patterns interpreted as time-dependent function?*

  The sequential pattern mining reveals relevant continuations, where the probability of the transition (confidence) remains constant. The key idea is that the characteristics of the confidence can be described by the distribution of the elapsed times between two events. In Chapter 5, I present a sequential pattern mining-based survival analysis where the confidences of sequences are represented by a time-dependent function. Moreover, its confidence interval can also be determined using the bootstrapping method.

The following section introduces the thesis findings derived from addressing these research questions.

1.3 Thesis findings

This section summarizes the contributions to the development of integrated survival analysis and machine learning algorithms. The new scientific results are the followings:

1. **I developed an integrated survival analysis and expectation - maximization - based clustering framework.**

   (a) I demonstrated, how heterogeneous survival models can be grouped into homogeneous models based on the similarity of survival times and explanatory variables by utilizing an expectation maximization algorithm.

This approach defines clusters and simultaneously identifies their survival probabilities using the Weibull distribution and the related continuous explanatory variables, leveraging multivariate Gaussian distributions.

(b) I presented, how the cluster memberships can be represented using Takagi-Sugeno fuzzy rules, providing a framework to determine the operating domain of continuous variables. With this approach, survival characteristics can be described by considering the domain of continuous variables. The method is versatile and can be applied for categorizing continuous variables.

Related publications: Róbert Csalódi, Zoltán Birkner, János Abonyi: Learning Interpretable Mixture of Weibull Distributions – Exploratory Analysis of How Economic Development Influences the Incidence of COVID-19 Deaths, Data, 2021. [24]
Róbert Csalódi, Zsolt Bagyura, János Abonyi: Mixture of survival analysis models - Cluster-weighted Weibull distributions, IEEE Access, 2021. [25]

2. **I developed an integrated survival analysis and frequent itemset-based association rule mining algorithm.**

(a) I demonstrated, how the probability of competing risks can be determined at specific time instances using event-driven frequent itemset-based association rules. This approach identifies relevant triggering events defined from time-dependent categorical variables, that lead to consequent events defined from competing risks. A sequence of frequent itemsets can be represented as a global, time-independent feature.

(b) I presented, how the cumulative incidence function of a specific competing risk can be directly determined for the population with a given sequence of frequent itemsets. The method segments the dataset for subjects that supports all the frequent itemset of the selected sequence and estimates the cumulative incidence function based on the modified rule supports and confidences.

(c) I introduced, how the student dropout rate can be estimated based on patterns of uncompleted subjects. The study has a sample curriculum that prescribes the recommended semester for each subject completion.

Inability to meet this requirement marks an uncompleted subject event, a crucial factor associated with subsequent student dropout.

Related publication: Róbert Csalódi, János Abonyi: Integrated Survival Analysis and Frequent Pattern Mining for Course Failure-Based Prediction of Student Dropout, Mathematics, 2021. [26]

3. **I developed an integrated survival analysis and sequential pattern mining algorithm.**

   (a) I demonstrated, how the time-dependent support and confidence functions of event transitions can be estimated using the integrated survival analysis and frequent sequence mining algorithm. The approach identifies relevant event continuations through frequent sequence mining and determines their support and confidence metrics. The temporal characteristics of the resultant rule confidences are determined using the Kaplan-Meier estimator. The multiplication of time distributions and rule confidences provides the time-dependent confidence function.

   (b) I presented, how sequential rule mining can be an alternative approach when handling a substantial volume of unique events that are poorly distributed. Unlike providing a continuation of events, this method identifies sets of antecedent events that may occur in any order before another set of consequent events that also may occur in any order. The determination of temporal characteristics of the resultant rules can be made using the previous approach.

   (c) I introduced, how the confidence intervals of the time-dependent confidences can be determined using the bootstrapping method. This involves randomly selecting input sequences and executing the method on this data. The process is executed repeatedly, resulting in a set of confidence functions from bootstraps. The percentile-based method estimates the confidence bounds in this set of functions, thereby establishing the confidence intervals

   Related publications: Róbert Csalódi, Zsolt Bagyura, János Abonyi: Time-dependent sequential association rule-based survival analysis: A healthcare application, MethodsX, 2024. [27]

Róbert Csalódi, Zsolt Bagyura, Ágnes Vathy-Fogarassy, János Abonyi: Time-dependent frequent sequence mining-based survival analysis, Knowledge-Based Systems, 2024. [28]

The proposed algorithms are described in separate chapters, each following a structured format. Every chapter commences with a comprehensive introduction that establishes the context for the algorithm being discussed. This section provides a clear overview of the problem the algorithm aims to address. It also critically assesses the limitations of recent approaches, emphasizing the need for novel methodologies to overcome existing challenges. Following the introduction, the chapter delves into a detailed mathematical description of the proposed algorithm, outlining its underlying principles, methodologies, and formulations. The goal is to provide a clear and rigorous understanding of the structure and functioning. The final section of each chapter showcases the practical application through several case studies. These case studies serve as real-world examples, demonstrating the efficiency of algorithms in solving specific problems. Moreover, this phase highlights their performance and ability to generate meaningful insights. The applied notations are summarized in Section list of notations.

Chapter 2

Formalization of survival analysis

The identified survival model manifests as a random variable, expressible through various functions such as hazard, probability density, probability distribution, or survival function [29]. These functions are a mathematically interconnected framework, that allow for the representation of one in terms of the other [30]. This interconnection is illustrated with the example of exponentially distributed samples and the recommended identification methods in Figure 2.1. The cumulative hazard function can be obtained by integrating the hazard function and can be expressed by the following function:

$$\lambda(t) = \int_0^t h(t)ds \tag{2.1}$$

where $\lambda(t)$ denotes the cumulative hazard function and $h(t)$ stands for the hazard function. Survival function is derived by taking the exponential of the negative cumulative hazard:

$$S(t) = \exp\big(-\lambda(t)\big) \tag{2.2}$$

where $S(t)$ denotes the survival function. The cumulative distribution function is obtained by subtracting the survival function from one expressed as:

$$F(t) = 1 - S(t) \tag{2.3}$$

where $F(t)$ stands for the cumulative distribution function. The probability density function is derived by taking the derivative of the cumulative distribution function and can be described mathematically as follows:

$$f(t) = \frac{dF(t)}{dt} \tag{2.4}$$

where $f(t)$ denotes the probability density function. The hazard function is calculated by dividing the probability density function by the survival function and can be expressed by the next equation:

$$h(t) = \frac{f(t)}{S(t)} \tag{2.5}$$

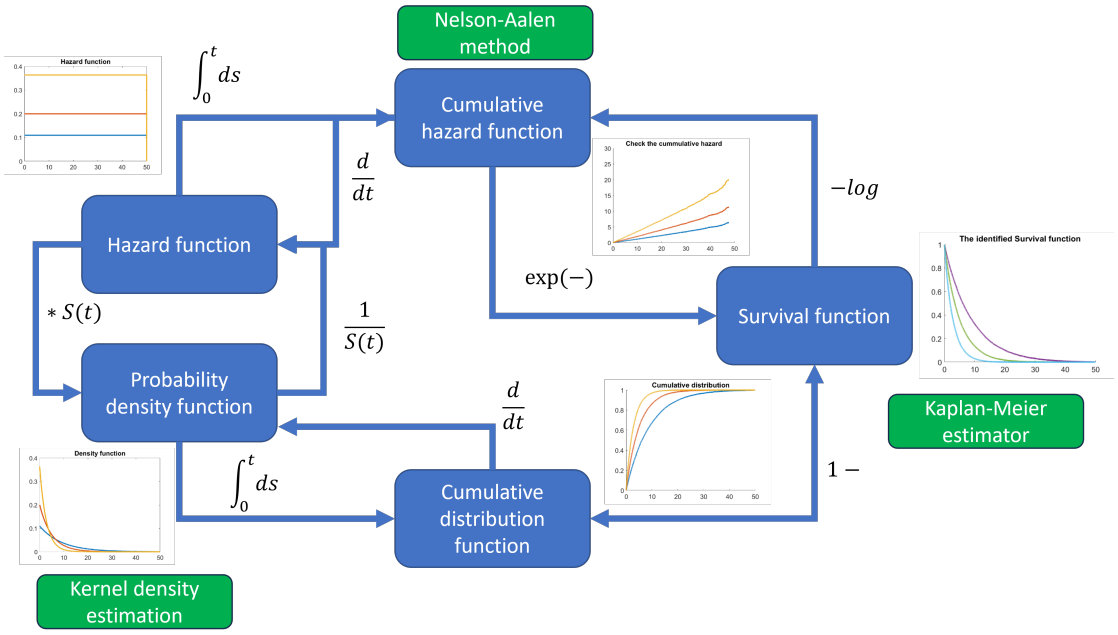

FIGURE 2.1: The interconnected mathematical framework of the functions, that represent survival models. Moreover, the examples of exponentially distributed samples, along with figures and the recommended identification methods within the green boxes are also provided.

The survival function $S(t)$ illustrates the probability that a subject will survive beyond a certain time instance t and can be expressed by the next equation:

$$S(t) = P(T > t) \tag{2.6}$$

where T denotes the random variable that represents the survival time.

The traditional survival analysis incorporates the concept of censoring, particularly useful when a sample cannot be observed until the event of interest occurs. Consequently, the exact occurrence time becomes unknown, but the available observations still provide valuable information up to the moment of censoring. In the following sections, some frequently applied methods of survival analysis are presented. Section 2.1 presents the Kaplan-Meier product-limit estimator, that is extensively employed for analyzing time until event of interest data. Section 2.2 explains the Cox semi-parametric regression method. Finally, Section 2.3 covers the maximum likelihood estimation method used to identify parametric distributions.

2.1 Kaplan-Meier estimator

The Kaplan-Meier estimator [31], also known as the product-limit estimator, is a non-parametric method used to estimate the survival function from observed data in survival analysis. It is commonly employed to analyze time-to-event data, where the focus is on the time it takes for an event of interest to occur. The survival function can be expressed by the following formula [32]:

$$S(t) = P(T > t) = \prod_{f:t_f \leq t} \left(1 - \frac{m_f}{N_f}\right) \tag{2.7}$$

where N_f represents the number of event of interest that have not been occurred or censored at time instance t_f, while m_f denotes the number of event of interest occurred between periods of time instance t_{f-1} and t_f. An occurred event refers to an event of interest that has taken place among the subjects under observation during the study period. Conversely, a censored event occurs when the event of interest has not been observed within the study period, typically due to subjects being lost to follow-up. In equation 2.7, the term $1 - \frac{m_f}{N_f}$ within the equation delineates the probability of surviving the time interval from t_{f-1} to t_f, while the multiplication of such terms signifies the overall survival probability.

An example of applying the Kaplan-Meier method is demonstrated in Table 2.1. Let us consider time instance 3, where the remaining event of interest is calculated by subtracting the occurred event of interest and censored event of interest from the remaining event of interest at the previous time instance, such as $72-8-4 = 60$.

The occurred event of interest and the censored event of interest are derived from the dataset. The probability of surviving the time interval from time instance 2 to 3 can be calculated as $1 - \frac{10}{60} = 0.8333$. The overall survival probability is the cumulative product of the probabilities of surviving the time intervals from time instance 1 to 3. The survival function derived from the column overall survival probability is illustrated in Figure 2.2. In this example, the probability that the event will last longer than 2 seconds is 0.80, while the probability that the event will last longer than 6 seconds is 0.33.

TABLE 2.1: An example of applying Kaplan-Meier method. In this dataset, if the occurred event of interests (3rd column) is 12 at time instance (1st column) one, then the database contains this specific sample one 12 times.

Time instance	**Remaining event of interests**	**Occurred event of interests**	**Censored event of interests**	**Probability of surviving the time interval**	**Overall survival probability**
1	80	8	0	0.9000	0.9000
2	72	8	4	0.8889	0.8000
3	60	10	0	0.8333	0.6667
4	50	6	0	0.8800	0.5867
5	44	10	5	0.7727	0.4533
6	29	8	0	0.7241	0.3283
7	21	16	0	0.2381	0.0782
8	5	5	0	0.0000	0.0000

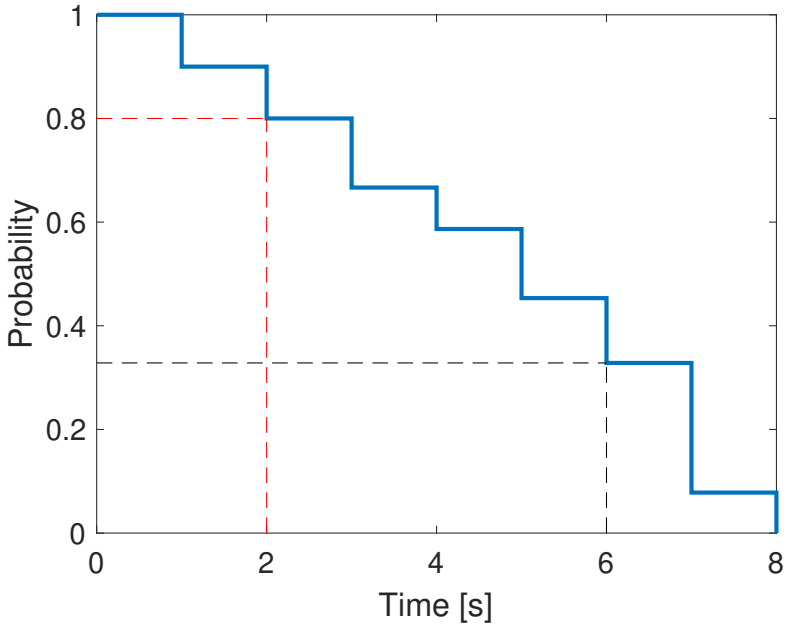

FIGURE 2.2: Example of Kaplan–Meier empirical survival function. In this example, the probability that the event will last longer than 2 seconds is 0.80, while the probability that the event will last longer than 6 seconds is 0.33.

Survival analysis can be enhanced by incorporating explanatory variables. Managing these variables with the Kaplan-Meier estimator necessitates the stratification of data, which divides the dataset according to the values of the explanatory variables. However, a more direct approach for incorporating explanatory variables in survival analysis is offered by the Cox semi-parametric regression model, which is presented in the next section.

2.2 Cox regression

The Cox regression is a semi-parametric regression method that handles explanatory variables [10]. This model directly calculates the cumulative hazard function, from which the survival function can be expressed straightforwardly based on Equation 2.2. It consists of two components: the baseline cumulative hazard function and an exponential term. The baseline cumulative hazard function is empirically derived to establish the population mean, while a parametric exponential term modifies the baseline function to reflect the cumulative hazard related to the explanatory variables. The parameters are estimated using the maximum likelihood method, as discussed in Section 2.3. The Cox proportional hazard model can be expressed by the following equation [33]:

$$\lambda(t, \boldsymbol{Z}) = \lambda_0(t) \exp\Big(\sum_{f=1}^{N_Z} \gamma_f Z_f\Big) \tag{2.8}$$

where $\lambda_0(t)$ is the baseline cumulative hazard function, N_Z stands for the number of explanatory variables, $\boldsymbol{\gamma}$ denotes the vector of model parameters and $\mathbf{Z}$ represents a vector of the relevant explanatory variables. The baseline cumulative hazard function can be estimated using the Nelson-Aalen estimator, which is calculated by the following equation [34]:

$$\lambda_0(t) = \sum_{f:t_f \leq t} \frac{m_f}{N_f} \tag{2.9}$$

where N_f represents the number of event of interest that have not been occurred or censored at time instance t_f, while m_f is the number of event of interest occurred between periods of time instance t_{f-1} and t_f.

From the cumulative hazard function, the survival function can be expressed as shown in Figure 2.1. Substituting the Equation 2.8 into Equation a 2.2 the following equation determines the survival function:

$$S(t, \boldsymbol{Z}) = S_0(t)^{\exp\left(\sum_{f=1}^{N_Z} \gamma_f Z_f\right)} \tag{2.10}$$

The application of Cox regression comes with an important assumption known as the Proportional Hazards Assumption (PHA). This assumption is crucial for the validity and reliability of the Cox proportional hazards model. The PHA essentially asserts that the hazard ratio between any two individuals remains constant over time. Deviations from this assumption can lead to biased estimates and compromised model accuracy. The hypothesis can be statistically checked with the examination of Schoenfeld residuals [35]. The survival times need to be ranked and if these ranked variables and residuals are not correlated, then the PHA assumption is satisfied. If the PHA assumption does not hold for any explanatory variable, a so-called stratified Cox model should be identified [36]. In this approach, a separate baseline cumulative hazard function is created for each predictor variable. Naturally, in this case, the stratified variable will not appear among the predictors.

In the next section, the identification of parametric probability distribution models is presented through Maximum Likelihood Estimation.

2.3 Identification of parametric probability distribution models

Maximum Likelihood Estimation is a statistical method used for estimating the parameters of a model, commonly employed in identifying parameters in probability distributions [37]. The fundamental idea is to determine the parameter values that maximize the likelihood of observing the given data under the assumed probability distribution. The likelihood function describes how well the observed data is explained by the model.

Let $X = \{x_1, \ldots, x_N\}$ denote the dataset, and $\boldsymbol{\Theta} = [\theta_1, \ldots, \theta_{N_\Theta}]$ represent the estimating parameter vector. In the initial step, the user assumes the type of

probability distribution to fit, often guided by studying the histogram of the data X. Assuming independence and identical distribution, the likelihood function of the model is expressed as a product of probability density function values for each individual observation [9]:

$$\mathcal{L}(\boldsymbol{\Theta}) = f(\boldsymbol{\Theta}|x_1) \dots f(\boldsymbol{\Theta}|x_{N_x}) \tag{2.11}$$

Multiplying a large number of samples of probabilities often results in an extremely small number, challenging to represent precisely as a double floating-point number. In practice, the logarithms of the individual likelihoods are summed:

$$\log\big(\mathcal{L}(\boldsymbol{\Theta})\big) = \log\big(f(\boldsymbol{\Theta}|x_1)\big) + \ldots + \log\big(f(\boldsymbol{\Theta}|x_{N_x})\big) \tag{2.12}$$

In maximum likelihood estimation, the goal is to find the parameter values Θ that maximize the likelihood function. However, optimization algorithms typically find minimum points rather than maximum points. Therefore, it is common practice to work with the negative log-likelihood function, as minimizing its value is equivalent to maximizing the likelihood function:

$$\underset{\boldsymbol{\Theta}}{\operatorname{argmin}}\Big(-\log\big(\mathcal{L}(\boldsymbol{\Theta})\big)\Big) \tag{2.13}$$

The next chapters present the proposed integrated survival analysis and machine learning algorithms.

Chapter 3

Integrated survival analysis and expectation-maximization-based clustering: a collection of case studies

This chapter introduces an integrated survival analysis and expectation - maximization - based clustering method designed to cluster a heterogeneous population based on the similarity of survival times and explanatory variables. Section 3.1 provides a clear overview of the challenges the algorithm aims to address and briefly introduces the algorithm. Section 3.2 presents a detailed mathematical description of the proposed algorithm. Section 3.3 demonstrates the practical application of the algorithm through several case studies, including the estimation of student dropout rates, remaining useful life of Li-ion batteries, survival chances of patients with prostate cancer, and mortality rates per 100K population of countries related to the COVID-19 pandemic.

3.1 Introduction

Traditional survival models often treat populations as homogeneous entities [38], neglecting the heterogeneity that may exist within the dataset. In many cases, populations comprise a mixture of distinct survival models, each reflecting diverse underlying dynamics [39]. The methodological foundation of understanding the survival of individual populations is closely tied to explanatory variables [40]. These variables are key elements in perception and modeling heterogeneous survival outcomes. A heterogeneous survival model comprises several homogeneous survival models, each referred to as a local model. Every local model has their designated operating domain in the variable space, and the interaction of these variables describes the factors that significantly impact survival. Whether calculating the duration until a critical event or predicting the time until the failure of a system [41], the inclusion of explanatory variables is a necessary methodological approach, not just a procedural step.

Handling explanatory variables with Kaplan-Meier estimator often necessitates the stratification of data, a straightforward task when dealing with discrete parameters. In such cases, distinct groups can be formed, enabling the estimation of survival functions for each category. However, the approach becomes notably intricate in case of continuous variables. The inherent challenge lies in effectively categorizing the data into intervals or ranges [42], a task that demands thoughtful consideration. The conventional approach involves creating individual survival curves for each stratum, representing how survival probabilities evolve across different levels of explanatory variables. While this approach is informative, it may not fully capture the nuanced dynamics of continuous variables [43]. Therefore, more sophisticated techniques are required that can identify differences within continuous parameters [44].

An alternative solution for incorporating explanatory variables in survival analysis lies in the application of the Cox proportional hazard regression model [45]. This method allows for the simultaneous examination of the impact of multiple variables on survival outcomes, by accommodating both discrete and continuous explanatory variables in a unified framework. However, it is important to note that the Cox regression model is limited to certain assumptions, most notably the proportional hazard assumption. This hypothesis posits that the hazard ratio for any two individuals is constant over time. Deviations from this assumption can

lead to biased estimates and compromised model accuracy. Although, the Cox regression model offers a robust methodology, the validity of the results needs extra effort by testing the proportional hazard hypothesis.

Beyond traditional methods, nonlinear models are also applied to handle explanatory variables, such as integrated survival analysis and machine learning frameworks. Within this expansive approaches, neural networks are frequently applied due to their exceptional accuracy. However, the significant drawback of the method lies in its interpretability [46]. The trade-off between accuracy and interpretability underscores the critical need for models that not only excel in predictive performance but also offer transparency in revealing the complex relationships between explanatory variables and survival outcomes.

The preceding discussions have underscored the challenges and limitations in handling explanatory variables within survival analysis methods. As the demand for precision and interpretability intensifies, it becomes apparent that existing methods may not fully meet the requirements posed by diverse populations and complex variables. Therefore, this chapter presents a novel integrated clustering and survival analysis method designed to create interpretable local models [47], that address these limitations comprehensively. Clustering heterogeneous subsets is a novel and under-explored method in the context of risk prediction [48]. This is an unsupervised machine learning method that explores hidden groups by partitioning data based on the similarity of objects. The key idea is that the similarity is measured based on the relationship between explanatory variables and survival times.

The proposed method operates by simultaneously identifying local models based on explanatory variables and survival data. In the continuous variable space, the local models are represented by multivariate Gaussian mixture models [49]. The discrete variables are represented by their proportions. Each local model is assumed to be characterized by distinct survival properties. The survival probability is modeled using parametric approaches, with a specific focus on the Weibull distribution [50], chosen for its widespread relevance and applications across various scientific fields [51]. Such a mixture of the Weibull model has already been built for modelling reliability [41] and time-to-event analysis [39]. The model parameters are determined by an iterative approach called the expectation-maximization algorithm [52], which estimates the parameters of multivariate Gaussian and Weibull distributions for each cluster. Furthermore, it is noteworthy that this approach

exhibits considerable flexibility. Its core elements, the probability distribution models, can be substituted to accommodate different probability characteristics in a given case study.

To enhance the interpretability of continuous variables, their impact on cluster membership are defined using fuzzy logic. In this approach, interpretable fuzzy if-then rules are employed to represent the domain of local models [49]. These rules consist of a rule antecedent and a rule consequent part. Specifically, Takagi-Sugeno fuzzy rules apply functions of the input variables in the rule consequent [53]. As the function within this fuzzy model, the Gaussian membership function is applied, which additionally describes the operating regions of the continuous variables. Consequently, the resulting fuzzy approach can be considered as a special case of Gaussian mixture model. Since, the domain of these function designates the range of continuous variables that indicate the same survival dynamics, the clusters also serve as categories for the continuous variables.

The proposed methodology results in a compact output, where local models are identified in clusters, each represented by a multitude of probability distributions. Furthermore, Gaussian membership functions not only aid in the interpretation of continuous variables but also serve to categorize them. Notably, this method effectively handles explanatory variables without the need for additional hypothesis testing, as seen in the case of Cox regression. The Gaussian fuzzy membership models contribute to an interpretable and transparent visualization of variable spaces. One of the limitation of the method lies in the fact that only parametric distribution models can be placed in the algorithm. Nevertheless, leveraging cluster membership values allows for the estimation of the empirical survival function, albeit indirectly and contingent on the results of the method. Another drawback is that the identified clusters highlight population differences, but these differences do not inherently imply causality. Therefore, additional causality analysis can be necessary. The specific contributions of the chapter can be outlined as follows:

- Heterogeneous survival models can be clustered into homogeneous models based on the similarity of survival times and explanatory variables by utilizing an expectation maximization algorithm. This approach defines clusters and simultaneously identifies their survival probabilities using the Weibull distribution and the related explanatory variables, leveraging multivariate Gaussian distributions.

- The cluster memberships of the continuous variables can be represented using Takagi-Sugeno fuzzy rules, providing a framework to determine the operating domain of continuous variables. With this approach, survival characteristics can be described by considering both the domain of continuous variables and the proportions of discrete variables. The method is versatile, and the clusters indicate categorization for continuous variables.

The effectiveness of the proposed algorithm is demonstrated across diverse case studies, showcasing its applicability across various domains. The student dropout rate is estimated based on the average grade during the first semester and enrollment scores. Additionally, the remaining useful life of Li-ion batteries is estimated based on their capacity, internal resistance, and charging condition. Furthermore, survival chances for patients with prostate cancer are determined based on their age, serum hemoglobin level, and treatment. Finally, the mortality rate per 100K population of countries related to the COVID-19 pandemic is estimated based on demographical and economical data.

The next section presents the mathematical description of the proposed methodology.

3.2 Description of the mixture of survival models

This section outlines the proposed method. Subsection 3.2.1 details the representation of the model, Subsection 3.2.2 covers the estimation of model parameters, Subsection 3.2.3 details a method for selecting the number of clusters, and finally, in Subsection 3.2.4, the algorithm is summarized.

3.2.1 Model representation

This method is based on a mixture of models, which have already been used in survival analysis [48]. The distribution of survival times is represented by their probability density function $f(t|\mathbf{\Theta})$, expressed as the weighted sum of component distributions [54]:

$$f(t|\mathbf{\Theta}) = \sum_{j=1}^{M} \omega_j f_j(t|\Theta_j) \tag{3.1}$$

where M stands for the number of components, Θ_j denotes the distribution parameters of the jth component, ω_j represents the weight of the jth component and $\sum_{j=1}^{M} \omega_j = 1$. Each component is characterized by distinct survival dynamics, assuming they have the same distributional form but varying parameters. The distribution of the components is further discussed using a clustering-based approach, where the dependence of Θ is not specified for the purpose of simplicity:

$$p(t) = f(t|\mathbf{\Theta}) = \sum_{j=1}^{M} p(t|j)p(j) \tag{3.2}$$

where $p(t|j)$ stands for the probability that the survival event occurs at time instance t, given that the sample belongs to cluster j. Meanwhile, $p(j)$ represents the probability that an element belongs to cluster j. This unconditioned cluster probability satisfies the conditions $\sum_{j=1}^{M} p(j) = 1$, and $0 \leq p(j) \leq 1$. The distribution of the components utilizes the principles of the law of total probability to model survival dynamics. In this approach, survival patterns are grouped into distinct clusters based on their distributional characteristics.

The conditional probability of the survival times $p(t|j)$ can be characterized by various distributions. In this study, the Weibull distribution is employed due to its wide-range applications across diverse fields such as reliability [55], lifetime [56] or risk analysis [57]. The distribution is represented by the parameter vector $\mathbf{\Theta}_j = [\theta_j, \beta_j]$, where θ_j denotes the scale parameter, and β_j the shape parameter [58]:

$$p(t|j) = f(t|\mathbf{\Theta}_j) = \frac{\beta_j}{\theta_j} \left(\frac{t}{\theta_j}\right)^{\beta_j - 1} \exp\left(-\left(\frac{t}{\theta_j}\right)^{\beta_j}\right) \tag{3.3}$$

By substituting (3.3) for (3.2), the entire model can be expressed as:

$$p(t|j) = \sum_{j=1}^{M} \frac{\beta_j}{\theta_j} \left(\frac{t}{\theta_j}\right)^{\beta_j - 1} \exp\left(-\left(\frac{t}{\theta_j}\right)^{\beta_j}\right) p(j) \tag{3.4}$$

The developed method is a cluster-weighted model [59], designed to partition the domain of variables $\mathbf{x}$ into local models. In this case, variables can be divided into continuous and discrete ones, represented as $\mathbf{x} = [\mathbf{x}^c, \mathbf{x}^d]$, where $\mathbf{x}^d = \left[x_1^d \ldots, x_{N_d}^d\right]^T$ denotes the vector of the discrete variables and $\mathbf{x}^c = \left[x_1^c, \ldots, x_{N_c}^c\right]^T$ stands for the vector of continuous variables. Assuming independence between continuous and discrete variables, the variable- and time-based probabilities can be calculated as:

$$p(t, \mathbf{x}) = p(t, \mathbf{x}^c, \mathbf{x}^d) = \sum_{j=1}^{M} p(t|j)p(\mathbf{x}^c|j)p(\mathbf{x}^d|j)p(j) \tag{3.5}$$

The fundamental concept is that each component of the mixture model operates within its designated domain, which depends on both continuous and discrete variables. Continuous variables are also represented by a probability distribution, particularly by their probability density function. In this paper, these covariates are modeled using a multivariate Gaussian distribution, where $p(\mathbf{x}^c|j)$ can be expressed as:

$$p(\mathbf{x}^c|j) = f(\mathbf{x}^c|\mathbf{\Theta}_j^*) = \frac{(|\mathbf{F}_j^{-1}|)^{1/2}}{(2\pi)^{M/2}} \exp\left(-\frac{1}{2}\left(\mathbf{x} - \mathbf{v}_j\right)^T \mathbf{F}_j^{-1}\left(\mathbf{x} - \mathbf{v}_j\right)\right) \tag{3.6}$$

This distribution is described by the parameter set $\mathbf{\Theta}_j^* = \{\mathbf{v}_j, \mathbf{F}_j\}$, where $\mathbf{v}_j$ denotes the center and $\mathbf{F}_j$ represents the covariance matrix of the jth Gaussian distribution. Moreover, this distribution defines the applicability domain of the jth model in the space of the $\mathbf{x}^c$ variables.

The discrete variables are assumed to be independent, so the probability $p(\mathbf{x}_q^d|j)$ is described by the following equation:

$$p(\mathbf{x}^d|j) = \prod_{q=1}^{N_d} p(x_q^d|j) \tag{3.7}$$

The covariance matrix of Gaussian functions is diagonal in a special case, when the variables are independent. Therefore, the Gaussian functions can also be interpreted as a fuzzy membership function. The rule of fuzzy logic can be formulated as:

r_j : If x_1^c is $A_{j,1}(x_1^c)$ and $\dots x_{N_c}^c$ is $A_{j,N_c}(x_{N_c}^c)$, then $\hat{y} = \exp\left(-\left(\frac{t}{\theta_j}\right)^{\beta_j}\right)$, $[\omega_j]$, where $[\omega_j]$ represents the weight of the j-th rule.

The linguistic variables of the fuzzy rules can be represented by Gaussian membership functions [60]:

$$A_{j,i}(\mathbf{x^c}) = \exp\left(-\frac{1}{2}\frac{(x_j^c - v_{j,i})^2}{\sigma_{j,i}^2}\right) \tag{3.8}$$

where i denotes the index of the explanatory variables, and σ stands for the variance of Gaussian functions. This representation provides an approximation of the identified mixture model, mapping the multiple-dimensional variable space into single dimensions. The reason for this transformation is that the multivariate Gaussian distribution becomes challenging to interpret in multiple dimensions. Consequently, simplifying the individual components becomes necessary for better visualization. However, this simplification comes at a cost that Equation 3.8 utilizes only the variance for interpretation, neglecting the potential influence of covariance values.

In this subsection, the model structure has been presented. In the next subsection, it will be shown how the parameters of the proposed model can be estimated.

3.2.2 Estimation of the model parameters

In the context of estimating the parameters of probability distributions, the aim is to maximize the log-likelihood. However, many optimization algorithms are conventionally designed as minimizers. Therefore, the negative log-likelihood function is employed, as minimizing its value corresponds to maximizing the log-likelihood. The negative log-likelihood of the data set can be expressed as:

$$\mathcal{L} = -\sum_{n=1}^{N} \log p(t_n, \mathbf{x}_n^c, \mathbf{x}_n^d) = -\sum_{n=1}^{N} \log\left(\sum_{j=1}^{M} p(t_n|j)p(\mathbf{x}_n^c|j)p(\mathbf{x}_n^d|j)p(j)\right) \tag{3.9}$$

where N denotes the number of independent observations in the given data set, where every sample is characterized by a survival time, along with a set of continuous and discrete parameters. Additionally, log stands for the natural logarithm.

The proposed method employs the expectation-maximization algorithm for estimating the model parameters. The method is an iterative optimization method used to estimate parameters in statistical models and involves two main steps: the E-step (Expectation step) and the M-step (Maximization step). In the E step, the algorithm calculates the posterior probabilities. These probabilities signify the likelihood that a specific data point is originated from a particular cluster, thereby serving as a membership value. This concept does not only reflect the association of a data point with a cluster but also establishes a soft clustering. The cluster membership value can be calculated accordingly:

$$p(j|t_n, \mathbf{x}_n^c, \mathbf{x}_n^d) = \frac{p(t_n, \mathbf{x}_n^c, \mathbf{x}_n^d|j)}{p(t_n, \mathbf{x}_n^c, \mathbf{x}_n^d)} = \frac{p(t_n|j)p(\mathbf{x}_n^c|j)p(\mathbf{x}_n^d|j)p(j)}{\sum_{l=1}^{M} p_l(t_n|l)p(\mathbf{x}_n^c|l)p(\mathbf{x}_n^d|l)p(l)} \tag{3.10}$$

In the M-step, the algorithm updates the model parameters to maximize the expected log-likelihood based on the estimated membership values from the E-step. Given that the E-step altered the membership values, the model parameters must be recalculated. Initially, the unconditional cluster probability needs to be recalculated. This probability corresponds to the weighted parameter ω_j in Equation 3.1, representing the significance of a given cluster. The calculation of the unconditional cluster probability is expressed by the following equation

$$p(j) = \frac{1}{N} \sum_{n=1}^{N} p(j|t_n, \mathbf{x}_n^c, \mathbf{x}_n^d) \tag{3.11}$$

The recalculated unconditional cluster probability can be employed to compute the new mean (or center) of the clusters in the multivariate Gaussian model:

$$\mathbf{v}_j = \frac{1}{Np(j)} \sum_{n=1}^{N} \mathbf{x}_n^c p(j|t_n, \mathbf{x}_n^c, \mathbf{x}_n^d) \tag{3.12}$$

where $\mathbf{v}_j$ denotes the center of jth cluster. Similarly, the weighted covariance matrix can be calculated through the following steps:

$$\mathbf{F}_j = \frac{1}{Np(j)} \sum_{n=1}^{N} (\mathbf{x}_n^c - \mathbf{v}_j) (\mathbf{x}_n^c - \mathbf{v}_j)^T p(j|t_n, \mathbf{x}_n^c, \mathbf{x}_n^d) \tag{3.13}$$

where $\mathbf{F}_j$ stands for the covariance matrix of the jth cluster. The above equations collectively constitute the new parameter set of the multivariate Gaussian mixture distribution. To determine the parameters of the local Weibull distributions, minimization of (3.9) is employed. The formula is expressed as follows:

$$\underset{\beta_j,\theta_j}{\operatorname{argmin}}(\mathcal{L}) = -\sum_{n=1}^{N} \log\left(\sum_{j=1}^{M} \frac{\beta_j}{\theta_j}\left(\frac{t}{\theta_j}\right)^{\beta_j-1} \exp\left(-\left(\frac{t}{\theta_j}\right)^{\beta_j}\right) p(\mathbf{x_n}|j)p(j)\right) \quad (3.14)$$

This formula can be considered as a cluster-weighted estimation of M Weibull distributions. The parameters can be estimated by solving the following system of nonlinear equations:

$$\theta_j = \left(\frac{1}{Np(j)} \sum_{n=1}^{N} t_n^{\beta_j} p(j|t_n, \mathbf{x}_n^c, \mathbf{x}_n^d)\right)^{\frac{1}{\beta_j}} \quad (3.15)$$

$$\beta_j = \left[\frac{\sum_{n=1}^{N} t_n^{\beta_j} \log(t_n) p(j|t_n, \mathbf{x}_n^c, \mathbf{x}_n^d)}{\sum_{n=1}^{N} t_n^{\beta_j} p(j|t_n, \mathbf{x}_n^c, \mathbf{x}_n^d)} - \frac{\sum_{n=1}^{N} \log(t_n) p(j|t_n, \mathbf{x}_n^c, \mathbf{x}_n^d)}{p(j)}\right]^{-1} \quad (3.16)$$

The conditional probabilities $p(x_q^d|j)$ are also regarded as parameters of the jth clusters, where the parameters $p(j|x_q^d = z)$ indicate that the qth discrete variable takes on a specific value z. These parameters can be calculated using Bayes' theorem:

$$p(x_q^d = z|j) = \frac{p(j|x_q^d = z)p(x_q^d = z)}{p(j)} \quad (3.17)$$

The calculation of the conditional probability $p(j|x_q^d = z)$ can be expressed as:

$$p(j|x_q^d = z) = \frac{1}{N} \sum_{n=1}^{N} I(x_q^d = z) p(j|t_n, \mathbf{x}_n^c, \mathbf{x}_n^d) \quad (3.18)$$

where $I(x_d^q = z)$ denotes a characteristic function that is one if $x_q^d = z$, but zero otherwise. The unconditional probability $p(x_q^d)$ is estimated based on the relative frequency of cases when the discrete variable x_q^d takes on a certain value z:

$$p(x_q^d = z) = \frac{1}{N} \sum_{n=1}^{N} I(x_{n,q}^d = z) \tag{3.19}$$

The importance of selecting the proper number of clusters is underscored by its impact on the effectiveness of clustering algorithms. In the next section, the significance of this decision will be explored, and methodologies for determining the appropriate number of clusters will be discussed.

3.2.3 Determining the number of clusters by Akaike information criterion

One of the hyperparameters of the algorithm is the number of clusters that must be identified. The proper selection of this number is a crucial step, as the number of clusters has a significant effect on accuracy. Information criteria can assist in choosing the most suitable number of clusters, and several different types of such criteria are available [61]. For selecting an adequate number of components in the mixture model, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion are commonly applied [62]. The calculation of these criteria is described by the following equations [63]:

$$AIC = 2K - 2\mathcal{L} \tag{3.20}$$

$$BIC = \log(N)K - 2\mathcal{L} \tag{3.21}$$

where K denotes the total number of estimated model parameters, log represents the natural logarithm, and N stands for the number of samples. The parameter K can be determined based on the dimensions of the parameter vectors and matrices. The matrix $\mathbf{v}$ is $M \times N_c$ and of matrix $\mathbf{F}$ are $N_c \times N_c \times M$; the vector θ is calculated by $1 \times M$ and the vector β by $1 \times M$ the unconditional cluster probability $\mathbf{p}$ is a vector $1 \times M$ and the probability matrix $\mathbf{p}(\mathbf{x^d}|\mathbf{j})$ are $m \times N_d^*$, where N_d^* denotes the total number of possible combinations of $\mathbf{x}^d$. The elements of every parameter must be summed as follows:

$$K = MN_c + N_cN_cM + M + M + M + MN_d^* \tag{3.22}$$

Both criteria excel in selecting the optimal number of clusters. However, In a context where each cluster has the same amount of data, AIC is more practical as it is independent of the amount of data. It is important to note that AIC provides relative information between models [64]. Even though it is unable to estimate how well a model explains the data, this criterion can suggest which option is the most suitable. The model with the preferred number of components has the lowest AIC value.

3.2.4 Summary of the proposed method

The method outlined in the preceding sections can be briefly summarized as follows: The proposed approach applies an expectation-maximization algorithm. The algorithmic process is visually depicted in the flowchart provided in Figure 3.1. The gray arrows represent feedback information, while the yellow arrows denote iterative steps. The algorithm receives inputs derived from the data conditioning process, encompassing survival times, discrete, and continuous explanatory variables. The purpose of the expectation-maximization algorithm is to determine the parameters of the corresponding distributions $(\theta_j, \beta_j\, \mathbf{v}_j, \mathbf{F}_j)$. During the iteration, these parameters alternate from their initial values to the accepted solution. The expectation-maximization algorithm consists of two main steps, the E and M steps.

In the E-step, cluster membership values, $p(j|t, \mathbf{x}^c, \mathbf{x}^d)$, are assigned to each data point based on Equation 3.10. These values are the key for clustering the observations. Equation 3.10 is derived from Equation 3.5, which comprises four components corresponding to survival times, continuous variables, discrete variables, and cluster probabilities. The probability of survival time, $p(t|j)$, follows a Weibull distribution and can be determined using Equation (3.3). Similarly, the probabilities of continuous variables, $p(x^c|j)$, are identified based on the multivariate Gaussian distribution using Equation 3.6. Finally, the probabilities of discrete variables, $p(x^d|j)$, are determined by Equation (3.7). During the E-step, the membership values of the data points change. The M-step tunes the Weibull $\{\theta_j, \beta_j\}$ (see (3.15) and (3.16), respectively) and Gaussian parameters $\{\mathbf{v}_j, \mathbf{F}_j\}$ (see (3.12) and (3.13), respectively). Finally, the probability of the cluster itself is based on Equation 3.7. Now, the next iteration step begins once more using these tuned parameters.

Given the iterative nature of the algorithm, initialization of Weibull and Gaussian parameters for each cluster, as well as membership values for each cluster and sample, is required. The hyperparameters of the algorithm include the number of clusters and the maximum number of iterations. The number of clusters can be estimated by the Akaike Information Criterion. The next section presents a collection of case studies to demonstrate the effectiveness of the proposed algorithm.

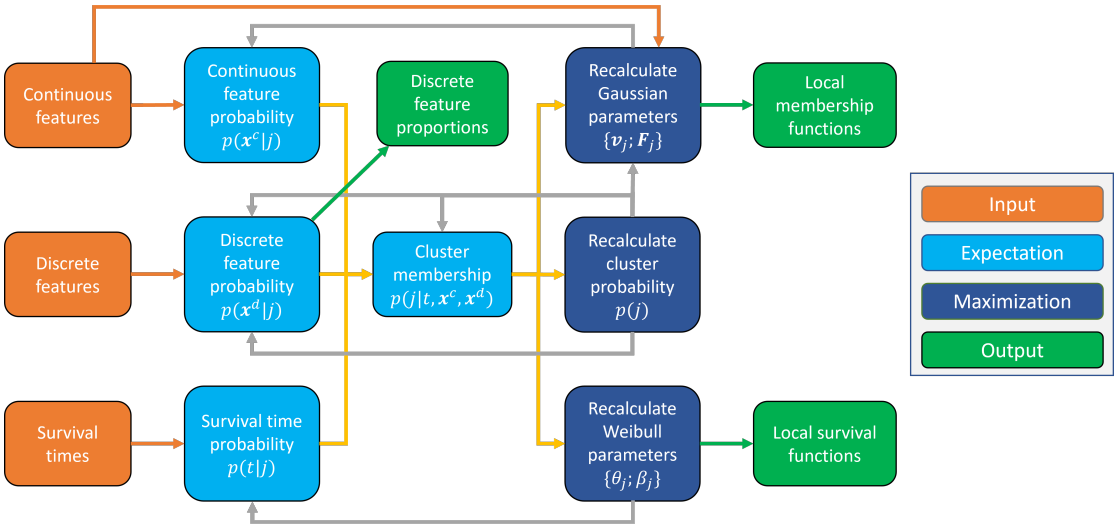

FIGURE 3.1: Flowchart of the proposed expectation-maximization algorithm. In the expectation step, every cluster membership value is determined for each data point. Data points undergo changes in their membership values during this step. The maximization step adjusts the Weibull and Gaussian parameters based on the recalculated new membership values. The gray arrows indicate the feedback information, while the yellow arrows denote the progression of the iteration step.

3.3 Case studies

In this section, the application of the proposed methodology is presented through four different case studies. Subsection 3.3.1 addresses student dropout, utilizing data from a Hungarian university, with estimates relying on the average grade during the first semester and enrollment scores. Subsection 3.3.2 focuses on estimating the remaining useful life of Li-ion batteries, with the dataset available on the MATLAB website. The calculation is based on the capacity, internal resistance, and charging condition of Li-ion batteries. Subsection 3.3.3 discusses the survival rate of patients with prostate cancer, considering factors such as age, serum

hemoglobin level, and treatment. Finally, Subsection 3.3.4 demonstrates an analysis of the distribution of the COVID-19 mortality rate based on demographical and economical data.

3.3.1 Analysis of student dropout

The proposed method is applied to examine the dropout rate of chemical engineering students at the University of Pannonia in Hungary. The goal of the examination is to present a simple example, how the graduation chances of the students are influenced by their performance. Therefore, performance-relevant variables such as the average grade of subjects in the first semester and the enrollment score are utilized. No discrete variables are covered in this case. The examination consists of data from 349 students. The data representing the university path of students is integrated from three different databases. This includes completion records of students based on the subjects taken and grades achieved, enrollment data, as well as sample curricula. Integration of these three databases provides the following information:

- how many semesters the student spent at the university (survival time)
- the outcome of his / her training was dropout or graduation (competing risks)
- what the average grade of subjects completed during the first semester was (continuous variable)
- how many enrollment scores he/she had (continuous variable)

Grades are rated on a scale of 1 to 5, with 1 being the worst and 5 being the best. The enrollment score can reach a maximum of 500. A high enrollment score indicates that the student has performed well in their previous high school endeavors. The available survival times indicate how many semesters a student spent at the university. The training outcome can result in two ways: graduation or dropping out. Survival analysis is typically suited for handling a single outcome. Therefore, traditional survival analysis should be extended to handle the outcome as competing risks for estimating the exact dropout rate [65]. However, since handling competing risks with the proposed methodology is beyond the scope, students

who graduated were censored during the identification of Weibull distribution. The handling of competing risks is addressed in a later chapter.

The applied continuous variables are the average grades and enrollment scores of students during the first semester. The less correlated, the more complex influences the variables have on the dropout probabilities. The correlation coefficient between the two variables is 0.65, indicating a certain degree of independence. The histogram of the average grades and enrollment scores during the first semester can be seen in Figure 3.2. These histograms can be compared later with the Gaussian membership functions.

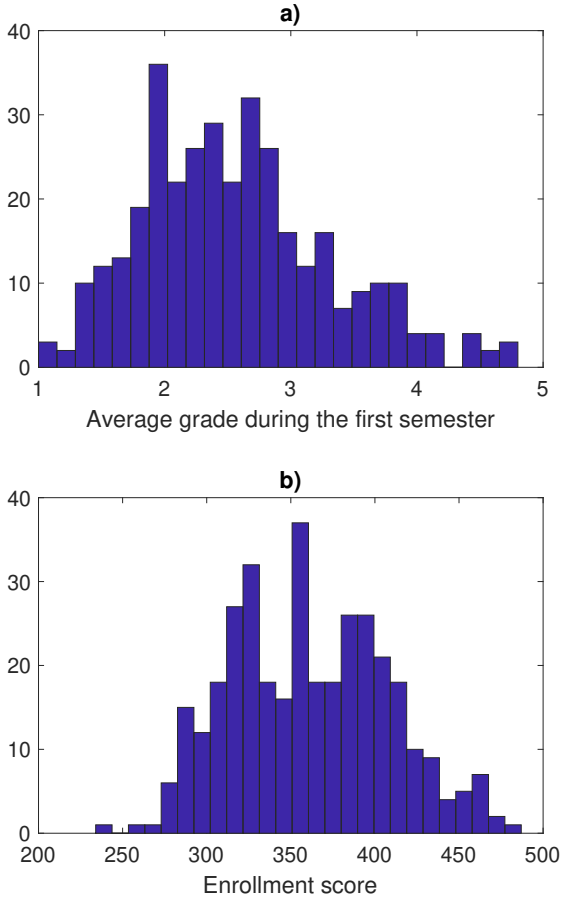

FIGURE 3.2: Histogram of the average grades of students during the first semester *a*) and enrollment scores *b*), which can be compared to the Gaussian membership functions. The peaks of the membership functions must match the peaks of the histogram.

Given that the analysis continues with the estimation of the optimal number of components, the value of the Akaike Information Criterion is computed and presented in Table 3.1. For this estimation, A maximum of 5 sub-models is considered, as convergence problems may arise during the calculation of $\mathcal{L}$ if more than five

sub-models are accounted for. The smallest AIC value corresponds to the optimal number of components. Therefore, 3 local models are required.

TABLE 3.1: The values of the Akaike information criterion. The component number with the smallest AIC value shall be used. In this case study, 3 local model is required.

No of components	**One**	**Two**	**Three**	**Four**	**Five**
AIC value	12721	12620	12566	12568	12584
Log-likelihood	-6350	-6288	-6250	-6240	-6237

The survival function of the resulting clusters can be seen in Figure 3.3. The chemical engineering training lasts for seven semesters in Hungary. However, due to the Bologna system and government laws, a student can spend a maximum of 11 semesters in the training, so functions are plotted only up to the 11th semester. The Gaussian membership functions can describe the operating domain of certain continuous variables for a specific cluster, which can be observed in Figure 3.4. The proportion of the cluster members is also collected in Table 3.2.

TABLE 3.2: The proportions of the clusters. Describes how many of the students belong to a given cluster. Cluster 1 dominates in terms of membership, to which 85 percent of students belong.

	Cluster 1	**Cluster 2**	**Cluster 3**
Cluster proportions	0.8503	0.0456	0.1040

The Gaussian membership functions show that the three clusters divide students into three groups according to their probability of graduation. Cluster 1 represents the worst performing students who have the lowest probability of graduation (see in Figure 3.3 and 3.4). These students also have the lowest grades and enrollment scores. Cluster 2 represents the set of the best performing students. These students exhibit the highest probability of graduation, as evidenced by their top enrollment scores and average grades in the first semester. Accordingly, Cluster 3 represents the cohort of the students showing medium performance and graduation probability.

The Weibull survival function describes the probability that a student continues to study beyond a certain semester. Upon examining the survival function, the probability of low-performing students continuing to study for more than four semesters is 60%, while this figure exceeds 90% in the case of clusters representing medium and high-performing students. Empirical evidence suggests that the

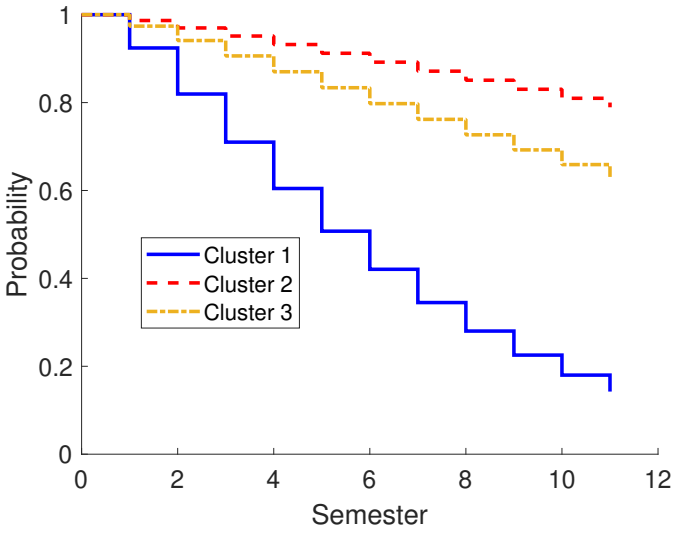

FIGURE 3.3: The survival function of students at the university. By examining the results, the probability that a low-performing student studies for more than four semesters is 60%, while such a probability is above 90% in clusters of the medium and best-performing ones. The graph can be compared to the Gaussian membership functions. The peaks of the membership functions must align the peaks of the histogram.

worst-performing students tend to drop out sooner than those who perform better. This empirical observation supports the conclusion that the model effectively captures the distribution of dropout.

The Weibull parameters in Table 3.3 provide valuable insights. The parameter β determines how the dropout rate changes over time. For all three clusters, this parameter is greater than 1, indicating that the dropout rate increases over time. The magnitude of the parameter β determines the significance of this increase. In Cluster 1, representing students who perform worse, this value is higher, suggesting a more significant increase in the dropout rate over time. Since the best performing students are represented in Cluster 2, the lowest value of β would be expected for this cluster. Surprisingly, Cluster 3, representing students with medium performance, has the lowest value of β. Although, the difference is negligible, this phenomenon might be caused by disproportional memberships (most of the students belong to Cluster 1).

The overall survival can be estimated by the weighted arithmetic mean of the local survival functions using Equation 3.2. A comparison of overall survival is conducted with the baseline function of the Cox model and the empirical survival function determined by the Kaplan-Meier method, as illustrated in Figure 3.5. It is important to note that the overall distribution may deviate from the empirical survival

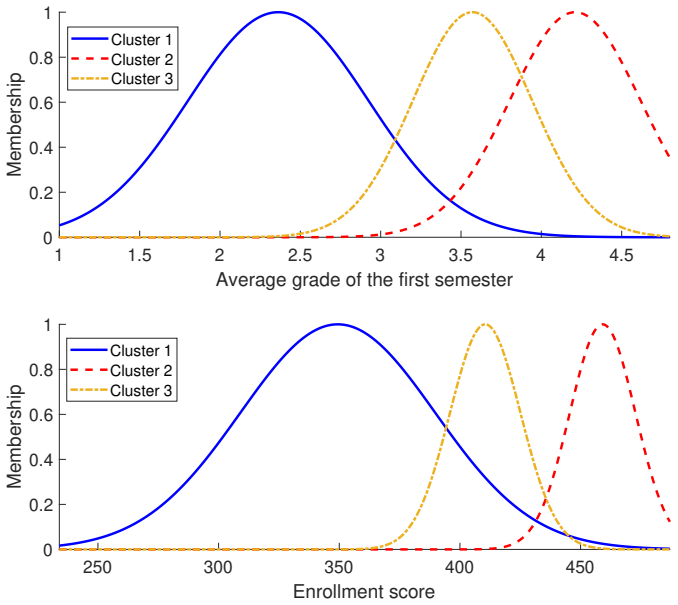

FIGURE 3.4: The Gaussian membership function of the clusters. Three clusters can be observed. The blue function describes the worst-performing students who belong to Cluster 1 (low average grade, low enrollment score), the yellow line represents medium ones who belong to Cluster 3, and the red curve denotes the best students, who belong to Cluster 2.

TABLE 3.3: The Weibull parameters. The parameter β is higher for students who perform worse (Cluster 1), suggesting that the dropout rate increases more significantly in that cluster over time. The best- and medium-performing students are characterized by almost the same values.

Weibull parameter	**Cluster 1**	**Cluster 2**	**Cluster 3**
θ	6.6819	36.5908	20.7377
β	1.3381	1.1994	1.1988

function due to the inherent limitation of the proposed Weibull distribution-based algorithm, which is related to its parametric nature. Although, the deviation is not significant, it demonstrates the limitation of the method, as the flexibility of the local functions significantly determine the accuracy of the resulted mixture model.

The local models were further compared to the Cox Proportional Hazard Model. The local survival functions were calculated in the local mean values of the Gaussian distribution. The results can be seen in Figure 3.6. The local mixture models and the estimates of the Cox survival functions slightly differ, which difference is also caused by the fact that the local functions are based on the less flexible Weibull distributions. Moreover, the Cox regression assumes proportional hazards, a condition that is not met in this case due to the identified β parameters.

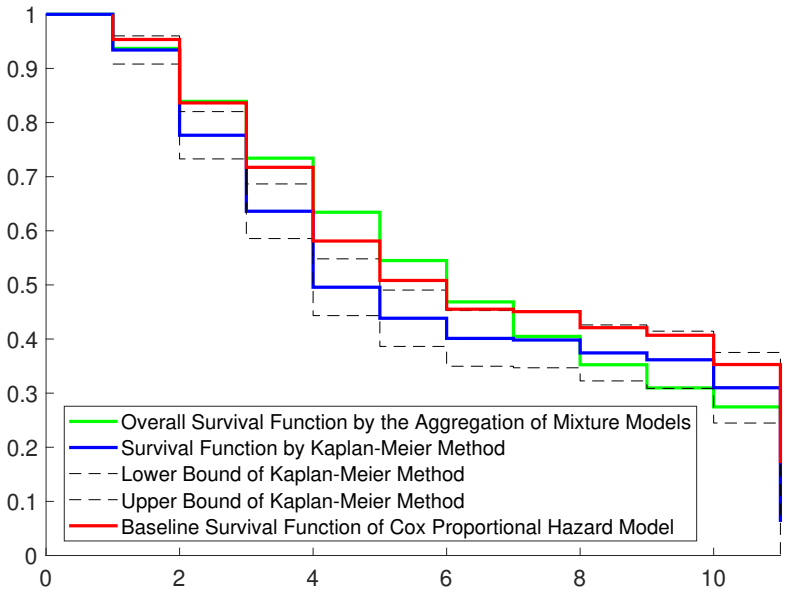

FIGURE 3.5: The overall survival functions, estimated by the weighted arithmetic mean of the local survival functions; the empirical survival function, calculated by the Kaplan-Meier method; the baseline function of the Cox regression model.

The distinct values of these parameters indicate that the hazard functions in those clusters are not proportional.

After some refinements, the model can be applied to determine the distributions of student dropout. This model could alert students if they enter an "at-risk" state during their studies.

3.3.2 Estimation of remaining useful life of Li-ion batteries

Models of survival analysis play a significant role in estimating the remaining useful life in the technical field. This method is beneficial when there is a need to alert for potential future failures based on explanatory variables, that describe the operational circumstances. In this section, the proposed methodology is applied to estimate the Remaining Useful Life (RUL) of lithium-ion batteries, where the charge-discharge cycles serve as the main indicator of survival. Establishing a remaining useful life approach requires a comprehensive understanding of the given technology for finding the most significant features that influence the survival. However, this time, such expert knowledge is not available. Therefore, data preprocessing and feature selection have been performed based on a prior study [66].

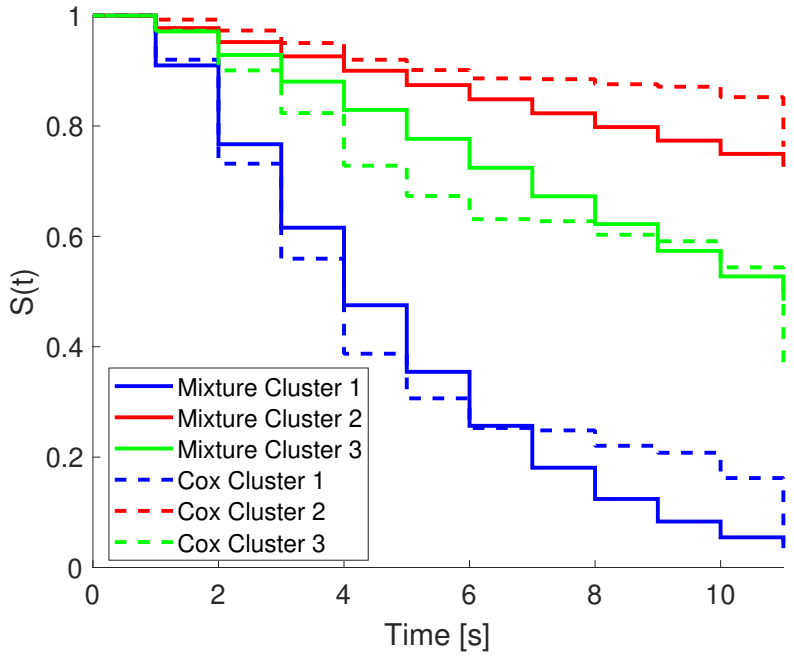

FIGURE 3.6: The local survival models and the estimates of the Cox proportional hazard model. The small difference between the local models and the related estimates is caused by the fact that the local functions are based on less flexible Weibull distributions.

This case study serves as a benchmark problem and is also accessible in the predictive maintenance section of MATLAB. The dataset can be downloaded from the following link: `https://data.matr.io/1/`.

This dataset comprises time series measurements of charge-discharge cycles from 124 lithium-ion batteries, each with a nominal capacity of 1.1 Ah and a nominal voltage of 3.3 V. These collected data were divided into training (81) and validation data (43). The independent variable is the charge cycle, and functions that describe their discharging circumstances in time series are preprocessed and transformed into aggregated constant features, including linear and nonlinear transformations. This approach is necessary as the proposed method cannot handle time series data directly but requires constant features. The change in discharge voltage curves between cycles is a highly effective predictor of the lifetime. Specifically, the difference in discharge capacity as a function of voltage between the 100th and 10th cycles is examined. For practical reasons, these values are transformed into logarithms. Both the variance and the minimum of these functions are taken into consideration. The following variables describe the remaining useful life of Li-ion batteries:

- $Q_{100-10(V)}$ log variance (QLV)

- $Q_{100-10(V)}$ log minimum (QLM)
- Slope of linear fit to capacity fade curve between cycles 2 and 100 (SCFC)
- Intercept of linear fit to capacity fade curve between cycles 2 and 100 (ICFC)
- Discharge capacity during cycle 2 (DC2)
- Average charging time over the first 5 cycles (ACT5)
- Minimum Internal Resistance between cycles 2 and 100 (MIR)
- Internal Resistance difference between cycles 2 and 100 (IRD)

The AIC can also be used to quantify the contribution of individual variables to the goodness of the model, serving as feature selection. After executing the algorithm on the dataset consisting only of a subset of all features, the resulting AIC value should be compared to the AIC value of the model identified based on the full set of variables. If the AIC value decreases upon elimination, it suggests a worsening of the model, whereas an increase indicates an improvement. Based on the AIC analysis, all the variables should be included in the model when creating three clusters. In this example, there were no discrete variables.

After running the clustering algorithm with the selected number of components, the resultant distributions and Gaussian membership functions are depicted in Figure 3.7 and 3.8, respectively. The parameters of the Weibull distribution and the resulting cluster proportions are presented in Table 3.4 and 3.5. The available variables cover different areas. Some variables refer to the initial moment of the devices, i.e. the starting and production conditions (DC2, ICFC, ACT5), while others cover the initial operating period (SCFC, IRD, QLM, MIR, QLV). By examining the resulting clusters, it can be conducted that the median survival time (the one corresponding to the 50% probability) is the smallest in Cluster 1 and the largest in Cluster 3. Therefore, battery life is maximized under Cluster 3. This information can also be detected based on the parameter θ of the Weibull distribution.

The features related to the initial operating period provide valuable insights into the characterizing changes over time. The underlying concept is that the useful life is longer when it maintains its initial state or experiences minimal deviations from it. Since QLM and QLV are logarithmized values, smaller values indicate

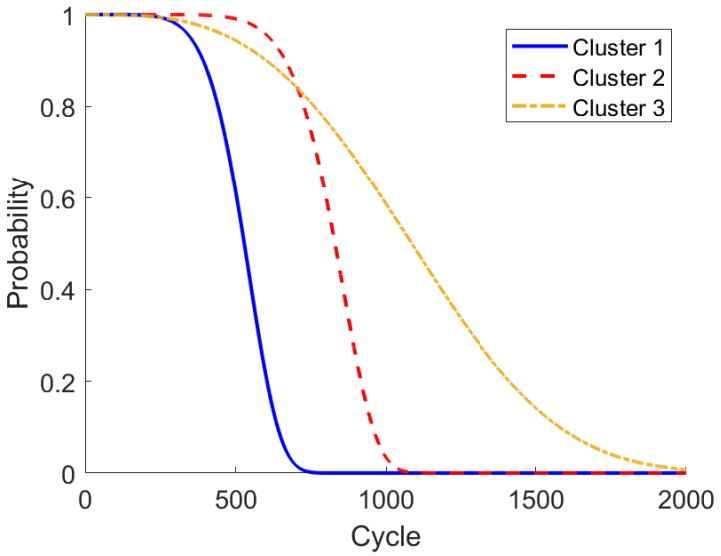

FIGURE 3.7: The Weibull survival function of the failure rates. It can be assumed that batteries in Cluster 1 have the shortest lifetime, while those in Cluster 3 have the longest.

TABLE 3.4: The Weibull parameters. The parameter θ denotes the order of failure time order. It can be assumed that batteries in Cluster 1 have the shortest lifetime, while those in Cluster 3 have the longest. If the degradation process is initially significant, the β denotes the failure rates.

Weibull parameter	**Cluster 1**	**Cluster 2**	**Cluster 3**
θ	560.4124	866.7754	1217.1980
β	6.3060	8.5254	3.2010

TABLE 3.5: The proportions of the clusters, which describe the proportion of the batteries that belong to a given cluster. Cluster 3 dominates in terms of membership, to which almost 50 percent of the batteries belong.

	Cluster 1	**Cluster 2**	**Cluster 3**
Cluster proportions	0.3333	0.1709	0.4959

smaller differences in reality. Therefore, it can be hypothesized that the cluster with the longest lifetime has the lowest expected QLM. The variance characterizes how significant the differences are during discharge between the nominated cycles. Smaller fluctuations in the difference function suggest a longer expected lifetime. Additionally, it is logical to assume that a more rapid decrease in the initial phase of SCFC leads to earlier battery failure.

The MIR function suggests that a lower minimum internal resistance correlates with a longer battery lifetime. Therefore, IRD is less sensitive to survival time during the initial phase. The results also indicate in case of ACT5 variable, there

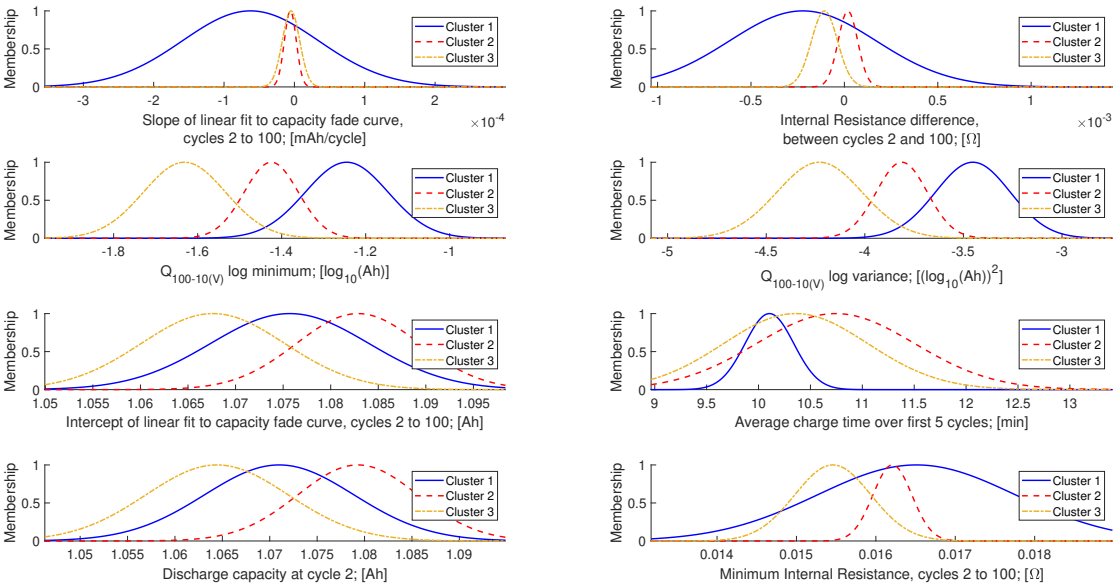

FIGURE 3.8: The Gaussian membership functions of the continuous variables. The variables can be categorized into those indicating the initial moments of the device, such as starting and production conditions (DC2, ICFC, ICFC), and those covering the initial operating period (SCFC, IRD, QLM, MIR, QLV). Variables related to the initial operating period provide insights into how batteries maintain their initial condition and deviate from it. The longer the initial state is maintained, the longer the useful life of batteries. In the case of ACT5, an optimal value of the average charge over the first five cycles is observed. Deviating from this value, both in positive and negative directions, adversely affects the survival. Based on the MIR function, lower minimum internal resistance is associated with longer battery lifetimes. However, drawing logical conclusions about the impact of ICDC and DC2, characterizing initial capacity, on survival time is not apparent.

is an optimal value for the average charge during the first five cycles, that maximizes the useful life. Deviating from this optimal value, both in the positive and negative directions, has adverse effects on the survival time. While ICDC and DC2 characterize the initial capacity, it is challenging to draw a logical conclusion about how they affect survival.

Data validation is achieved by comparing the resultant Weibull distribution for the training data with the empirical distribution for the validation data. The empirical distribution is estimated using the Kaplan-Meier method, which can be extended to handle the probabilities $p(j|t, \mathbf{x}^c, \mathbf{x}^d)$ as weights. The results are shown in Figure 3.9. The training and validation data are closely situated in Clusters 1 and 2. However, they deviate from each other in Cluster 3, likely due to the limited amount of data in the validation set for Cluster 3.

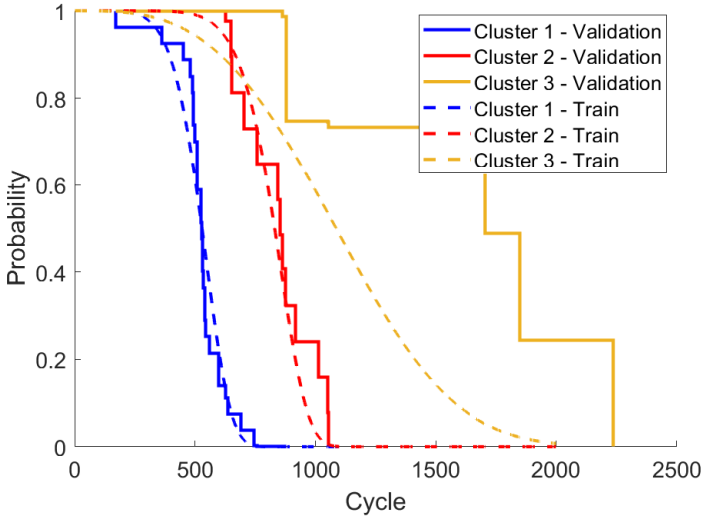

FIGURE 3.9: The validation of the results. The results are validated by comparing the resultant Weibull distribution for the training data with the empirical distribution of the validation data. It can be observed that the training and validation data are very similar in Cluster 1 and Cluster 2. The deviation in Cluster 3 is a result of very few data points in its validation set.

This section has demonstrated that the proposed method can be excellently used for calculating the remaining useful life, even when many variables are taken into consideration. The algorithm effectively identifies degradation processes that lead to failures. In terms of online applications, this method shows promise in detecting whether a battery belongs to a degradation cluster and predicting the remaining useful life.

3.3.3 Estimation of the survival of patents with prostate cancer

The applications of survival analysis are often found in healthcare case studies. The applicability of the proposed method is demonstrated using data concerning prostate-cancer patients [67]. These data were obtained from a randomized clinical trial comparing three types of treatments and placebo for patients suffering from prostate cancer. The trial was double-blinded and the treatments include placebos, 0.2 mg of diethylstilbestrol (DES), 1.0 mg of DES and 5.0 mg of DES, all administered orally on a daily basis. Although 12 variables were proposed in the database, only two continuous variables and one discrete variable are considered in the examination. The continuous variables are age (measured up until 89, so 89

denotes that the patient is at least 89 years old) and the concentration of serum hemoglobin in g/l. In addition, the type of treatment is considered as a discrete variable, which has a uniform distribution.

Although the database initially consisted of 506 patients, only 485 records were utilized for the analysis after removing 21 records due to missing data. In this study, survival times indicate how long a patient participates in the examination. The outcome variable can take on nine different values, where eight denote death due to different causes, and one signifies the recovery of the patient. Since the goal of the analysis is to showcase the applicability of the proposed method, estimating diverse outcomes is beyond the scope of the presentation. Therefore, the focus is on all-cause mortality. Based on the AIC, five local models accurately describe the survival times. The resulting distribution and Gaussian membership functions are considered in Figure 3.10 and 3.11, respectively. The values of the discrete probabilities are presented in Table 3.6 and the cluster proportions in Table 3.8.

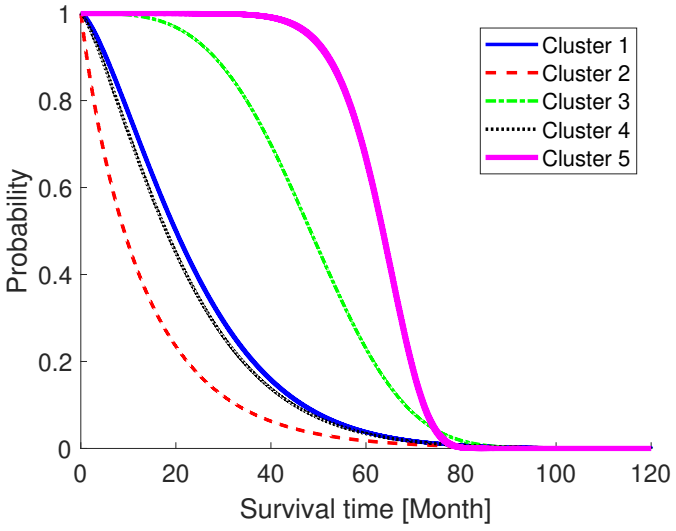

FIGURE 3.10: Survival time of the patients. Patients in Clusters 2 and 5 exhibit the shortest and longest survival, respectively.

The variable concerning the concentration of serum hemoglobin is exceptionally valuable in this study because it strongly and obviously influences the survival time. Inadequate hemoglobin concentration can have fatal implications. However, it is important to note that a low concentration of serum hemoglobin is not the direct cause of death, but an indicator of the patient's poor health. The results align well with this experience, and it is evident that a higher expected concentration of serum hemoglobin correlates with longer survival. An interesting observation arises in Clusters 3 and 5, hereafter referred to as the "serum hemoglobin

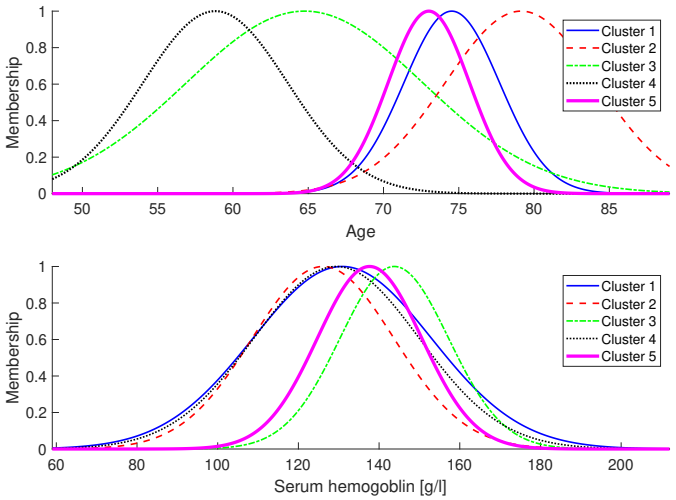

FIGURE 3.11: Gaussian membership functions of the continuous variables. The concentration of serum hemoglobin has trivial consequences. An inadequate concentration of hemoglobin can be fatal. It is observable that the higher the expected concentration of serum hemoglobin in a patient, the longer their survival tends to be. The exceptions to this rule are Clusters 3 and 5. The age of patients in their clusters do not appear to significantly affect survival times. After all, it is challenging to interpret the relationship between age and survival time based on these results.

TABLE 3.6: Values of discrete probabilities. Although, it can be observed that the applied treatments are mostly uniformly distributed over the different clusters. Cluster 2 is an exception in this regard, as the majority of patients were treated with 5 mg of DES. Patients in this cluster were likely in a critical condition and died the soonest. In Cluster 3, although the majority of patients were also treated with 5 mg of DES, the significance is not as pronounced as in Cluster 2. Following the relationship observed in Cluster 2, it suggests that a significant number of critically ill patients may have been present here, who died earlier.

Cure	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Shannon Entropy**
Placebo	0.3119	0.0515	0.2097	0.2418	0.2397	1.7202
0.2 mg of DES	0.2931	0	0.1140	0.5432	0.2936	1.7199
1.0 mg of DES	0.2656	0.0004	0.2760	0.2150	0.2652	1.7291
5.0 mg of DES	0.1294	0.9482	0.4002	0	0.2015	1.9525

phenomenon", where this line of reasoning appears contradictory. However, it is noticeable that the expected values do not differ significantly in these two clusters, and their standard deviations are nearly identical. While the lack of data may contribute to this phenomenon, it is also plausible that an unexplored factor might lead to the observed results.

The variable age does not exhibit a clear correlation with survival time. While age is generally an important factor that can indicate the death of a patient,

TABLE 3.7: The Weibull parameters. The parameter θ concerns the order with regard to the time of death. It can be assumed that patients in Cluster 2 have the shortest lifetime, while those in Cluster 3 can survive the longest.

Weibull parameter	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
θ	25.89	13.41	53.66	23.71	66.01
β	1.41	0.93	3.48	1.31	9.42

TABLE 3.8: The proportions of the clusters, which describe the proportion of the patients that belong to a given cluster. Cluster 1 dominates in terms of membership, to which almost 43 percent of the patients belong.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Cluster proportions	0.4331	0.0577	0.2216	0.0596	0.2280

several other factors can significantly influence the survival. Despite expectations based on medical studies, the concentration of serum hemoglobin contradicts what might be anticipated in terms of age. For instance, the highest concentration of serum hemoglobin is found in Cluster 3, where patients, despite being younger, have a shortest survival than those in Cluster 5. Obtaining more accurate and interpretable results would likely require more detailed explanatory variables.

Concerning the discrete variables, it is observed that the applied treatments are mostly uniformly distributed across the different clusters. An exception is Cluster 2, where the majority of patients are treated with 5 mg of DES. Patients in this cluster were likely in a critical condition and died the soonest. Treatment-related dominance was also observed for patients in Cluster 4, who had second fastest expected mortality, with the majority receiving the lowest treatment dose, which may have been ineffective. In Cluster 3, although the majority of patients were also treated with 5 mg of DES, this proportion is not as considerable as in Cluster 2. Following the pattern observed in Cluster 2, a significant number of critically ill patients who died earlier may have been situated in Cluster 3. This might be a possible explanation for the phenomenon concerning the concentration of serum hemoglobin.

The probabilities of discrete variables seem to be very uncertain for determining the cluster belonging. In order to further analyze the effect of dose on the cluster belonging, the base 2 conditional Shannon entropy is calculated for the probabilities $p(j|x_d^q = z)$ with the following Equation [68]:

$$H = \sum_{j=1}^{M} p(j|x_d^q = z) \log_2(p(j|x_d^q = z)) \tag{3.23}$$

where H denotes the Shannon entropy. Analyzing the results, Lower entropy values indicate less uncertainty, meaning that the treatment provides more information about the cluster membership. Higher entropy values indicate more uncertainty. Placebo, 0.2 and 1.0 mg of DES seem to have similar effects on clustering, as their conditional entropies are very close. On the other hand, 5.0 mg of DES stands out with a notably higher conditional entropy compared to the others. It suggests that this treatment is associated with more variability in cluster membership among patients.

In this case study, despite attempting to demonstrate how discrete variables could also be applied, the limited information content of the dataset only allowed for the trivial result that the concentration of serum hemoglobin is related to the chances of survival.

3.3.4 Analysis of the distribution of the COVID-19 mortality rate

The analysis aimed to highlight how the COVID-19 mortality rates varies in different countries according to explanatory variables related to the economic situation, urbanization, and the health condition of the citizens. The dataset has been compiled from several sources. The integrated dataset contains the number of death cases per 100K population and nine explanatory variables. The mortality rates were downloaded from the web page of Johns Hopkins University [69]. The explanatory variables are detailed in Table 3.9. All countries for which any of the variables were not available were removed from the analysis. Finally, data from $N = 117$ countries were studied.

The most important aspects in selecting the variables were their relevance and coverage (the number of countries where a given indicator is published). The selected variables mainly describe the health of the population, influenced by cultural effects and many other factors, including the state of the health system. To provide a comprehensive background of countries, variables related to the economic situation and urbanization were also included. From the initial dataset, some of the

Table 3.9: The economic, urban, and health condition explanatory variables, and their dates, time interval, and sources.

Sector	**Variable name**	**Time interval**	**Downloaded**	**Source**
Economic	GDP per capita (current US$)	01.01.2019 31.12.2019.	27.09.2021.	[70]
Health	Adolescent fertility rate (births per 1000 women ages 15–19)	01.01.2019 31.12.2019.	27.09.2021.	[70]
Economic	GHG emission/capita (CO_2 equivalent)	01.01.2018 31.12.2018.	27.09.2021.	[70]
Urban	Rural Population (% of population)	01.01.2020 31.12.2020.	27.09.2021.	[70]
Health	Diabetes prevalence (% of population ages 20–79)	01.01.2019 31.12.2019.	27.09.2021. 27.09.2021.	[70]
Health	Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age)	01.01.2018 31.12.2018.	27.09.2021.	[70]
Health	Life expectancy at birth (years)	01.01.2019 31.12.2019.	27.09.2021.	[70]
Health	Prevalence of current tobacco use (% of adults)	01.01.2018 31.12.2018.	27.09.2021.	[70]
Health	Obesity Rate (% of population)	01.01.2021 31.12.2021.	11.10.2021.	[71]

variables were eliminated as they were not significant in the studied models. The goodness of the feature selection was also confirmed by Cox regression, which will be presented in this section as well.

The application of the method requires selecting the number of clusters. In this study, three clusters were identified. The results of the clustering are depicted in Figure 3.12, where the identified Weibull survival functions are shown in the first subplot in the top row, along with the Gaussian membership functions of all the variables. The soft clustering results demonstrate that countries are highly likely (at least 0.99 probability) to belong to one of the identified clusters. Figure 3.13 depicts histograms of the variables corresponding to the clusters, revealing a similarity in shape between the histograms and Gaussian membership functions. The frequency of these occurrences also serves as an indicator of variable importance and goodness of fit.

The first subfigure of Figure 3.12 presents the resultant Weibull survival functions, highlighting substantial differences in the distributions of deaths caused by COVID-19 across the identified three country clusters. In Cluster 1, the probability that the number of death cases is beyond 100 per 100K population is nearly 0, indicating that members of this cluster have the best chances of surviving the

pandemic. Conversely, in Cluster 3, this probability is 0.55, suggesting that citizens in countries assigned to Cluster 3 have the worst chances of surviving the COVID-19 pandemic.

The comparison of the Gaussian membership functions of the clusters can provide insights into the key differences between countries that significantly affect the mortality chances of COVID-19. As shown in Figure 3.12, Cluster 1 exhibits the smallest number of smokers, while Cluster 3 has the highest number of smokers. A similar pattern is observed in the incidence of diabetes. These results are logical, as both smoking [72] and diabetes [73] are known to be associated with increased mortality of COVID-19.

The clusters exhibit wide and overlapping distributions for variables such as alcohol consumption, GHG emission per capita, and obesity rate, indicating that these factors may not distinctly characterize the death cases. A notable driver of cluster segmentation is the GDP per capita variable. Surprisingly, countries with low GDP fall into Cluster 1, while prosperous nations are divided between Cluster 2 and 3. Since life expectancy correlates with GDP per capita [74], wealthier countries tend to have more older citizens. Given that COVID-19 poses a higher risk to the senior population [75], it implies that fewer people die in countries with a lower proportion of older adults. However, this statement holds true only up to a certain level of wealth. The wealthiest countries have made greater efforts and possess more advanced healthcare systems, allowing them to protect their older population more efficiently.

The method can explore correlating variables. The Gaussian membership functions of the GHG emission per capita and GDP per capita behave similarly, reflecting correlation. This result aligns with the Kuznets theory, which suggests that countries emit more GHG gas as GDP per capita grows. However, after reaching a certain level of development, countries often have the resources to invest in technologies that reduce emissions [76], indicating a turning point in the Kuznets function. According to the Gaussian membership functions, emissions are lower in poorer countries and higher in richer ones.

Moreover, the variance of GDP per capita is much more comprehensive for the most prosperous countries because some countries have already passed the turning point. Recent studies suggest a link between air pollution and the COVID-19

mortality rate [77]. Given the connection between GDP and GHG, it becomes apparent that there is a minor association between air pollution and the COVID-19 mortality rate, acting as an indirect relationship, as both variables are influenced by GDP. The results illustrate that GDP per capita is the most significant factor in the modeling problem. Moreover, GDP per capita correlates with urbanization, as wealthier countries tend to have higher levels of urbanization and more centralized populations. This alignment suggests that the richer a country is, the more urbanized and centralized its population, potentially leading to easier and more direct virus spread [78].

Furthermore, wealthier countries tend to have a higher prevalence of overweight individuals due to better access to food resources, and these citizens carry a greater risk of infection [79]. Additionally, alcohol consumption is generally higher in wealthier countries, potentially weakening the immune systems of individuals and contributing to a higher risk factor [80]. The geographical distribution of the countries of different clusters are shown in Figure 3.14. Moreover, the COVID mortality rates per 100K population are also provided for certain countries to underscore their range within the specific clusters.

The proposed method has been compared to semi-parametric Cox regression, a widely applied technique for survival analysis. The resultant Cox model was evaluated at the cluster means, and the extracted distributions were compared to the local models identified by the proposed method. As shown in Figure 3.15, the local distributions are nearly identical to the distributions estimated by the Cox regression model.

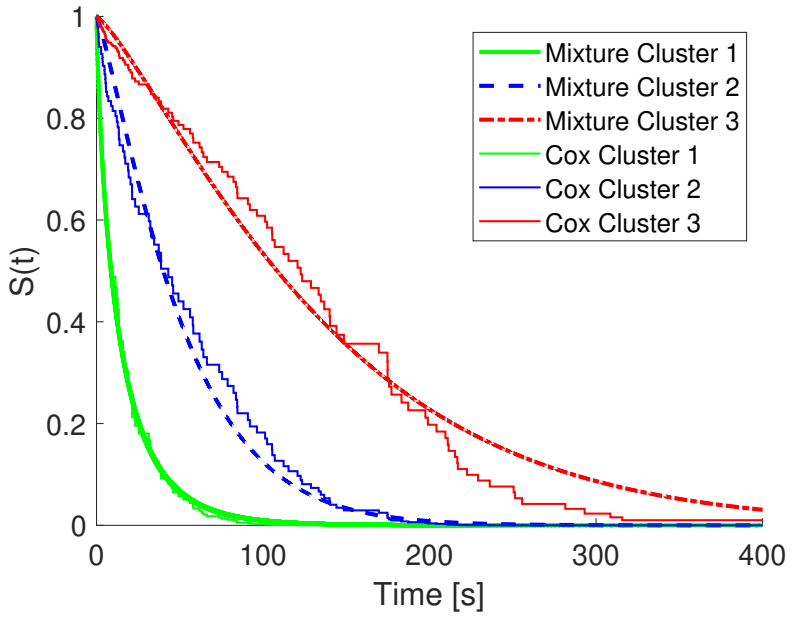

FIGURE 3.15: The comparison of the proposed methodology and the Cox regression. The survival functions are calculated with the Cox regression method at the point of the cluster means, and they are compared with the resultant distributions by the proposed method. The proposed method describes the distribution the same way as the Cox regression. The contribution of the variables can be measured by the Cox regression parameters.

This section has demonstrated that the proposed method can be excellently used for calculating the probability distribution of mortality rates per 100K population of countries related to the COVID-19 pandemic. The algorithm effectively identifies which factors causing an increased mortality rate.

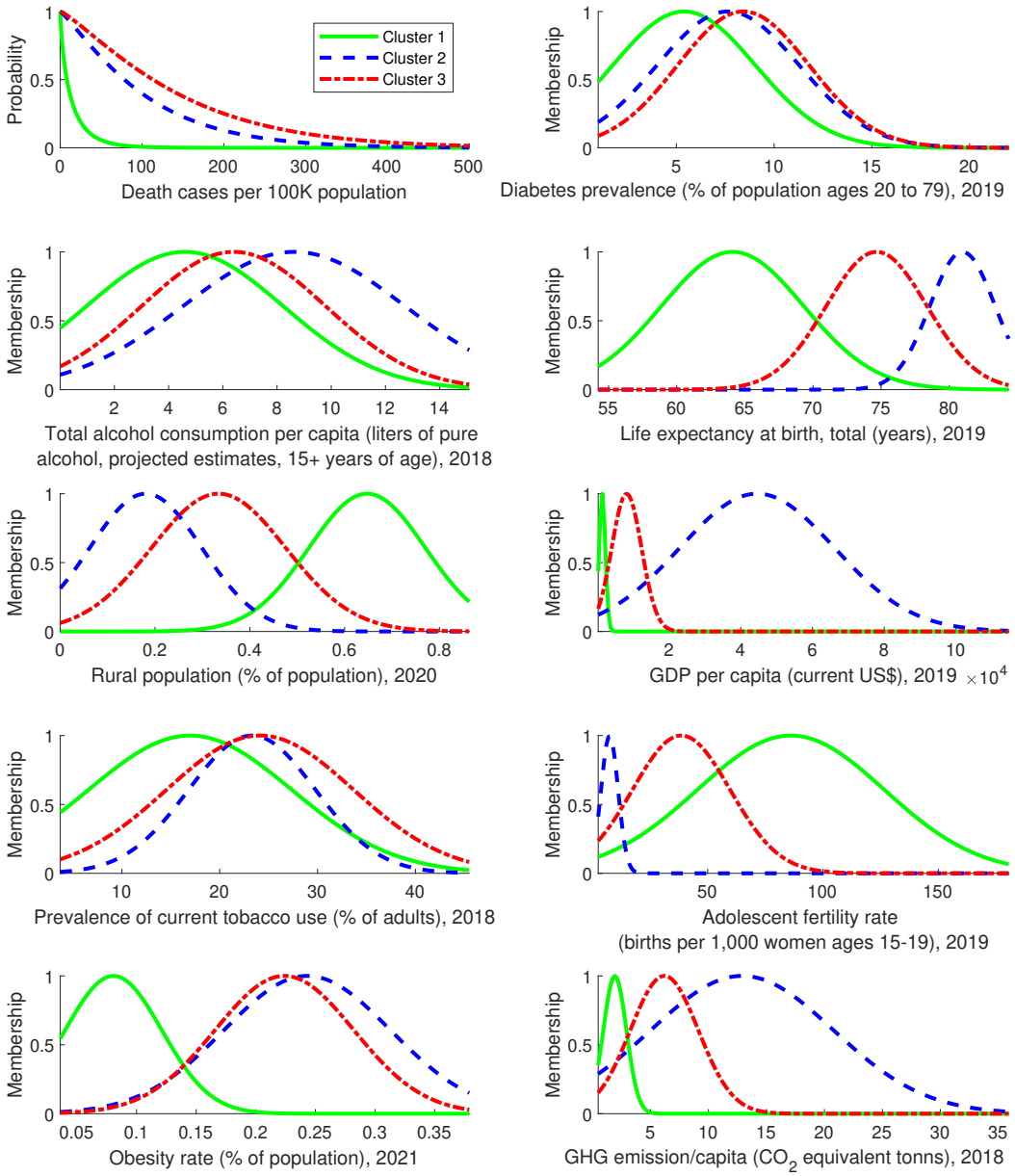

FIGURE 3.12: The left upper subfigure shows the resultant Weibull survival functions. The distributions of deaths caused by COVID-19 significantly differ in the identified three groups of countries. While Cluster 1 shows a nearly 0 probability that the number of death cases is beyond 100 per 100K population, Cluster 3 exhibits a substantially higher probability of 0.55%. This distinction underscores that members of Cluster 1 have notably better chances of surviving the pandemic compared to those in Cluster 3. Other subfigures depict the Gaussian membership functions that describe the operating regions of each variable. The main driving force of the analysis is the GDP per capita. The method reflects correlating variables. Life expectancy correlates with the GDP per capita. COVID-19 is riskier for the older population, and this implies that fewer people die in countries where fewer older adults live.

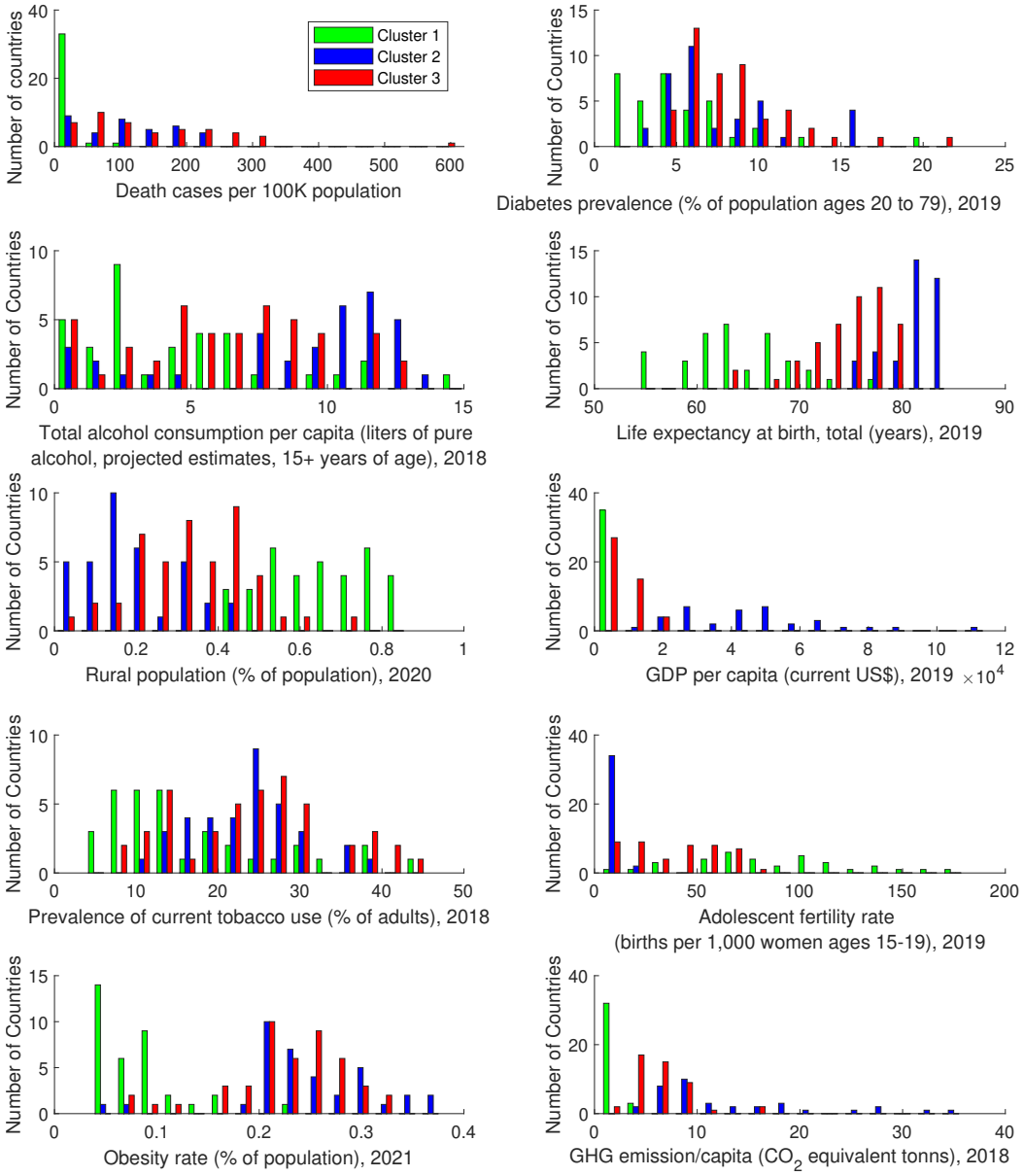

FIGURE 3.13: The histogram of the variables according to the clusters. It can be observed that the shape of the histograms and Gaussian membership functions are analogous. However, the number of these incidences can also indicate the weight of the variable and the goodness of fit.

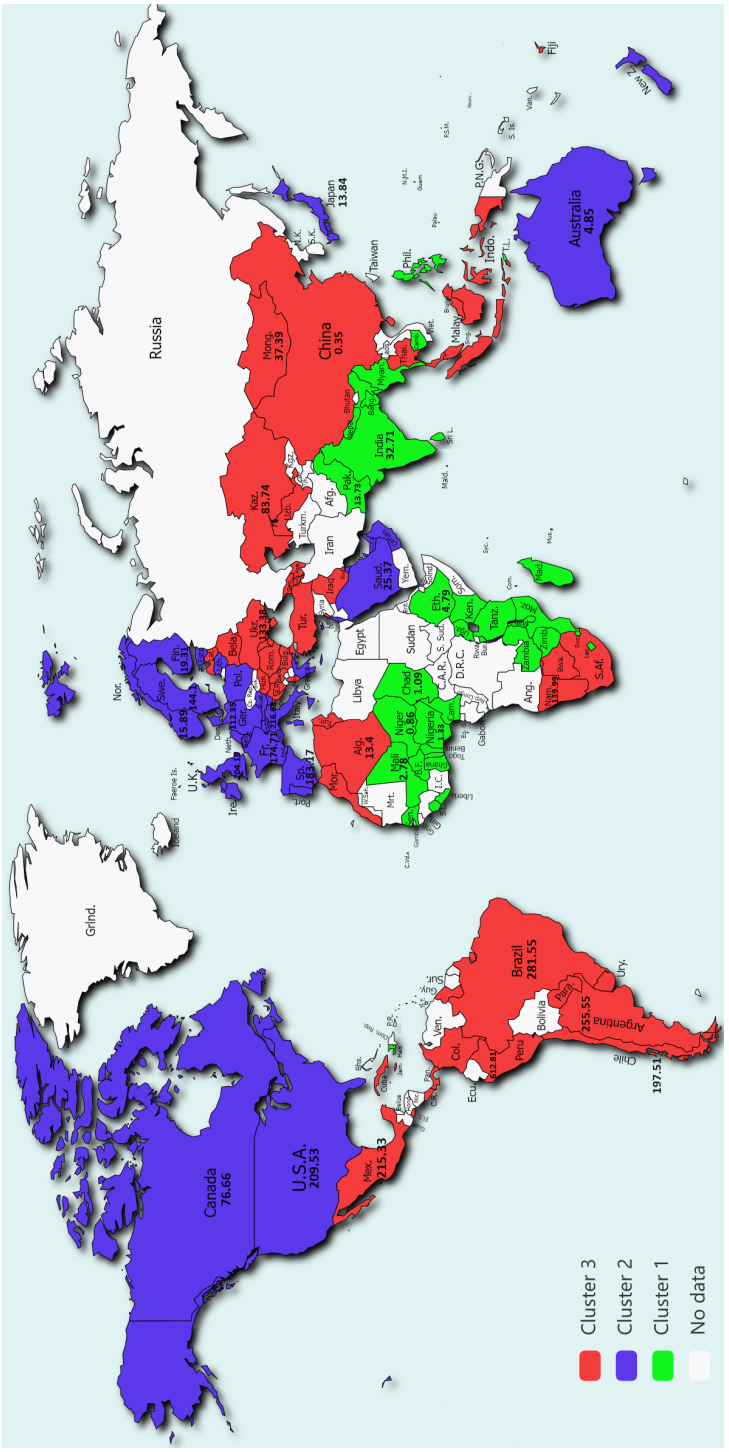

FIGURE 3.14: The cluster membership of the countries. It can be observed that the clustering was indeed based on whether they were developing or developed countries. Moreover, the COVID mortality rates per 100K population are also provided for certain countries to underscore their range within the specific clusters.

3.4 Summary of the chapter

In this chapter, a novel integrated survival analysis and expectation-maximization-based clustering method was introduced. The method focused on separating datasets based on the similarity of survival times and explanatory variables, aiming to group heterogeneous survival models into homogeneous ones. The approach utilized the Weibull distribution and multivariate Gaussian distributions to define clusters and simultaneously identified their survival probabilities. The cluster memberships were represented using Takagi-Sugeno fuzzy rules, offering a framework to determine the operating domain of continuous variables. This approach allowed for describing survival characteristics by considering both the domain of continuous variables and the proportions of discrete variables. The method demonstrated versatility and is applied for categorizing continuous variables. The determination of the number of clusters was achieved by employing the Akaike Information Criterion.

The effectiveness of the proposed algorithm was demonstrated across diverse case studies, showcasing its applicability across various domains. The student dropout rate was estimated based on the average grade during the first semester and enrollment scores. After some refinements and incorporation of additional approaches, the model could be developed into a personalized forecasting system that alerts students if they become "at-risk" during their studies. Additionally, the remaining useful life of Li-ion batteries was estimated based on their capacity, internal resistance, and charging condition. The algorithm effectively identified degradation processes that led to failures. In terms of online applications, this method proved to be promising in detecting whether a battery belongs to a degradation cluster and predicting the remaining useful life. Furthermore, survival chances for patients with prostate cancer were determined based on their age, serum hemoglobin level, and treatment. Despite attempting to demonstrate how discrete variables were applied, the limited information content of the dataset only allowed for the trivial result that the concentration of serum hemoglobin is related to the chances of survival. Finally, the mortality rate per 100K population of countries related to the COVID-19 pandemic is estimated based on demographical and economical data. This section demonstrated that the proposed method can be excellently used for calculating mortality rates per 100K population of countries related to the COVID-19 pandemic. The algorithm effectively identified which factors caused an increased mortality rate.

Chapter 4

Integrated survival analysis and frequent itemset-based association rule mining: a course failure-based prediction of student dropout

This chapter introduces an integrated survival analysis and frequent itemset-based association rule mining that identifies relevant triggering events defined from time-dependent categorical variables, that lead to consequent events of competing risks. Section 4.1 provides a clear overview of the challenges the algorithm aims to address and briefly introduces the algorithm. Section 4.2 presents a detailed mathematical description of the proposed algorithm. Section 4.3 demonstrates the practical application of the algorithm through the estimation of student dropout rates.

4.1 Introduction

In traditional survival analysis, the primary focus revolves around the time until a singular event occurs. However, real-world scenarios frequently involve a spectrum of possible outcomes, referred to as competing risks [81]. Competing risks arise when multiple distinct outcome can potentially characterize the survival, and the occurrence of one event precludes the occurrence of others. To illustrate, consider healthcare studies, where the survival time often denotes the duration until a subject leaves the examination. Reasons for leaving can be attributed to various competing events, such as cardiac death or non-cardiac death [82]. Similarly, in educational contexts, the survival time might represent the duration until the student is present at the university, and their outcomes could include graduation or dropout. Despite the prevalence of competing risks in diverse fields, there remains a gap in their comprehensive integration within analytical frameworks. While the concept of competing risks is acknowledged in research [83], its consistent application is often overlooked [84]. Failing to account for competing risks leads to biased estimates.

A conventional but naive approach to addressing competing risks involves applying the Kaplan-Meier estimator while censoring samples related to competing events. However, this simplistic method violates the independent censoring assumption, leading to biased estimates [13]. In the present of competing risks, the survival function is inadequate and a so-called cumulative incidence function needs to be given instead. The estimation of the cumulative incidence function is based on two distinct approaches: the cause-specific hazard and the sub-distribution hazard model [85] that are introduced in Equations 4.3, 4.6, respectively. The estimation of the cumulative incidence function based on cause-specific hazard model requires both the all-cause survival function and the hazard function of the anticipated outcome with censored competing events [86]. Conversely, the sub-distribution hazard model takes a different perspective by assuming that subjects who experience competing events are destined to face an infinite risk. In this scenario, subjects with competing events are not eliminated from the risk set while estimating the outcome [87].

Survival models commonly incorporate explanatory variables that stratify subgroups, that have distinct survival probabilities and also cumulative incidences [86]. The straightforward approaches for handling explanatory variables have already

been presented in the previous chapter, without considering competing risks. However, in the presence of competing risks, managing explanatory variables becomes particularly challenging. Parametric regression models such as Cox regression were designed for this purpose and have become the predominant way to address these variables [88]. In this context, cause-specific hazard models lose their ability to interpret survival functions because they lack a direct connection to competing events [89]. To overcome this issue, the Fine-Gray sub-distribution regression model was designed, based on sub-distribution hazard models.

Explanatory variables in survival models are typically time-independent. However, scenarios exists wherein the explanatory variables depend on time, which can be further categorized into external and internal time-dependent variables [90]. External variables remain unaffected by the failure mechanism, maintaining their independence from its occurrence [91]. On the other hand, internal variables dynamically respond to the failure mechanism, encapsulating factors directly influenced by the underlying failure conditions. The Fine-Gray sub-distribution hazard regression model has limitation related to the nature of explanatory variables [30]. In the case of external time-dependent variables, interpreting the resultant model parameters requires careful consideration [92], but the cumulative incidence function can be interpreted. In the case of internal time-dependent variables, the cumulative incidence function cannot be estimated. Moreover, the model also needs to meet the proportional hazard assumption [93].

In addressing limitations associated with time-dependent covariates in competing risks, this chapter introduces an integrated frequent itemset-based association rule mining and survival analysis framework that converts the relevant characteristics of time-dependent variables into a global, time-independent variable. The proposed method operates within a discrete time domain [94], where survival is characterized by categorical explanatory variables at every time instance. The time-dependent categorical variables are conceptualized as triggering events, acting as precursors, initiating consequent events that signify the competing outcomes of the survival process. In conclusion, events in general are composite structures, comprising both triggering and consequential elements. Frequent itemset mining is employed to identify relevant co-occurrences of events at each discrete time instance, and association rules are formulated to discern those triggering events leading to specific consequent events. The approach segments the dataset for subjects characterized by a sequence of frequent itemsets consisting of specific combinations of triggering

events. Consequently, individuals with these characteristics can be identified by a global, single-variable, time-independent feature. The population, characterized by this independent feature, exclusively exhibits those time-dependent characteristics, and the resulting cumulative incidence function precisely describes this combination of variables.

The values of the cumulative incidence function for the specific outcome of interest can be directly determined within these groups, each having a specific sequence of frequent itemsets. This is achieved by modifying supports and confidences according to the segmented database [95], emphasizing frequent itemsets that also include the designated consequent event. By utilizing these confidences, the cumulative incidence function can be determined directly based on the rules [96]. This involves cumulating the confidences in a manner similar to the manual approach in the sub-distribution hazard model. Alternatively, the application of the sub-distribution hazard model is feasible by stratifying the population. Moreover, the Fine-Gray hazard regression model can also be applied. These, along with the direct determination of the cumulative incidence function based on the association rules, provide three independent approaches for comparing and validating the results.

The developed method is demonstrated based on the data of 348 students in the chemical engineering undergraduate program at the University of Pannonia in Hungary. Student dropout is a pervasive issue with significant financial and prestige implications in various countries, such as the United States [97] and Chile [98]. Since, dropping out not only affects the student but also the university and its community [99], all three parties should pay close attention. Numerous research aspects explore why students become at-risk or drop out of university. Previous studies have primarily focused on estimates based on high school performance [100]. Factors of interest include background information combined with semester performance [101], financial considerations [102], and family background [103]. In addition to these factors, there are universal aspects in the studies, such as demographics and personal characteristics [104].

The impact of artificial intelligence solutions on educational research has already been demonstrated [105]. However, it remains a relatively novel and unfamiliar area for many researchers and educators, presenting unsolved challenges [106]. One notable deficiency in current studies is the predominant focus on early prediction methodologies. Consequently, a future challenge lies in developing a robust and

comprehensive early warning system capable of effectively predicting and identifying 'at-risk' students in the distant future [107]. While traditional machine learning algorithms, such as clustering and classification, are adept at partitioning students, they fall short in providing a predictive model [108]. On the other hand, certain machine learning tools excel in determining the outcome of a given semester [109]. Survival analysis proves valuable for predicting the probability of dropout over several semesters [20]. This framework has demonstrated utility in predicting the success of online education [110] and detecting dropout factors [111]. The proposed method aims to integrate the principles of the previously mentioned solutions, leveraging their advantages to offer a compact and comprehensive solution.

Due to the complex nature of the field, it becomes imperative to formulate specific objectives and methodologically develop strategies [112]. Thus, beyond algorithm development, this chapter endeavors to introduce novel factors capable of predicting student dropout. A thorough literature review revealed a gap in existing studies, particularly in the absence of estimations based on patterns of uncompleted subjects of students. Consequently, this research aims to identify a model capable of discerning regularities in frequently uncompleted subjects using available performance data. The potential outcomes in this analysis encompass either graduation or dropout, treated as competing events. Notably, the study not only unveils frequently occurring association rules predicting student dropout through association rule mining algorithms by directly estimating the cumulative incidence function but also provides the associated Kaplan-Meier estimate of the empirical distribution of dropout times based on the sub-distribution hazard model. Furthermore, the research demonstrates the identification and application of the Fine-Gray sub-distribution regression model. The specific contributions of the chapter can be outlined as follows:

- The probability of competing risks can be determined at specific time instances using event-driven frequent itemset-based association rules. This approach identifies relevant triggering events defined from time-dependent categorical variables, that lead to consequent events defined from competing risks. A sequence of frequent itemsets can be represented as a global, time-independent feature.
- The cumulative incidence function of a specific competing risk can be directly determined for the population with a given sequence of frequent itemsets.

The method segments the dataset for subjects that supports all the frequent itemset of the selected sequence and estimates the cumulative incidence function based on the modified rule supports and confidences.

- The student dropout rate can be estimated based on patterns of uncompleted subjects. The study has a sample curriculum that prescribes the recommended semester for each subject completion. Inability to meet this requirement marks an uncompleted subject event, a crucial factor associated with subsequent student dropout.

The next section presents the mathematical description of the proposed methodology.

4.2 Integration of frequent itemset-based association rule mining and survival analysis

The proposed method integrates frequent itemset-based association rules and survival analysis, creating a comprehensive framework to estimate the probability of competing risks, as illustrated in Figure 4.1. The method commences with the integration of diverse data sources essential for event identification. This process also divides events into triggering and consequential ones. The next step involves employing frequent itemset mining to identify relevant events at each temporal instances. Subsequently, association rules are defined to identify sets of triggering events, revealing how specific consequential events unfold. The resulting supports and confidences are then utilized to estimate the cumulative incidence function of a designated competing event. When focusing solely on consequence events, the baseline cumulative incidence function can be estimated. However, the cumulative incidence function can be delineated by specific triggering events that lead to a consequent event. To estimate cumulative incidence functions for subjects displaying distinctive characterizing triggering events, the dataset is segmented according to the supported subjects of the selected frequent itemsets sequence. This segmentation includes only subjects that support all the selected frequent itemsets, and the cumulative incidence function is subsequently estimated based on this specific set.

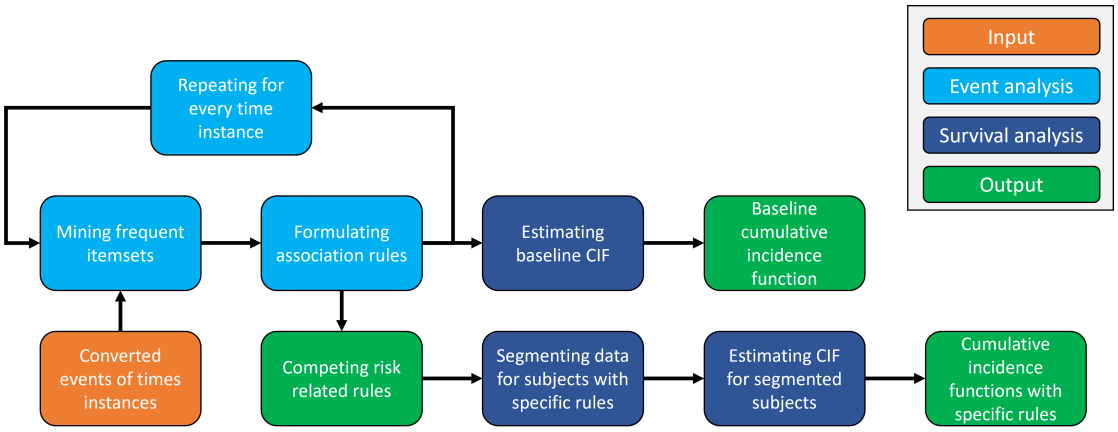

FIGURE 4.1: The steps of the proposed method. The method starts with event definition. The next step utilizes frequent itemset mining to identify events at each time instances. Association rules are defined to identify sets of triggering events that lead to consequence events. The resulting supports and confidences are then utilized to estimate the Cumulative Incidence Function (CIF).

The following subsections provide the details of the method. In Subsection 4.2.1, the estimation of the cumulative incidence function for competing risks is explained. Subsection 4.2.2 introduces the concept of frequent itemset-based association rules. In Subsection 4.2.3, the estimation of the baseline cumulative incidence function is presented, and finally, Subsection 4.2.4 details the estimation of the cumulative incidence function for specific sequences of frequent itemsets.

4.2.1 Empirical cumulative incidence functions for competing risks

The proposed method studies the non-parametric empirical distribution of the occurrence of events in ordered discrete occurrence times: $\mathbf{t} = [t_1, \ldots, t_k, \ldots, t_N]$. The $S(t_k)$ survival function represents the probability that an event occurs later than the discrete time instance t_k:

$$S(t_k) = P(T > t_k) \tag{4.1}$$

The proposed comprehensive analytical framework considers competing risks that arise when multiple distinct outcomes influence the survival process. In the presence of competing risks, the estimated survival function by Kaplan-Meier method

introduces bias, necessitating the provision of a cumulative incidence function. This marginal probability distribution for a given outcome can be given based on two different hazard models: the cause-specific and sub-distribution hazard model. The cause-specific hazard model censors the samples of competing events, and the instantaneous hazard of the interested outcome c can be described by the next equation [86]:

$$h^c_{CS}(t_k) = \lim_{\Delta t_k \to 0} \frac{P\big(t_k \leq T < t_k + \Delta t_k, cause = c | T > t_k\big)}{\Delta t_k} \tag{4.2}$$

From this hazard model, the cumulative incidence function for outcome c can be expressed as follows [113]:

$$CIC^c_{CS}(t_k) = \sum_{f=1}^{k} h^c_{CS}(t_f) S(t_{f-1}) \tag{4.3}$$

On the other hand, the sub-distribution hazard model assumes that samples of competing events remain indefinitely in the risk set. The instantaneous hazard of the designated outcome c can be described by the following equation [87]:

$$h^c_{SD}(t_k) = \lim_{\Delta t_k \to 0} \frac{P\big(t_k \leq T < t_k + \Delta t_k, cause = c | T > t_k \cup \{T \leq t_k, cause \neq c\}\big)}{\Delta t_k} \tag{4.4}$$

Utilizing this hazard model, the Kaplan-Meier estimator can be employed to estimate the cumulative incidence function, accounting for the time-to-event data of competing events occurring in infinity [94]:

$$\phi_i = \begin{cases} T_i, & \text{if } cause = c \\ \infty, & \text{if } cause \neq c \end{cases} \tag{4.5}$$

After this modification, the cumulative incidence function for cause c can be expressed as follows:

$$CIC^c_{SD}(t_k) = 1 - P(\phi > t_k) = 1 - \prod_{f=0}^{k} \left(1 - \frac{m_f}{N^*_f}\right) \tag{4.6}$$

where, N_f^* is the corrected number of risk set at time instance f. One of the significant advantages of the presented empirical distribution is its straightforward application complexity, even when managing competing risks. Both cause-specific and sub-distribution hazard models can be utilized to estimate the cumulative incidence function of an interested outcome in the absence of explanatory variables. However, a drawback of this method is its treatment of the entire dataset without providing additional information on individual cases, such as the impact of different triggering events. In scenarios where various causes may lead to an outcome, it is advisable to explore the influence of sets of possible causes and their contribution to the risk of an outcome. The following subsection details how frequent itemsets of events and association rules can be explored to understand these contributing factors.

4.2.2 Frequent itemset-based association rule mining

Similarly to the survival times, the studied events can occur at discrete time instances $\mathbf{t} = [t_1, \ldots, t_k, \ldots, t_N]$. Consider a set of distinct events denoted as $E = \{e_1, \ldots, e_M\}$, where M denotes the number of unique events. These unique events can be structured within cases at every time instant, forming an unordered set of events: $\mathcal{A}_i(k) = \{a_{i,1}(k), \ldots, a_{i,j}(k), \ldots, a_{i,N_{i,k}}(k)\}$, where $a_{i,j}(k)$ denotes the occurrence of the jth event in the ith case at the kth time instant and $a_{i,j}(k) \in E$. In this analysis, a case is the set of uncompleted subjects of a specific student. The collection $\mathcal{A}(k) = \{\mathcal{A}_1(k), \ldots, \mathcal{A}_i(k), \ldots, \mathcal{A}_{N_k}(k)\}$ of these cases serves as the input of frequent itemset mining at the kth time instance, where N_k is the number of cases at the kth time instance.

The purpose of this frequent itemset mining problem is to select $X_p(k)$ set of relevant event patterns (itemsets) at the kth time instant. These event patterns are further structured into a collection $X(k) = \{X_1(k), \ldots, X_p(k), \ldots, X_{P_k}(k)\}$ of frequent sequences, where P_k is the number of frequent itemsets at the kth time instance. This collection serves as the output of frequent itemset mining at the kth time instance. The relevancy of an itemset $X_p(k)$ is determined based on its frequency in $\mathcal{A}(k)$. The relative number of cases in which the itemset $X_p(k)$ occurs in $\mathcal{A}(k)$ is called support denoted as $supp\big(X_p(k)\big)$ and can be expressed by the next equation [114]:

$$supp\big(X_p(k)\big) = \frac{|\{\xi \in \mathcal{A}(k) : X_p(k) \subseteq \xi\}|}{N_k} \tag{4.7}$$

The itemset $X_p(k)$ is considered frequent, if its support exceeds a specific threshold value: $supp(X_p(k)) \geq minsup$. A higher $minsup$ value leads to a smaller number of identified frequent itemsets, while a smaller $minsup$ value yields more itemsets. Consequently, decreasing the $minsup$ value results in increased computational complexity, which can be quite significant. Generally, frequent itemset mining algorithms aim to discover all frequent itemsets. However, alternative approaches exist that limit the resulting outputs by disregarding those whose information is somehow involved in another frequent itemset. These alternative approaches include mining top-k, closed, or maximal frequent itemsets, which can help mitigate the algorithm's complexity without significant loss of information when using a low $minsup$ value [115]. These techniques eliminate irrelevant rules based on different principles, and the approach selection is highly case-study specific.

To involve competing risks into the analysis, the goal is to discover frequent itemsets that can be grouped into a set of triggering events and a consequence event as follows: $X_p(k) = \{X_{p*}(k), e_c(k)\}$. In this structure, the antecedent part of the $X_{p*}(k) \Rightarrow e_c(k)$ association rule is the $X_{p*}(k)$ set of triggering events and the $e_c(k)$ consequence part is the triggered consequence event. The length of the association rule $X_p(k)$ is defined as the number of items in $X_{p*}(k)$ and denoted as $L_{X_p} = |X_{p*}(k)|$.

The confidence of the $X_{p*}(k) \Rightarrow e_c(k)$ association rule is represented by the $P(e_c(k)|X_{p*}(k))$ conditional probability, that describes the probability that the $X_{p*}(k)$ set of triggering events causes the $e_c(k)$ consequential event:

$$conf\big(X_{p*}(k) \Rightarrow e_c(k)\big) = P\big(e_c(k)|X_{p*}(k)\big) = \frac{supp\big(X_p(k)\big)}{supp\big(X_{p*}(k)\big)} \tag{4.8}$$

Given the support and confidence measures of the association rules, the probability of the consequential events can be calculated, as detailed in the following subsection.

4.2.3 Determining the marginal probability of competing events based on event analysis

This section explores the idea that the cumulative incidence function for a designated outcome can be directly estimated based on the identified patterns. The function is determined based on the support and confidence measures derived from the frequent itemsets and association rules [95]. The estimation of the cumulative incidence function based on the cause-specific hazard model requires both the all-cause survival function and the hazard function of the anticipated outcome, factoring in the censoring of competing events. The all-cause survival function can be given as:

$$S(t_k) = \prod_{f=1}^{k} \left(1 - \sum_{q=1}^{C} supp\big(e_q(f)\big) \right) \tag{4.9}$$

Frequent itemset mining at every time instances manually do the modification of risk-set comes from censoring. Therefore, using the 4.3 and the 4.9 Equations, the following formula can be given

$$CIC^{c}_{CS}(t_k) = \sum_{f=1}^{k} e_c(f) \prod_{f'=1}^{f-1} \left(1 - \sum_{q=1}^{C} supp\big(e_q(f')\big) \right) \tag{4.10}$$

The estimation of the cumulative incidence function based on subdistribution hazard model requires the modification of the risk set so that subject died by competing risks are not excluded from the risk set.

$$CIC^{c}_{SD}(t_k) = \prod_{f=1}^{k} \left(1 - \frac{supp\big(e_c(f)\big) N_f}{N_f + \sum_{f'=1}^{f-1} \sum_{q=1,\ q \neq c}^{C} supp\big(e_q(f')\big)} \right) \tag{4.11}$$

The next subsection presents, how the cumulative incidence function characterised by a specific sequence of frequent itemsets can be calculated that leads to a consequent event, thereby advancing the method towards a predictive framework.

4.2.4 Estimating the cumulative incidence functions with specific patterns based on event analysis

In the previous section, the estimation of the baseline cumulative incidence function of a specific consequent event is presented. This section presents, how the cumulative incidence function characterized by specific triggering events can be calculated that leads to a consequent event, thereby advancing the method towards a predictive framework. The first step in this extension is to formulate a sequence of frequent itemsets at the interested time instant, denoted as $\langle X_{p_1}(1) \rightarrow \ldots \rightarrow X_{p_k}(k) \rightarrow \ldots \rightarrow X_{p_N}(N)\rangle$. These itemsets represent informative patterns of events associated with the survival process.

To estimate cumulative incidence functions for subjects exhibiting specific characterizing triggering events, the dataset is segmented based on the supported subjects of selected frequent itemsets sequence. Only subjects, that supports all the selected frequent itemsets are included in this segmentation. Let $\mathcal{D}$ represent the dataset comprising subjects that exhibit the identified sequence of frequent itemsets. With the segmented dataset $\mathcal{D}$, the cumulative incidence functions can be estimated using the subdistribution hazard model as follows:

$$CIC^{c}_{SD}\big(t_k, X_{p_1}(1) \rightarrow \ldots \rightarrow X_{p_k}(k)\big) = \prod_{f=1}^{k}\left(1 - \frac{supp\big(X_{p_f}(f) \Rightarrow e_c(f)\big)N_{f,\mathcal{D}}}{N_{f,\mathcal{D}} + \sum_{f'=1}^{f-1}\sum_{q=1,\ q\neq c}^{C} supp\big(X_{p_{f'}}(f') \Rightarrow e_q(f')\big)}\right) \tag{4.12}$$

where $N_{k,\mathcal{D}}$ is the number of subjects at kth time instance in the segmented dataset $\mathcal{D}$.

This methodology opens the door to developing a predictive system. By estimating cumulative incidence functions for subjects with specific characterizing triggering events, the model can provide insights into the chances of a consequential event given that the subject has experienced the triggering events at the given time instances. This predictive capability enhances the practical utility of the integrated method in understanding and forecasting event occurrences in complex systems such as student dropout.

4.3 Application to student dropout prediction

This section outlines the practical application of the proposed method in analyzing course completion data relating to former chemical engineering students at the University of Pannonia. Specifically, the focus is on students who have either successfully graduated or been expelled from the university. The study excludes both active and passive students due to the lack of available outcome information. Additionally, re-applied students are excluded from the analysis. The data used is entirely de-identified. The input for the method is constructed through the integration of student log files and a sample curriculum. The recorded data span from 2011 to 2018, encompassing 348 students. During data processing, students were carefully excluded who had applied and dropped out before 2011, as their re-application after this period could introduce confounding factors, such as students seemingly graduating prematurely for reasons that are not readily understandable. Formulating patterns for each case of uncompleted subject failure proved to be a particularly challenging aspect of the analysis.

The subsequent subsections illustrate the application of the proposed method. In Subsection 4.3.1, the dataset under analysis is described. Subsection 4.3.2 delves into the survival analysis with competing risks, while Subsection 4.3.3 demonstrates the analysis of frequent itemset-based association rules. In Subsection 4.3.4, the estimation of the baseline cumulative incidence function is presented. Finally, Subsection 4.3.5 details the estimation of the cumulative incidence function for specific sequences of frequent itemsets.

4.3.1 The description of the analyzed dataset of course completions

The integrated student log file consists of two components. The student database records each attempt to complete a subject as an elementary event. Additionally, a binary variable is utilized to depict graduation or unsuccessful completion (dropout). Through the combination of these components with information extracted from the sample curriculum, an integrated student log file is generated. A sample of this log file is illustrated in Table 4.1. Based on the integrated student log file,

the empirical survival function can be specified by both the cause-specific and sub-distribution hazard models. However, for more complex event analysis, conversion steps must be included.

TABLE 4.1: A sample for the student log file, which integrates the student-specific data and the sample curriculum. For instance, in the first semester, student No. 3 did not complete subject ID 4, whereas in the second semester, he/she successfully completed the subject.

Student ID	**Educational status**	**Subject ID**	**Subject status**	**Attempted semester**	**Recommended semester**
1	graduated	1	completed	1	1
1	graduated	2	failed	2	2
1	graduated	3	failed	2	2
1	graduated	2	failed	3	2
1	graduated	3	completed	3	2
1	graduated	2	completed	4	2
2	dropped out	1	failed	1	1
2	dropped out	4	failed	1	1
2	dropped out	5	failed	1	1
2	dropped out	6	failed	1	1
2	dropped out	2	failed	2	2
2	dropped out	3	failed	2	2
3	graduated	4	failed	1	1
3	graduated	2	completed	2	2
3	graduated	3	failed	2	2
3	graduated	4	completed	2	1
3	graduated	3	completed	3	2
3	graduated	7	failed	3	3
3	graduated	7	failed	4	3
3	graduated	7	failed	5	3
3	graduated	7	completed	6	3

Student subject failures are portrayed as events, illustrated in the Gantt chart in Figure 4.2. Let $\hat{\omega}_j$ denote the semester in which a student is expected to complete the jth subject according to the sample curriculum, and $\omega_{i,j}$ represents the semester in which the first successful completion of the jth subject is recorded for the ith student. The ith student does not complete the jth subject according to the sample curriculum, if $\hat{\omega}_j < \omega_{i,j}$. In this case, the student is characterized by uncompleted subject patterns $a_{i,j}(k)$ where $k \in \{\hat{\omega}_j, \hat{\omega}_j + 1, \ldots, \omega_{i,j} - 1\}$. The consequence events $e_{fail}(k)$ indicate instances when the student does not proceed with studies in the $k + 1$ semester, resulting in university dropout. As detailed in the next subsection, this event will be treated as the competing risk when assessing the probability of the student continuing their studies.

				Semester					
Student ID	Subject ID	Missing starts	Missing ends	1	2	3	4	5	6
1	1	-	-						
	2	2	3						
	3	2	2						
	4	-	-						
	5	-	-						
	6	-	-						
	7	-	-						
2	1	1	2						
	2	2	2						
	3	2	2						
	4	1	2						
	5	1	2						
	6	1	2						
	7	-	-						
3	1	-	-						
	2	-	-						
	3	2	2						
	4	1	1						
	5	-	-						
	6	-	-						
	7	3	5						

FIGURE 4.2: A Gantt chart illustrating the missing subjects of students which have been not performed until it was advised by the sample curriculum. The semesters in which the subject should have been already completed are indicated by dark blue.

4.3.2 Investigation of student dropout with survival analysis taking into account the competing risks

Upon examining the academic path of a student, it becomes evident that successful graduation precludes any alternative outcome such as dropout for that individual. Consequently, unsuccessful program completion and successful graduation emerge as competing risks that necessitate careful consideration. For students who interrupt their studies or face dismissal for various reasons, there exists the possibility of re-enrollment. However, these particular students are excluded from the study. Determining the cumulative incidence function for the unfavorable case allows for the precise estimation of the dropout rate among students.

The comparison between the all-cause survival function estimated by the Kaplan-Meier method and the cumulative incidence function of dropout estimated by Equation 4.3 or 4.6 is depicted in Figure 4.3. The beginning of competing risks becomes apparent in the seventh semester, coinciding with the introduction of an additional outcome option—graduation. Typically, this study program spans seven semesters. In the absence of other competing risks, the cumulative incidence

function aligns with the empirical distribution from Kaplan-Meier, as evident in the figure up to the seventh semester. Beyond this point, the two functions begin to diverge. The difference between these functions specifically identifies the subset of students who have successfully graduated.

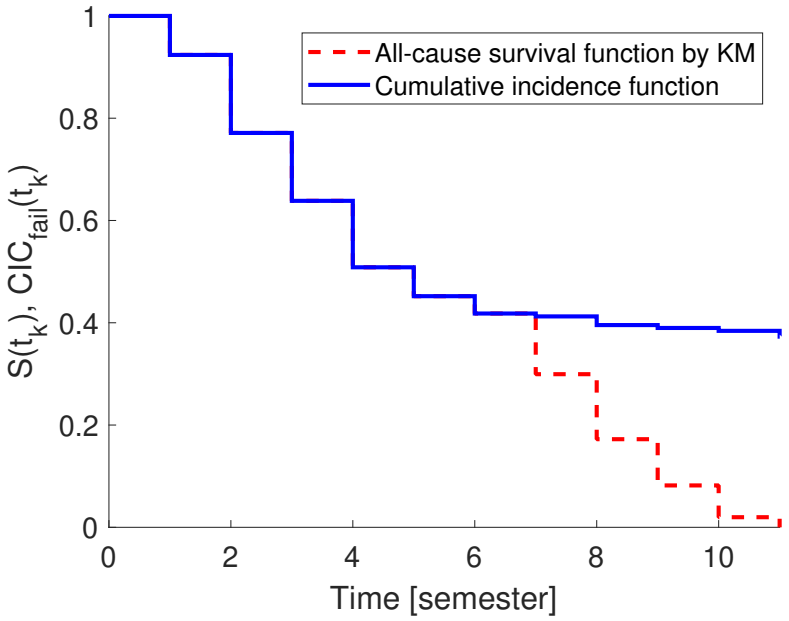

FIGURE 4.3: The empirical distribution by the Kaplan-Meier method (KM) (red) and the cumulative incidence function of non-graduate students (blue). As the competing risk of graduation only emerges from the seventh semester onward, it is anticipated that the two functions will exhibit differences starting from this semester. The distinction between these functions precisely identifies the subset of students who have successfully graduated.

A conventional but naive approach to addressing competing risks involves applying the Kaplan-Meier estimator while censoring samples related to competing events. However, this simplistic method violates the independent censoring assumption, leading to biased estimates. In the present of competing risks, the survival function is inadequate and a so-called cumulative incidence function needs to be given instead. The estimation of the cumulative incidence function is based on two distinct approaches: the cause-specific hazard and the sub-distribution hazard model. The comparison of the three functions is depicted in Figure 4.4.

The drawback of the Kaplan-Meier model, particularly when dealing with competing risks, is its tendency to describe only the entire population collectively. However, it is crucial to acknowledge that significant variations may exist when students traverse distinct pathways in completing or failing subjects during their university years. For example, the impact of failing in mathematics or chemistry during the first semester can differ significantly, emphasizing the introduction of

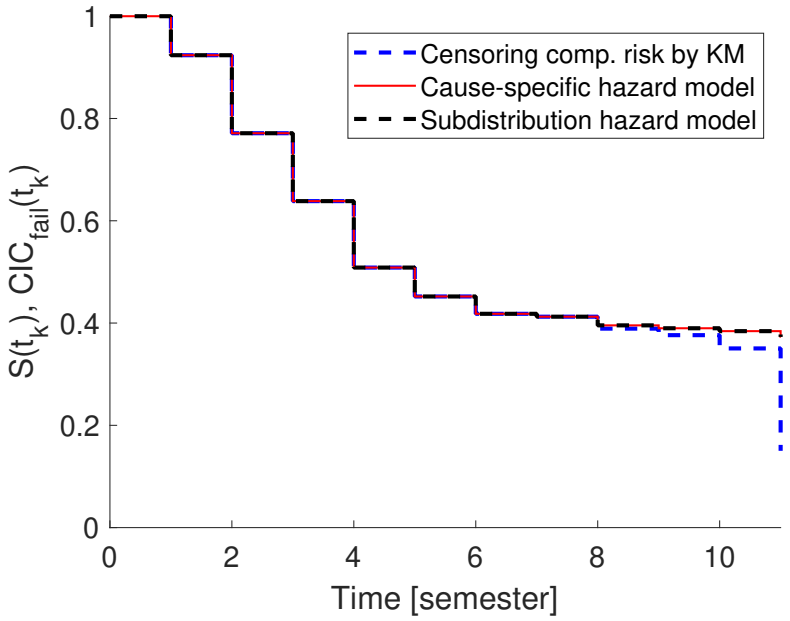

FIGURE 4.4: Comparison of cause-specific and subdistribution hazard models against conventional Kaplan-Meier (KM) based approach. Observations indicate a bias in the simplistic method.

event analysis through the utilization of frequent itemset and association rule mining methods, that are presented in the next section.

4.3.3 Event analysis with frequent itemset-based association rule mining

Building on the concepts presented in this case study, the event $a_{i,j}(k)$ signifies the non-completion of the jth subject for the ith student in the kth semester. Conversely, $\mathcal{A}_i(k) = \{a_{i,1}(k), \ldots, a_{i,j}(k), \ldots, a_{i,N_{i,k}}(k)\}$ represents the pattern of incomplete subjects for the ith student in the kth semester, while $\mathcal{A}(k) = \{\mathcal{A}_1(k), \ldots, \mathcal{A}_i(k) \ldots, \mathcal{A}_{n_k}(k)\}$ denotes the collection of missing subject completions for all students in the kth semester. It is crucial to note that the set $\mathcal{A}_i(k)$ is expanded to include triggered consequential events $e_c(k)$. These events can be specified as $e_{fail}(k)$ when a student fails at the end of the kth semester or $e_{graduate}(k)$ when the student successfully graduates. The extended set of events is denoted as $\mathcal{A}_i(k) = \{a_{i,1}(k), \ldots, a_{i,j}(k), \ldots, a_{i,N_{i,k}}(k), e_c(k)\}$.

Mining all the frequent itemsets leads to increased computational complexity in the context of this student dropout study. This arises because certain scenarios may occur where the support for particular uncompleted subjects equals the support

of their superset (the same subjects combined with others). In such instances, the additional subjects do not significantly impact dropout prediction and introduce unnecessary complexity. Therefore, they can be eliminated. To address this issue, the closed frequent itemset Mining method is applied, as it is adept at handling precisely this situation [116].

As each case study encompasses diverse types of relevant information, it is crucial to emphasize that, in the context of student dropout, specific conditions must be established while mining frequent itemsets. Certain outcomes may arise where the support for a particular uncompleted subject equals the support for that subject combined with others. In such instances, the additional subjects may not significantly impact dropout and might introduce unnecessary complexity. To mitigate this issue, the closed frequent itemset mining method is applied [116].

This method incorporates a critical hyperparameter—the minimum support for the frequent itemset mining algorithm. Adjusting this parameter allows for the fine-tuning of the complexity of the rules. A lower support threshold yields a higher number of rules, providing flexibility to optimize the complexity of the rules based on specific requirements. In this study, the minimum support parameter was set to 3%.

After performing the steps of frequent itemset-based association rule mining, the analysis of the results is the next step. The five most critical rules of every semester are summarized in Table 4.2. A rule can be considered critical if its confidence level is notably high. This metric serves as an indicator of the probability that, if a student fails to complete subjects that are in the rules, the consequential dropout event will occur with a probability equivalent to the confidence value. The seventh semester stands out with the absence of critical dropout rules. This anomaly can be attributed to the fact that only two students dropped out during this semester, signifying that a substantial portion of students prone to dropout had already done so in preceding semesters. The length of the rules exhibits a positive correlation with the progression of semesters. This can be attributed to the increasing number of potential subjects available to the student, encompassing those from preceding semesters that may remain uncompleted.

Guided by dropout rules that involve only one triggering event, critical subjects are identified, and their corresponding names are compiled in Table 4.3, utilizing both subject IDs and names as detailed in the Appendix. Notably, each semester

TABLE 4.2: The critical dropout rules of given semesters. It shows the pattern of uncompleted subjects, which should definitely be avoided by active students. A rule can be considered critical if its confidence level is notably high. The seventh semester stands out with the absence of critical dropout rules. The length of the rules exhibits a positive correlation with the progression of semesters.

Rule ID	**Rule length**	**Support (%)**	**Confidence (%)**	**Rule ID**	**Rule length**	**Support (%)**	**Confidence (%)**
1st Semester				2nd Semester			
259	6	4.02	56.00	1244	12	6.52	50.00
286	3	4.31	55.56	1246	11	6.83	50.00
197	7	3.45	54.55	1126	13	4.66	48.39
225	7	3.45	54.55	1242	13	3.73	48.00
252	4	3.74	52.38	1125	14	3.11	47.62
3rd Semester				4th Semester			
60	22	5.97	88.89	665	23	6.31	93.33
59	23	5.22	87.50	42	23	5.86	92.87
39	23	4.85	86.67	335	25	5.86	92.87
71	23	4.85	86.67	504	25	5.86	92.87
57	24	4.48	85.71	660	24	5.86	92.87
5th Semester				6th Semester			
33	28	3.41	100	13	24	3.21	100
75	29	3.41	100	26	24	3.21	100
84	28	3.41	100	117	24	3.21	100
86	28	3.41	100	131	23	3.21	100
87	27	3.98	100	189	23	3.21	100
7th Semester				8th Semester			
-	-	-	-	78	10	3.77	66.67
-	-	-	-	312	6	4.72	62.5
-	-	-	-	310	8	3.77	57.14
-	-	-	-	311	7	3.77	57.14
-	-	-	-	306	3	3.77	57.14
9th Semester				10th Semester			
2	12	3.28	100	10	4	3.45	100
-	-	-	-	4	9	3.45	100
-	-	-	-	5	6	6.90	100
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-

presents subjects that are seemingly critical. For instance, foundational engineering subjects like mathematics, physics, and chemistry emerge as critical subjects. Additionally, certain uncompleted subjects persist across multiple semesters. Notable examples include the comprehensive chemistry exam, surfacing from the fifth semester onwards, identified as critical in three semesters. Further critical subjects also occur, but in conjunction with other subjects that don't appear as single rules because closed itemsets were mined, and these subjects alone have the same support as when considered together. For example, failing in Material science and

Computer science for engineers I. has 51.61% confidence. The supports and confidences of the rules can be utilized to estimate the cumulative incidence function of dropout, as presented in the next section.

TABLE 4.3: Critical subjects are identified from dropout rules that involve only one triggering event. Foundational engineering subjects like mathematics, physics, and chemistry emerge as critical subjects. Additionally, certain uncompleted subjects persist across multiple semesters.

Semester	**Subject ID**	**Name of subject**	**Confidence**
1	60	General and inorganic chemistry	15.49
1	23	Economics	13.97
1	7	Physics I.	13.37
1	46	Computer science for engineers I.	9.48
2	11	Physical chemistry I.	24.63
2	40	Numerical mathematics	22.73
2	62	Problem solving in general and inorganic chemistry II.	22.37
2	26	Methematical analysis II.	21.28
2	41	Statistics	21.12
3	36	Technical fluid mechanics	25.73
3	52	Transportphenomena	19.23
5	20	Comprehensive exam in chemistry	12.82
6	20	Comprehensive exam in chemistry	9.09

4.3.4 Estimating the marginal probability of student dropout

The cumulative incidence function can be estimated based on the supports and confidences of the rules. When focusing solely on consequence events, the baseline cumulative incidence function can be derived both with cause-specific and subdistribution hazard models using Equations 4.10 and 4.11. The resulting function is validated by comparing it to the cumulative incidence function generated from traditional survival analysis, as depicted in Figure 4.5. The results of the proposed model notably approximate the cumulative incidence function from survival analysis very effectively.

Gaining admission to an engineering course in Hungary, even those supported by the government, is notably accessible due to the high demand for this profession. Consequently, a substantial number of students enroll with the intent to complete the program. However, many soon encounter the challenges involved and find it

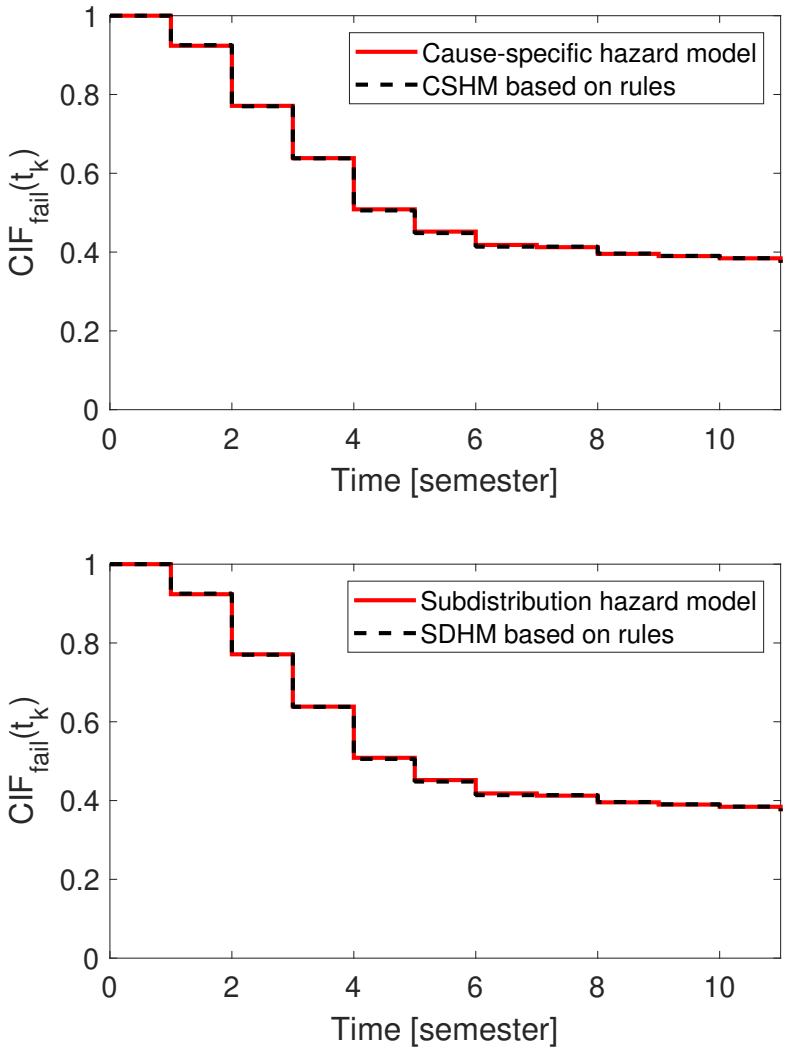

FIGURE 4.5: Comparison of the cumulative incidence function formed from frequent itemset-based association rules and the traditional approach for non-graduate students. The comparison is performed for both Cause-Specific Hazard Model (CSHM) and Subdistribution Hazard Model (SDHM) and have the same output for baseline function. The functions illustrate well that the proposed methodology is able to predict the dropout of a student.

challenging to persist. Notably, over half of the students leave their studies by the end of the fourth semester. In the initial two semesters, a relatively smaller number of students leave by their own decision upon realizing the difficulty of the program. In contrast, the dropout rate significantly increases in the third semester due to specific requirements for continuing the studies. To progress, every student must successfully complete all subjects recommended by the sample curriculum within the first semester. Nevertheless, there is a provision known as a 'fairness request', permitting students to complete one subject in the fourth semester in exceptional cases. The dropout rate in the fourth semester typically impacts those who have not effectively managed the fairness request. Additionally, there is an

extra requirement in the fifth semester, contributing to a less significant dropout rate. Once students reach the 5th semester, they become less likely to drop out after this point. In conclusion, based on the 11th semester, it can be stated that approximately 40% of students can graduate on their first attempt.

To present the effectiveness of the developed method from several perspectives, a comparative analysis was performed with the Naive Bayes Classification method. Based on the results, it can be observed that the classifier struggles to accurately estimate dropout based on uncompleted subjects. The cumulative incidence function of the Naive Bayes classifier and survival analysis is compared in Figure 4.6. It becomes evident that the Naive Bayes model tends to overestimate the number of failures, suggesting a weakness in its predictive capabilities. However, in identifying failed students, the model demonstrated accuracy, indicating potential suitability as an alerting system.

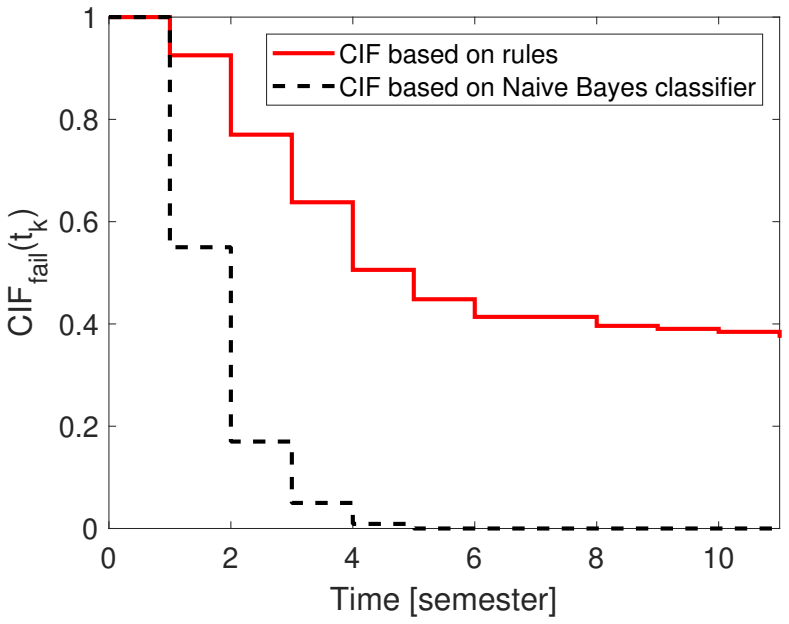

FIGURE 4.6: Cumulative Incidence Functions (CIF) formed from the Naive Bayes classifier and survival analysis for non-graduate students. The functions illustrate well, that the Naive Bayes classifier can poorly predict the dropout of a student.

Having different uncompleted subject patterns in the different semesters have different influence on the dropout rate. The estimated cumulative incidence function may deviate significantly from the baseline. The estimation of the cumulative incidence function in case of specific uncompleted subject patterns is presented in the next section.

4.3.5 Estimating probability of student dropout with specific uncompleted subject patterns

To estimate cumulative incidence functions for students that are characterized by distinctive uncompleted subject patterns, the dataset is segmented according to the supported subjects of the selected frequent itemsets. Consider two scenarios characterized by rules with a high amount of uncompleted subjects (scenario 1) and fewer uncompleted subjects (scenario 2) up to the second semester. In Scenario 1, students fail to complete the following subjects in the first semester: $X_1(1) = \{7, 8, 23, 24, 46, 60, 61\}$, while in the second semester: $X_1(2) = \{7, 8, 9, 10, 11, 23, 26, 27, 37, 40, 41, 43, 60, 61, 62, 63\}$. In Scenario 2, students fail to complete the following subjects in the first semester: $X_2(1) = \{7, 24, 46\}$, while in the second semester: $X_2(2) = \{9, 11, 26\}$. Scenario 1 includes 14 students, while scenario 2 comprises 49. The estimated cumulative incidence functions are illustrated along with the baseline function in Figure 4.7.

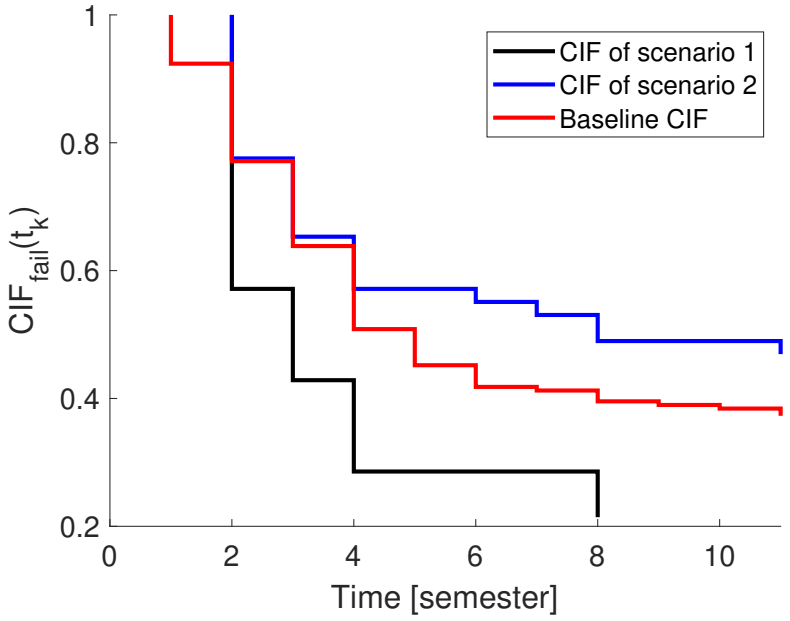

FIGURE 4.7: The Cumulative Incidence Functions (CIF) of different student path scenarios. Students in scenario 1 have the worst chances of obtaining a diploma. Analyzing the curves, it can be asserted that almost 80% of the students characterized by scenario 1 will drop out, while 53% of the students characterized by scenario 2 will experience dropout.

After the first semester, students have a probability of 1 to stay in the program. This occurs because students have uncompleted subjects in the second semester, indicating that they must definitely continue until the second semester. It is evident that the length of the rule positively correlates with the dropout chances. This

correlation is logical as students, in such cases, have additional work to complete for these subjects, and most of them struggle to meet these demanding requirements. Students in scenario 1 have the worst chances of obtaining a diploma. Analyzing the curves, it can be asserted that almost 80% of the students characterized by scenario 1 will drop out, while 53% of the students characterized by scenario 2 will experience dropout.

The proposed method is well-suited for estimating the probability of dropout for an active student still in training based on their current uncompleted subjects. Predictions can be generated by matching the missing subject completions with the determined rules. Additionally, a personalized forecasting system can be established, considering additional variables such as general, demographic, or enrollment data.

4.4 Summary of the chapter

In this chapter, a novel integrated survival analysis and frequent itemset-based association rule mining method was introduced. The probability of competing risks were determined at specific time instances using frequent itemset-based association rules. The proposed method operated within a discrete time domain. The survival was characterized by categorical explanatory variables at every time instance. The time-dependent categorical variables were conceptualized as triggering events, acting as precursors, initiating consequent events that signify the competing outcomes of the survival process. This approach identified relevant triggering events, that lead to a competing risk. The approach also segmented the dataset for subjects characterized by a sequence of frequent itemsets consisting of specific combinations of triggering events. Consequently, individuals with these characteristics were identified by a global, single-variable, time-independent feature. The cumulative incidence function of a specific competing risk were directly determined for the population with a given sequence of frequent itemsets.

The practical application of the proposed method was demonstrated by analyzing course completion data from former chemical engineering students at the University of Pannonia. Student dropout is a significant phenomenon that leads to economic loss and social tension. This chapter illustrated that survival analysis based on competing risk models effectively provides an estimate of the probability

of graduation. However, by incorporating the proposed method, the root causes of student dropout were more deeply explored. The study integrated the sample curriculum that prescribes the recommended semester for each subject completion. Inability to meet this requirement marks an uncompleted subject event, a crucial factor associated with subsequent student dropout. Although many researchers have analyzed this issue, no method has been developed thus far to predict the academic success of students based on uncompleted subject patterns. This framework was applied to mine frequent itemset-based association rules, highlighting the probability of dropout with specific uncompleted subject patterns. Using the modified support and confidences directly, cumulative incidence functions were estimated for students with specific sequences of frequent itemsets.

Another dropout phenomenon is that students may decide to reapply for the course at any time. This decision is often made in an attempt to improve their chances by erasing previous unfavorable results and resetting the requirement system. As the method under examination focuses solely on the first attempt at the training, these students are also considered as having dropped out. Previous studies indicate that only a few students successfully complete the training after reapplying. Despite this, the general experience suggests that reapplying may not be worth the effort due to a significant rate of failure.

The obtained results suggest the necessity for the university management to reconsider certain functional elements. First, it would be essential to reschedule the subjects in the sample curriculum. Some subjects, designed to build primary skills for later semesters, currently appear in the earlier part of the curriculum. Given that a significant number of students dropped out in the 3rd semester due to specific requirements, a reassessment of the timing of these requirements becomes important. Additionally, it is observed that there is a correlation between the success of students and the instructor teaching a particular subject. In this context, organizing targeted training for these educators, following Section 1.5 of the European Standards and Guidelines [117], would be crucial.

The proposed method has proven to be well-suited for estimating the probability of dropout for an active student still in training based on their current uncompleted subjects. Predictions can be generated by matching the missing subject completions with the determined rules. Additionally, a personalized forecasting system can be established, considering additional variables such as general, demographic, or enrollment data. The model may also be suitable for examining a wide class of

problems. An important characteristic of the applications is the presence of overlapping process steps and the occurrence of transitions caused by the triggering phenomenon. Examples include the development activities, so the method seems to be suitable to support capability maturity model integration processes, which will be one of the future research avenues.

Chapter 5

Integrated survival analysis and sequential pattern mining: a healthcare application

This chapter focuses on integrating sequential pattern mining with survival analysis to provide insights into event co-occurrences and their temporal relationships. In this framework, two different but similar algorithms are presented. Section 5.1 provides a clear overview of the challenges the algorithm aims to address and briefly introduces the algorithm. Section 5.2 presents the detailed mathematical description of the proposed integrated survival analysis and frequent sequence mining algorithm, while Section 5.3 demonstrates the detailed mathematical description of the proposed integrated survival analysis and sequential rule mining algorithm. Section 5.4 showcases the practical application of the integrated survival analysis and frequent sequence mining algorithm through a medical case study. Finally, Section 5.5 presents the practical application of the integrated survival analysis and sequential rule mining algorithm by establishing a medical decision support system.

5.1 Introduction

Sequential pattern mining is a robust technique for identifying event sequences within a sequential database, offering a deeper understanding of temporal order and event relationships [118]. In this chapter, frequent sequence mining and sequential rule mining algorithms are applied in this context. The frequent sequence mining algorithms typically use a support threshold, which is the minimum number of times (or minimum frequency) a sequence must occur in the dataset to be considered frequent. The resulting frequent sequences can be characterized by probabilities, namely support and confidence, based on the frequency of the input sequences. The support value is usually represented as a percentage, signifying the probability of occurrence of a specific sequence. On the other hand, the confidence value represents the probability of a specific sequence continuing, given the occurrence of previous events. The previous and the continuing events are called antecedent and consequent events, respectively.

Frequent sequence mining can be implemented using several algorithms [119]. In this study, a co-occurrence map-based algorithm called as CM-SPAM [120] was applied, which enables faster mining of all frequent sequences without constraints. Typically, constraints such as the mining of top k, closed, maximal [115] or nonoverlapping sequential patterns [121] are applied to reduce computation time, especially for large datasets. However, the proposed algorithm requires all possible frequent sequences to be used for calculating confidence values in the further step of the algorithm. Due to this requirement, constraints are not employed, as they would restrict the inclusion of potential sequences. Consequently, the frequent sequences exhibit significant redundancy, which necessitates the selection of relevant sequences for further analysis. To accomplish this requirement, association rules [122] are formulated to identify the rules with antecedent events of interest. The objective of the algorithm is to capture all potential consequent events associated with the identified antecedent events, thereby facilitating the prediction model for possible future event occurrences.

In traditional frequent sequence mining algorithm, support and confidence are typically considered constant. However, in certain contexts, they can be time-dependent. For instance, it matters whether event A is followed by event B after 1 day or 10 year. The probability of event A being followed by event B is lower

after 1 day but higher after 10 years. Traditional frequent sequence mining approaches cannot precisely capture this temporal relationship. To address this limitation, this study utilizes survival analysis to characterize probabilities as a function of time. The frequent sequence mining supplies the IDs of supporting input sequences. These sequence ID-s can be used to calculate elapsed times at the relevant continuations, providing input for the survival analysis to derive the survival function, which is the baseline of the temporal characteristics. The unbiased temporal dynamics are derived by multiplying the survival function with rule confidence. The integration of frequent sequence mining and survival analysis adequately accounts for multiple potential event continuations. Traditional survival analysis can only estimate one occurrence of events, otherwise the handling of competing risks becomes mandatory. Furthermore, reliable confidence intervals are provided using the bootstrapping technique, which, despite its conceptual simplicity, proves to be a powerful tool within the proposed framework.

Two major limitations are associated with the proposed method. Firstly, the resulting event continuations lack causal relationships. Consequently, validating the outcomes necessitates additional knowledge-based information to prevent irrelevant associations. Secondly, dealing with a substantial volume of unique events can make every trace highly individual. The potential frequent sequences also have a vast number of possible permutations, resulting in long sequences unable to become frequent due to the uniqueness of traces. Consequently, the selection of relevant events becomes important and necessitates the incorporation of additional knowledge-based information. Additionally, the application of frequent itemset mining can be considered for merging events, thereby reducing the number of events, which may constitute a further research direction. Reducing the minimum support hyperparameter can also be an option, but it comes at the cost of significantly increased computational time.

An approach that can help to handle large number of unique events is the sequential rule mining. Sequential rule mining stands out as a potent technique that capitalizes on such sequential databases to spotlight meaningful continuations of event itemset [123]. The method involves identifying potential itemset of antecedent events and their corresponding itemset of consequent events, illuminating the associations between these occurrences. In this case, the order of events is not explored; instead, sets of events that may occur in any order before another set of events that also occurred in any order are considered. Consequently,

the permutation of potential outcomes becomes more limited. A sequential rule can be considered relevant if its support and confidence values exceed a certain threshold [124]. The confidence value of the rule indicates the probability of the occurrence of consequent events, given that antecedent events have already occurred.

The dynamic temporal aspect of event sequences is frequently overlooked in recent publications in this domain. In the proposed comprehensive framework, the integration of frequent sequence mining and survival analysis addresses this gap, allowing for the capture of relevant event occurrences and the provision of comprehensive temporal information between these events. The specific contributions of the chapter can be outlined as follows:

- The time-dependent support and confidence functions of event transitions can be estimated using the integrated survival analysis and frequent sequence mining algorithm. The approach identifies relevant event continuations through frequent sequence mining. The temporal characteristics of the resultant rule confidences are determined using the Kaplan-Meier estimator. The multiplication of these two terms provides the time-dependent confidence function.

- Sequential rule mining can be an alternative approach when handling a substantial volume of unique events that are poorly distributed. Unlike providing a continuation of events, this method identifies sets of antecedent events that may occur in any order before another set of consequent events that also may occur in any order. The determination of temporal characteristics of the resultant rule can be made using the same approach.

- The confidence intervals of the time-dependent confidences can be determined using the bootstrapping method. This involves randomly selecting input sequences and executing the method on this data. The process is executed repeatedly, resulting in a set of confidence functions from bootstraps. The percentile-based method estimates the confidence bounds in this set of functions, thereby establishing the confidence intervals.

The developed frequent sequence mining-based survival analysis method is demonstrated using clinical data, where registered diseases are treated as sequences for

each patient. The mined frequent sequence of consecutive diseases can be divided into antecedent and consequent parts, meaning current and potential future disorders, respectively. Although representing the medical history of patients as a sequence is already a strong strategy, further developments are needed [125]. In this regard, the current study also considers the confidence function between current and potential future disorders, as the time-dependent probability of disease development. The proposed method has the potential to establish a decision support system for healthcare professionals, enabling the prediction of future potential disorders. However, its implementation should include additional factors describing the condition of patients, such as laboratory data. The combination of diagnoses and laboratory events results in an increased number of unique events that are poorly distributed. Moreover, the significantly fewer unique laboratory events with the same amount of records as diagnose events, the diagnosis events become infrequent compared to laboratory events. Therefore, the application of sequential rule mining is necessary to generate more representative results.

The next section presents the mathematical description of the proposed methodology.

5.2 Description of the frequent sequence mining-based survival analysis method

The proposed method integrates frequent sequence mining with survival analysis, creating a comprehensive framework for capturing the temporal dynamics of event sequences, as depicted in Figure 5.1. The algorithm utilizes the CM-SPAM algorithm, a co-occurrence map-based approach, to mine frequent sequential patterns while also assessing their supports and confidence values. Since the resultant frequent sequences often exhibit significant redundancy, association rules are employed to identify rules with the relevant antecedent events. This step discovers all potential consequent events, enabling the construction of prediction models for future event occurrences. In addition, the relevant events can be ranked based on confidence values to highlight the most crucial ones. The temporal characteristics of rules are determined by survival analysis, where the Kaplan-Meier estimator determines the distribution of the elapsed time between two events. The confidences of the rules are then multiplied by the estimated distribution functions,

resulting in an unbiased confidence function that quantifies the probability of the consequent event occurring within a specified timeframe. Estimating confidence intervals for the resultant confidence functions poses a notable challenge due to the uncertainties associated with the rule support and the survival function. To address this challenge, the bootstrapping method is applied, which is well-suited for such complex analytical calculations. The hyperparameter of the algorithm is the threshold value for the minimum support of the sequences.

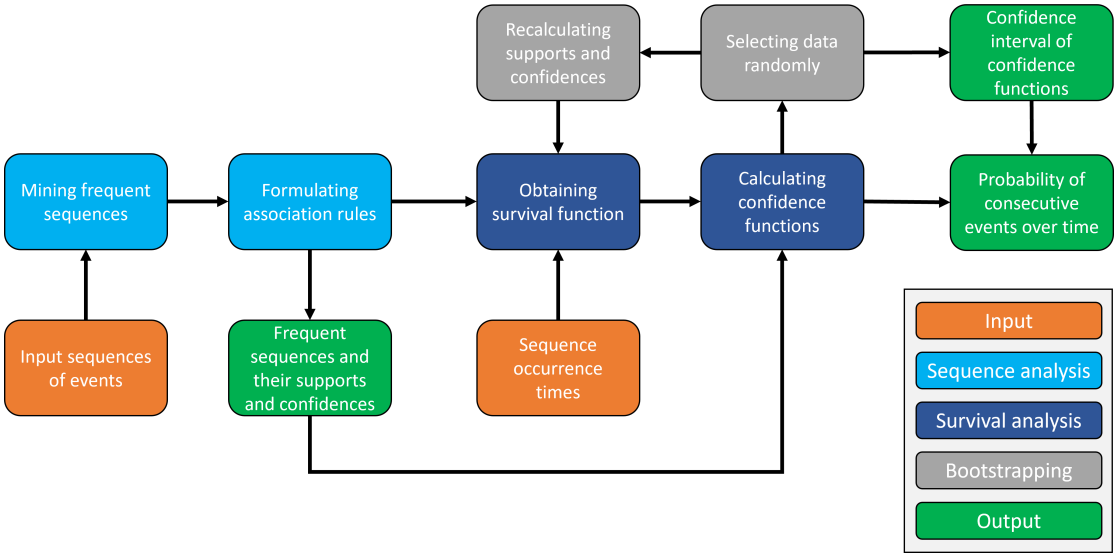

FIGURE 5.1: The steps of the proposed method. The CM-SPAM algorithm is initially employed to mine frequent sequential patterns from the dataset. Subsequently, association rules are utilized to identify relevant antecedent events and determine potential consequent events. The temporal dynamics are captured using the Kaplan-Meier estimator. Finally, the confidence values are multiplied with the estimated distribution function, resulting in an unbiased confidence function for predicting event occurrences.

A detailed description of the algorithm is given below. Section 5.2.1 describes the formulation of sequence mining, while Section 5.2.2 introduces how the model can be extended by timestamps and explanatory variables of the traces. Section 5.2.3 presents how survival analysis can be applied to visualize the time-dependent nature of confidence and support values, as well as how it can be applied to predict a single consequent part. Section 5.2.4. describes how to predict multiple sequences. A statistical test is presented in Section 5.2.5 to compare continuations with subgroups based on explanatory variables. Section 5.2.6 introduces how the existing model can be extended by bootstrapping to determine the confidence intervals. In Section 5.2.7, the crucial execution steps of the algorithm are meticulously outlined, with each phase of the analytical process intricately

detailed. Finally, Section 5.2.8 provides context for the proposed methodology by outlining its foundational principles and related research directions.

5.2.1 Formulation of the frequent sequence mining problem

The problem of sequential pattern mining can be formulated accordingly [126]: Let $E = \{e_1, \ldots, e_M\}$, be a set of M unique events. In medicine, unique events are disorders or laboratory data. Similarly, in manufacturing, the analysis of process steps and quality inspections can be conducted as unique events. The ordered list of events related to the same case is represented in a trace. Patients diagnoses or manufacturing steps in ordered list can represent a trace. Every trace has a unique identifier, referred to as a trace ID (patient ID, or process ID), consisting of a sequence of events.

The raw data set of frequent sequence mining and the proposed post-processing method contains information about the trace IDs, the appearances of items, and the time stamps of the occurrences. Additionally, each trace is characterized by a vector of discrete explanatory variables, denoted as $\mathbf{Z}_i = [z_{i,1}, z_{i,2}, \ldots, z_{i,N_Z}]$. These variables describe the features of the events, and each trace is associated with a single vector. In the healthcare domain, features of a patient can include gender and age, while in the context of production, characteristics like temperature or humidity can be used for characterization. An example of the raw database without the explanatory variables of the traces can be seen in Table 5.1. One can see in the table that Trace ID 1 had an event associated with item e_1 on 03.02.2019 and with item e_2 on 05.02.2019.

TABLE 5.1: An example of the raw database. One can see, that Trace ID 1 had an event associated with item e_1 on 03.02.2019 and with item e_2 on 05.02.2019.

Trace ID	**Item**	**Time stamp**
1	e_1	03.02.2019.
1	e_2	05.02.2019.
2	e_3	12.02.2019.
2	e_2	20.07.2020.
2	e_3	01.04.2021.
3	e_1	15.04.2019.

The consideration of the trace IDs, events and timestamps defines a touple of events and leads to the formulation of ordered lists of event-timestamp duplet [127],

referred to as trace, denoted as $\psi_i = \langle (a_{i,1}, \tau_{i,1}) \rightarrow \ldots \rightarrow (a_{i,j}, \tau_{i,j}) \rightarrow \ldots \rightarrow (a_{i,N_i}, \tau_{i,N_i}) \rangle$, where $a_{i,j} \in E$ stands for the jth event of ith trace, while $\tau_{i,j}$ means the timestamp of jth event of ith trace and $\tau_{i,j-1} < \tau_{i,j}$. The collection of sequences is denoted by $\Psi = \{\psi_1, \ldots, \psi_i \ldots, \psi_N\}$, where N represents the total number of traces.

Frequent sequence mining aims to extract a set of sequences from such data that appear frequently. In this work, frequent sequence is represented as $\alpha = \langle \alpha_1 \rightarrow \ldots \rightarrow \alpha_l \rightarrow \ldots \rightarrow \alpha_b \rangle$, where α_l stands for the lth event of the frequent sequence. For a sequence α, if α_l occurs before α_b, it is denoted as $\alpha_l < \alpha_b$. The support of sequence α, denoted by $supp(\alpha)$, is the total number of traces, where the sequence α appears in the correct order, expressed by:

$$supp(\alpha) = \frac{\left|\{\xi \in \Psi : \alpha \preceq \xi\}\right|}{N} \tag{5.1}$$

where $\alpha \preceq \xi$ represents that sequence α is a subsequence of sequence ξ, if a one-to-one order-preserving function f exists that maps events in α to events in ξ, that is, $\alpha_l \subseteq f(\alpha_l)$, moreover, if $\alpha_l < \alpha_b$ then $f(\alpha_l) < f(\alpha_b)$. For example, sequence $\langle A \rightarrow E \rangle$ is a subsequence of $\langle A \rightarrow Q \rightarrow C \rightarrow E \rangle$. The sequence α is referred to as frequent if the value of its support is higher, than a predefined threshold, namely $supp(\alpha) > minsupp$. To provide a clear definition of support, an example is presented in Figure 5.2. In this example, events are denoted by letters, and timestamps are indicated by numbers. The frequent sequence $\langle A \rightarrow B \rightarrow E \rangle$ signifies that event A is followed by event B, and then by event E. In Trace 1, events A, B, and E are observed to occur in the correct order, thus supporting the frequent sequence $\langle A \rightarrow B \rightarrow E \rangle$. Following this logic, the sequential rule $\langle A \rightarrow B \rightarrow E \rangle$ is supported by Trace 2 but not by Trace 3.

Sequential association rules can be formed by splitting the sequences into antecedent and consequent parts. Therefore, the sequence α is represented as $\alpha = \langle \alpha_1 \rightarrow \ldots \rightarrow \alpha_l \rightarrow \ldots \rightarrow \alpha_b \rangle = \alpha^l \rightarrow \alpha^b$, where $\alpha^l = \langle \alpha_1, \rightarrow \ldots \rightarrow \alpha_l \rangle$ and $\alpha^b = \langle \alpha_{l+1}, \rightarrow \ldots \rightarrow \alpha_b \rangle$ $(1 < l < b)$, are called the rule antecedent and the rule subsequent, respectively. The superscript letter indicates a subsequence of α, which may contain multiple event transitions, while the subscript letters only denote one event. This form can be applied to infer future events when the occurrence of α^b is considered, since the antecedent of α^l has previously occurred in the case of a trace. In healthcare, such an association rule can be identified,

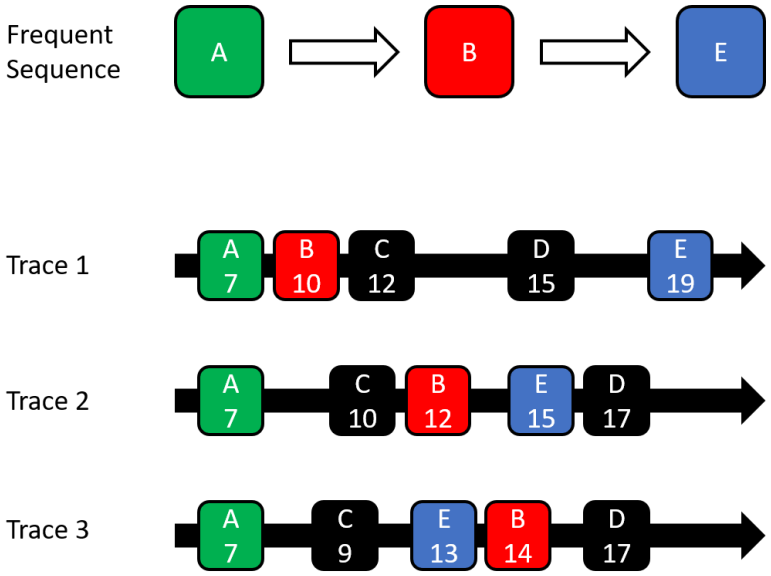

FIGURE 5.2: An example of the definition of support. The frequent sequence $\langle A \rightarrow B \rightarrow E \rangle$ signifies that event A is followed by event B, and then by event E. In Trace 1, events A, B, and E are observed to occur in the correct order, thus supporting the frequent sequence $\langle A \rightarrow B \rightarrow E \rangle$. Following this logic, the sequential rule $\langle A \rightarrow B \rightarrow E \rangle$ is supported by Trace 2 but not by Trace 3.

for example, by assuming that if a patient already has the condition Hyperthonia, then the occurrence of Diabetes becomes a potential continuation, as represented by the rule: $Hypertonia \rightarrow Diabetes$. This rule suggests a sequential relationship between these two medical conditions. The strength of the rule-consequent can be defined by a conditional probability known as confidence. The probability that α^l continues with α^b can be given by:

$$conf(\alpha^l \rightarrow \alpha^b) = \frac{supp(\alpha^l \rightarrow \alpha^b)}{supp(\alpha^l)} \tag{5.2}$$

The frequent sequences can be broken down into sequence event transitions, where the $(l+1)$th event of the sequence α, represented as α_{l+1}, follows the rule-antecedent α^l that already consists of l elements. Related conditional probabilities are referred to as confidences of continuations, denoted by $conf(\alpha^l \rightarrow \alpha_{l+1})$, and can be calculated as follows:

$$conf(\alpha^l \rightarrow \alpha_{l+1}) = P(\alpha_{l+1}|\alpha^l) = \frac{supp(\alpha^l \rightarrow \alpha_{l+1})}{supp(\alpha^l)} \tag{5.3}$$

The probability that the sequence $\alpha = \langle \alpha_1 \rightarrow \ldots \rightarrow \alpha_l \rightarrow \ldots \rightarrow \alpha_b \rangle$ occurs can be described as the chain rules of these confidences of continuations, or as cumulative confidence. This probability, denoted by $conf(\alpha)$, is called the confidence of the sequence α and can be calculated as follows:

$$conf(\alpha) = supp(\alpha_1) \prod_{f=1}^{b-1} conf(\alpha^f \rightarrow \alpha_{f+1}) \tag{5.4}$$

An example of calculating confidences and cumulative confidences can be seen in Table 5.2. Let us consider a frequent sequence $\langle A \rightarrow B \rightarrow C \rangle$. To compute the rule confidence, the support of frequent sequence $\langle A \rightarrow B \rangle$ should be calculated, and the support of event A.

TABLE 5.2: An example of calculating confidences and cumulative confidences

Frequent sequence	**Support**	**Confidence**	**Cumulative confidence**
A	100	1	1
$A \rightarrow B$	50	0.5	0.5
$A \rightarrow B \rightarrow C$	10	0.20	0.1

The following section presents how the temporal factors can be established based on timestamps and how the explanatory variables can enrich the analysis.

5.2.2 Taking into account the timestamps and the explanatory variables of the events

The main contribution of the work is that by analyzing these transition probabilities, the time instances and explanatory variables of the events are also taken into account to calculate the time dependencies. Accordingly, the frequent sequences are enriched by the knowledge of time stamp of their appearance as follows:

$$\alpha(\tau_i) = \langle (\alpha_1, \tau_{i,1}) \rightarrow \ldots \rightarrow (\alpha_l, \tau_{i,l}) \rightarrow \ldots \rightarrow (\alpha_b, \tau_{i,b}) \rangle \tag{5.5}$$

where $\alpha(\tau_i)$ represents the ordered list of frequent event-timestamp duplet of the frequent sequence α in the ith trace and $(\alpha_l, \tau_{i,l})$ denotes the lth frequent event-timestamp duplet in the ith trace. It should be noted that since α_l represents the lth event in the sequence α, $\tau_{i,l}$ denotes the time instant of α_l in the ith trace.

This information enables taking into account the time demand of the transitions, i.e., the elapsed time from sequence α^l to the occurrence of event α_{l+1}:

$$\Delta t_i(\alpha^l \to \alpha_{l+1}) = \tau_{i,l+1} - \tau_{i,l} \tag{5.6}$$

where $\Delta t_i(\alpha^l \to \alpha_{l+1})$ denotes the duration from α_l to α_{l+1} in the case of the ith supporting trace. Let us consider the example in Figure 5.2. The frequent sequence $\langle A \to B \to C \rangle$ can be represented as $\langle \alpha^2 \to C \rangle$, where $\alpha^2 = \langle A \to B \rangle$. As discussed earlier, this frequent sequence is supported by Trace 1 and 2. Therefore, Δt_1 and Δt_2 can be calculated as $\Delta t_1 = 19 - 10 = 9$ and $\Delta t_2 = 15 - 12 = 3$, respectively.

The distribution of these elapsed times can be studied, which helps in assessing the time-dependent characteristics of the support for the sequences. Therefore, Equation 5.1 is modified as follows:

$$supp(\alpha^l \to \alpha_{l+1}, t) = \frac{\left|\{\xi \in \Psi : \alpha^l \to \alpha_{l+1} \preceq \xi, \Delta t_\xi(\alpha^l \to \alpha_{l+1}) \leq t\}\right|}{N} \tag{5.7}$$

The proposed post-processing method can also be extended to incorporate explanatory variables. By considering the values of these explanatory variables, the database can be partitioned into subgroups and comparative analyses can be conducted within the subgroups. While forming subgroups based on discrete variables is a straightforward task, when dealing with continuous variables, it is advisable to employ discretization through clustering algorithms. The integrated survival analysis and expectation-maximization-based clustering presented in Chapter 3 stands out as a suitable option for this purpose.

The next section introduces how the temporal relation can be determined on the basis of survival analysis.

5.2.3 Kaplan-Meier empirical survival function-based analysis of the frequent sequences

This section details the extension to account for time-dependent confidences, in order to determine a function that describes the time distribution between two

events.

The survival function of the time elapsed between events α_{l+1} and α_l can be empirically determined. This function provides the baseline of the probability that the event α_{l+1} occurs later than a certain time, given that the sequence α^l has already occurred. The function can be introduced as follows [35]:

$$S(\alpha^l \to \alpha_{l+1}, t) = P(\Delta t(\alpha^l \to \alpha_{l+1}) > t) \tag{5.8}$$

The Kaplan-Meier estimator [31] is a straightforward technique that provides a foundation for calculating the survival function. The traditional survival analysis can incorporate the concept of censoring, primarily used when a transaction or event cannot be observed until the occurrence. As a result, the exact occurrence time becomes unknown, but the available observations can still provide valuable information up to the moment of censoring. The survival function can be expressed by the following formula [32]:

$$S(\alpha^l \to \alpha_{l+1}, t) = \prod_{f:\Delta t_f \leqslant \Delta t} \left(1 - \frac{m_f}{N_f}\right) \tag{5.9}$$

where N_f represents the number of events that have not been occurred or censored at time difference Δt_f, while m_f is the number of activities occurred between periods of time differences Δt_{f-1} and Δt_f.

The survival function and the confidence of the rule are multiplied to derive the confidence function. The confidence function describes the probability that consequent event α_{l+1} occurs earlier than a certain elapsed time, given that antecedent sequence α^l has already occurred as one of the competing risks. The Equation 5.2 is modified as follows:

$$\mathit{conf}(\alpha^l \to \alpha_{l+1}, t) = \big(1 - S(\alpha^l \to \alpha_{l+1}, t)\big)\, \mathit{conf}(\alpha^l \to \alpha_{l+1}) \tag{5.10}$$

If there are censored samples in the database, the confidence function can only be estimated by calculating the cumulative incidence function based on cause-specific or subdistribution hazard model, presented in the recent section.

The next section presents how the temporal relations can be derived for an entire frequent sequence.

5.2.4 Estimation of confidence functions at a given event continuation of a consequent sequence

The proposed post-processing method can estimate confidence functions for every continuation of the rule consequent α^b, not limited to just $\alpha^l \rightarrow \alpha_{l+1}$. Suppose that α_l has already occurred, the confidence function of the event α_{l+1} can be calculated based on the method presented in the previous sections. This section further extends the methodology to handle the estimation of the occurrence of event α_{l+1} followed by event α_{l+2}.

Let us consider the frequent sequence $\alpha = \langle \alpha^l \rightarrow \alpha^b \rangle = \langle \alpha^l \rightarrow \ldots \rightarrow \alpha_{l+h-1} \Rightarrow \alpha_{l+h} \rightarrow \ldots \rightarrow \alpha_b \rangle$, where α^l is the antecedent sequence, and $0 < h < b - l$. In the consequent sequence α^r, the continuation of the event of interest is denoted by "$\Rightarrow$". This double-lined arrow plays a role when the survival function is involved, as it delineates the boundaries for calculating the elapsed time. Each continuation can be characterized by a confidence function. The observation time of an event continuation begins when the last event of the antecedent sequence occurs and ends when the interested continuation is completed in the consequent sequence. The exact time differences between α_l and α_{l+h} can be calculated as follows:

$$\Delta t_i(\alpha^l \rightarrow \ldots \rightarrow \alpha_{l+h-1} \Rightarrow \alpha_{l+h} \rightarrow \ldots \rightarrow \alpha_b) = \tau_{i,l+h} - \tau_{i,l} \tag{5.11}$$

Let us consider the example in Figure 5.2. The frequent sequence $\langle A \rightarrow B \rightarrow C \rangle$ can be represented as $\langle \alpha^1 \rightarrow B \rightarrow C \rangle$, where $\alpha^1 = \langle A \rangle$. As discussed earlier, this frequent sequence is supported by Trace 1 and 2. Therefore, Δt_1 and Δt_2 can be calculated as $\Delta t_1 = 19 - 7 = 12$ and $\Delta t_2 = 15 - 7 = 8$, respectively.

The frequency of occurrence times is equal to $supp(\alpha^{l+h})$ and this parameter, denoted by $supp(\alpha^{l+h}, t)$ is also time-dependent. The empirical survival function based on Equation 5.8 can be given as follows:

$$\begin{aligned} S(\alpha^l \rightarrow \ldots \rightarrow \alpha_{l+h-1} \Rightarrow \alpha_{l+h} \rightarrow \ldots \rightarrow \alpha_b, t) = \\ P(\Delta t(\alpha^l \rightarrow \ldots \rightarrow \alpha_{l+h-1} \Rightarrow \alpha_{l+h} \rightarrow \ldots \rightarrow \alpha_b) > t) \end{aligned} \tag{5.12}$$

and the confidence function based on Equation 5.10 can be determined as:

$$conf(\alpha^l \to \ldots \to \alpha_{l+h-1} \Rightarrow \alpha_{l+h} \to \ldots \to \alpha_b, t) = \quad (5.13)$$
$$\big(1 - S(\alpha^l \to \ldots \to \alpha_{l+h-1} \Rightarrow \alpha_{l+h} \to \ldots \to \alpha_b, t)\big) \prod_{f=l}^{l+h} conf(\alpha^f \to \alpha_{f+1})$$

The next section presents how the resultant survival functions can be statistically analyzed.

5.2.5 Log-rank test-based comparison of the survival functions

The proposed post-processing method can be extended to explore the impact of the antecedent part on the consequent part across different sets of sequences, or vice versa. This approach can help determine whether there is a significant difference in the evolution of two distinct consequent events. If the two populations are equal, they can be handled simultaneously. This hypothesis can be checked by the log-rank test-based comparison of survival functions.

The log-rank test [128] is a popular chi-square-based method to assess the statistical differences between two or more survival functions. This non-parametric test is advantageous because it does not require assumptions about the shape of the curves [32]. The null hypothesis posits that there are no significant differences between the examined groups. If the null hypothesis is rejected, it indicates differences among the groups. However, to precisely determine which groups differ from each other, pairwise comparisons need to be conducted. Nonetheless, performing numerous statistical tests can elevate the risk of a type I error, necessitating the need for adjustment [129].

Moreover, traces can be characterized by discrete explanatory variables that may define subgroups. For the different subgroups, the log-rank test can be used to test the degree of similarity between the temporal occurrence of consequences that follow the same antecedent sequence.

5.2.6 Evaluation of the confidence intervals of the time-dependent rule confidences by bootstrapping

The output of the model is frequently subject to uncertainty, making it challenging to precisely describe the real values. Therefore, the range of possible values, in other words the confidence interval, is essential to be determined. The uncertainty primarily arises from two key sources: first, from identifying the frequent sequences and, secondly, from estimating the empirical distribution functions. Determining the confidence interval using exact approaches can be extremely challenging in such an integrated method. As a solution, this study recommends an alternative approach known as the bootstrapping method. The bootstrapping method is a resampling technique utilized to estimate statistics for a population by repeatedly sampling a replacement dataset [130]. An important advantage of bootstrapping is that it requires minimal additional effort to implement [131].

The method randomly selects $N_B < N$ input sequences, which are placed in a subset called a bootstrap. The algorithm is executed on this selected data, and then the data is replaced. These steps are repeated M_B times, meaning that a given input sequence may be an element of multiple bootstraps, and each sample maintains the same size (N_B). The bootstrapping method determines M_B bootstrap confidence functions for every association rule, which needed to be aggregated to an upper and lower limit. For this case, the percentile-based method is applied [131], which determines the $\hat{\alpha}$th and $1 - \hat{\alpha}$th percentile (lower and upper confidence bounds) of the bootstrap confidence functions at every point in the domain of confidence functions, where $\hat{\alpha}$ is the significance level.

5.2.7 The steps of the algorithm and their time and space complexities

In this section, the crucial execution steps of the algorithm are meticulously outlined, with each phase of the analytical process intricately detailed. Moreover, the time and space complexities are also discussed. The enumerated steps, from pattern extraction to the derivation of confidence intervals through sophisticated bootstrapping. These steps can be summarized as follows:

1. **Read patterns from input sources:** Read the initial patterns from the input sources, possibly stored in an Excel or text file.

2. **Convert input to sequential representation:** Transform the input file into a trace, arranging the data in a format suitable for sequence mining.

3. **Execute a sequence mining algorithm:** Use a sequence mining algorithm (e.g., CM-SPAM) to analyze the sequential data and extract frequent sequences.

4. **Process the resultant frequent sequences:** Process the output of the sequence mining algorithm by extracting the rules, transaction IDs, supports, and confidences. Store this information for further analysis.

5. **Define the interested rule antecedent parts and find association rules:** Identify association rules by finding the predetermined rule antecedent parts within the frequent sequences. This process also brings attention to potential consequent parts, providing insights into future events.

6. **Gather survival times related to selected rules:** Gather survival times associated with the association rules, based on transaction IDs.

7. **Obtain survival function based on survival times:** Compute the survival function using Kaplan-Meier estimator based on the gathered survival times. This process provides the survival function of elapsed time between rule antecedent and consequent parts.

8. **Obtain confidence function based on survival function and confidence:** Derive confidence functions by multiplying the survival functions and confidence values linked to each association rule, representing the time-dependent confidence function between rule antecedent and consequent parts.

9. **Execute bootstrapping to get confidence intervals:** Perform bootstrapping, a statistical resampling technique, to estimate confidence intervals for the confidence functions. This approach provides a robust assessment of the uncertainty associated with the estimated time-dependent confidence functions.

The overall complexity of the method depends on multiple factors characterized by the database, the hyperparameters and the predefined rule antecedent parts. These

factors encompass the number of unique items (M), the number of traces (N), the number of frequent sequences (N_α) and the quantity of discovered association rules ($N_{\alpha^l \to \alpha^b}$). To provide comprehensive description of the complexity, the average length of traces (L_τ), average length of frequent sequences (L_α) and the average support of association rules ($L_{S,\alpha^l \to \alpha^b}$) should also be considered.

In order to analyze complexity, every step of the algorithm should be examined and understand how they are influenced by the dependent factors. The contributing components affects the time and space requirements similarly, so their description can also be approached with the same idea. However, the users have the autonomy regarding the space complexity to retain only the necessary variables to liberate additional space. In contrast, there are no comparable shortcuts for the time complexity available.

The process of reading the database in Step 1 requires $O_1(NL_\tau)$ resources, as the algorithm scans through N traces, each with an average length of L_τ. Similarly, the transformation of the input sequences into a sequential database in Step 2 has comparable resource requirements, indicated by $O_2(NL_\tau)$. In Step 3, the extraction of frequent sequences presents a somewhat more intricate complexity, which is highly dependent on the mining technique chosen. Essentially, the complexity of frequent sequence extraction can be described as the product of the number of frequent sequences and the resources needed to process each pattern, denoted as $O_3(N_\alpha)O_3(N)$ [132]. Unlike more sophisticated analytical analysis, experimental studies provide detailed information on the complexity of various approaches [133].

The processing of resulting frequent sequences in Step 4 demands $O_4(N_\alpha L_\alpha)$ resources. This complexity can be described in a manner similar to Steps 1 and 2. The algorithm scans through N_α frequent sequences, each having an average length denoted as L_α. Similarly, the discovery of association rules within the frequent sequences in Step 5 necessitates a rescan of the rules, leading to a resource requirement of $O_5(N_\alpha L_\tau)$. In the worst-case scenario, the number of association rules corresponds to the number of frequent sequences, as follows:

$$N_{\alpha^l \to \alpha^r} = N_\alpha = 3^M - 2^{M-1} - 1 \tag{5.14}$$

The collection of survival times in Step 6 involves a subtraction operation along the association rules, resulting in a complexity of $O_6(N_{\alpha^l \to \alpha^b})$. In Step 7, the computation of the survival function is straightforward to express in terms of complexity.

The Kaplan-Meier method exhibits linear behavior based on the input samples of the algorithm and is directly influenced by the average support of association rules, manifesting as $O_7(L_{S,\alpha^l \to \alpha^b})$. Moving on to Step 8, obtaining the confidence function based on the survival function and confidence is similar to Step 6, necessitating a multiplication operation for every association rule. Therefore, the complexity for this step is $O_8(N_{\alpha^l \to \alpha^b})$. The overall complexity without considering the bootstrapping method is the sum of the individual steps, and it can be expressed as::

$$O = \sum_{i=1}^{8} O_i = \tag{5.15}$$
$$O_1(NL_\tau) + O_2(NL_\tau) + O_3(N_\alpha)O_3(N) + O_4(N_\alpha L_\alpha) +$$
$$O_5(N_\alpha L_\tau) + O_6(N_{\alpha^l \to \alpha^b}) + O_7(L_{S,\alpha^l \to \alpha^b}) + O_8(N_{\alpha^l \to \alpha^b})$$

The next section presents an overview of related works, providing insights into existing literature and contextualizing our approach within the broader research landscape.

5.2.8 Related works

The objective of this section is to provide context for the proposed methodology by outlining its foundational principles and related research directions. The proposed framework aims to integrate frequent sequence mining and survival analysis, bringing the element of time into sequence analysis.

Frequent sequence mining, presented in Section 5.2.1, is a widely used technique with a broad spectrum of potential applications [134], which still has promising challenges [16]. This technique has been applied to generate systematic test cases for software developers [135], discover patterns in traffic data [136], and even enhance image compression efficiency through integrated clustering and sequence mining algorithm [137]. Traditional sequence mining algorithms primarily focus on the order of event occurrences, often overlooking the associated time stamp information, which can be extracted based on Section 5.2.2. Nevertheless, the incorporation of time information can offer valuable insights into the dynamics

of event sequences [138]. Consequently, frequent sequence mining has evolved to include time stamp information, providing a method known as time-stamped frequent sequence mining [139].

In the time-stamped frequent sequence mining algorithm, time stamps are considered as input parameters alongside event IDs, enabling the simultaneous handling of temporal information during frequent sequence mining. However, time-stamped sequence mining methods may face challenges in identifying meaningful patterns, especially when events are sparsely distributed over a long time window. As a result, time stamps may lose their temporal information because the frequency of an event at a specific time stamp does not meet the minimum support threshold. Suppose there is a database containing time-stamped input sequences, and in a frequent scenario, the event A occurs at time instant 0 followed by the event B at time instant 10. However, in cases where the event B occurs at a different time stamp, such as 9 or 11, this information could be lost if these time-stamped input sequences do not meet the minimum support threshold, resulting in the loss of valuable time stamp information. Time-stamped frequent sequence mining can also be extended to operate within specific time intervals [140], potentially overlooking important temporal nuances that fall outside of those intervals, such as event B being followed at time instant 11.

These approaches may encounter limitations when a process unfolds over a long duration, as a particular event may occur at various time stamps, albeit infrequently. In healthcare and stock market analysis [141], time-stamped data plays a crucial role. For instance, monitoring disease progression and treatment effectiveness heavily relies on analyzing patient health data with precise time stamps. Conversely, in text classification tasks like social media sentiment analysis and basic text categorization, time information is often less critical, with the primary focus on content and sentiment rather than precise timestamps [142].

The proposed methodology aims to rectify the inherent drawback observed in time-stamped frequent sequence mining algorithms, which is associated with the loss of temporal information. By ensuring the preservation of complete temporal information, the anticipated outcome of the investigated case study becomes notably more precise. For this reason, this paper proposes a technique based on frequent sequence mining and survival analysis. The components have individual contribution to the complete outcome, where frequent sequence mining captures the occurrence of relevant events, while survival analysis provides comprehensive

temporal information between the captured events, as presented in Section 5.2.3 and 5.2.4.

Integrating survival analysis with frequent sequence mining, as presented in Section 5.2.3 and 5.2.4, is a promising idea, as machine learning models can significantly enhance the efficiency of survival analysis [20]. However, there is a gap in the literature concerning publications that address this fusion approach. Extensive research has been conducted in the field of student dropout [143], where student activity was represented by sequential engagement states in various courses. Hidden Markov models were employed to cluster engagement trajectories, which were further analyzed using frequent sequence mining to identify prevalent subsequences and their supports. While survival analysis was utilized to identify the overall survival function of dropout for students within a dedicated cluster of trajectories but not between the intermediate states. Another study introduced an approach that incorporates frequent sequence mining to determine standard treatments for cancer patients [144]. The frequent sequences were used as input, along with the clinical data, for classification algorithms that predict survival outcomes. However, traditional survival analysis methods are not considered in this context. Typical carrier paths have also been identified by sequence mining, and survival analysis has revealed the duration in weeks that an employee spent in a specific role before being assigned another role [145]. Nevertheless, survival functions were not estimated at the given event transitions.

The probability of occurrence of a consequent event is considered to be time-dependent, and its baseline function is determined using a nonparametric survival analysis method known as the Kaplan-Meier estimator [32], as presented in Section 5.2.3 and 5.2.4. Moreover, the survival analysis can be enhanced by explanatory variables, allowing the calculation of subgroup distributions based on specific features. For the analysis of the subgroups with the Kaplan-Meier estimator, the samples need to be separated according to the labels formed by the explanatory variables. While forming subgroups based on discrete variables is a straightforward task, when dealing with continuous variables, it is advisable to employ discretization through clustering algorithms. The integrated survival analysis and expectation-maximization-based clustering presented in Chapter 3 stands out as a suitable option for this purpose. Alternatively, a Cox semi-parametric regression model can also be identified, which can handle the explanatory variables, but certain assumptions, such as the proportional risk assumption, must be verified [146].

Conventional survival analysis methods typically assume that only a single outcome can occur. However, when multiple potential outcome exist, the handling of competing risks is required to avoid biased results [147]. The proposed integrated framework deals with this issue, since the frequent sequence mining separates the samples to cases that have only one possible outcome at the same time. Moreover, survival analysis can accommodate cases by censoring [148], where a given event has no more follow-up data [148]. In that case, the traditional Kaplan-Meier estimator fails, and cumulative incidence functions [35] need to be calculated instead. The confidence interval of the resultant confidence functions has two sources of uncertainty: the support of the rules and the survival function. Estimating these confidence intervals using analytical methods can be a very challenging task. Therefore, an alternative approach known as the bootstrapping method is used [130], as presented in Section 5.2.6.

The proposed approach aims to rectify a significant drawback of existing applications related to the loss of crucial temporal information. The precision of predicting outcomes is significantly enhanced by ensuring a thorough provision of temporal data. This enhancement is facilitated through a novel technique based on the integration of frequent sequence mining and survival analysis. Event occurrences are adeptly captured using frequent sequence mining, and comprehensive temporal data between these events are supplied through survival analysis.

To handle large number of unique events, that are poorly distributed, the sequential rule mining is applied. Sequential rule mining stands out as a potent technique that capitalizes on such sequential databases to spotlight meaningful continuations of event itemset. The method is presented in the next section.

5.3 Description of the sequential rule mining - based survival analysis method

The proposed method integrates sequential rule mining with survival analysis, creating a comprehensive framework for capturing the temporal dynamics of event sequences. The algorithm is summarized in Figure 5.3. The method utilizes training input sequences to extract sequential rules using minsupp and minconf hyperparameters. The block continues by identifying trace IDs that support these sequential rules, allowing for the extraction of occurrence times. Utilizing these

occurrence times, the algorithm calculates elapsed times and proceeds to estimate the survival function through the Kaplan-Meier method. Leveraging the empirical distribution and rule confidence values, it then computes the confidence function for the provided rules. The outcomes comprise trained elements: relevant sequential rules and their associated confidence functions.

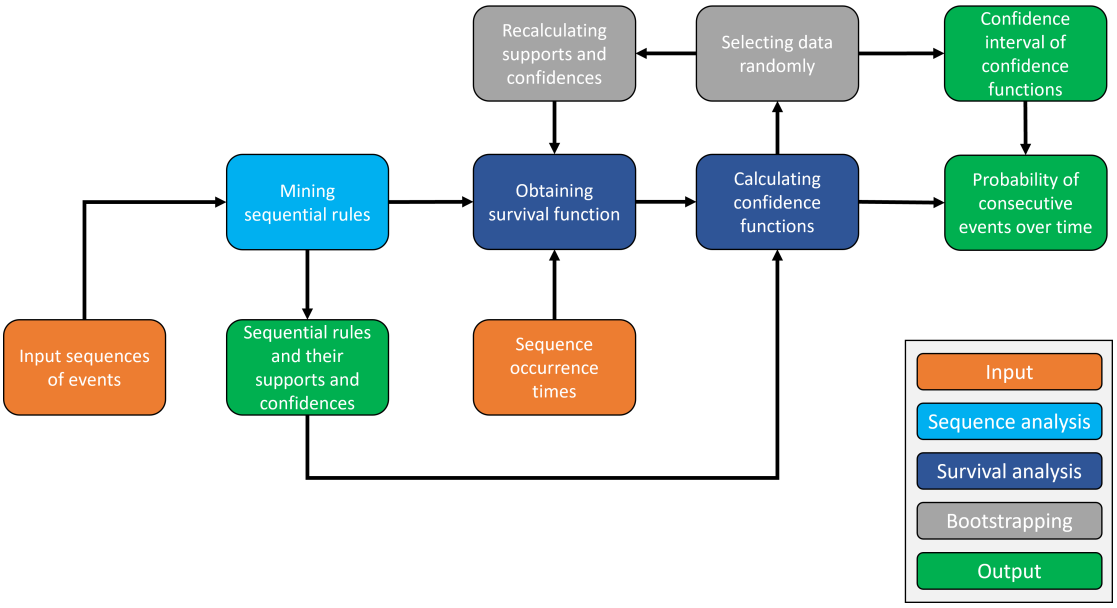

FIGURE 5.3: Steps of the algorithm divided into model training and prediction parts. The method uses training input sequences to extract sequential rules. Utilizing these occurrence times, the algorithm calculates elapsed times and proceeds to estimate the survival function through the Kaplan-Meier method. Leveraging the empirical distribution and rule confidence values, it then computes the confidence function for the provided rules. While this figure may resemble Figure 5.1, the difference lies in the sequence analysis part. This method employs sequential rule mining instead of frequent sequence mining.

5.3.1 Formulation of the sequential rule mining problem

From trace ψ_i, $X = \{(x_1, \tau_{i,1}), \ldots, (x_{b_x}, \tau_{i,b_x})\}$ and $Y = \{(y_1, \tau_{i,1}), \ldots, y_{b_y}, \tau_{i,b_y}\}$ unordered set of events can be selected, where $b_x = |X|$ and $b_y = |Y|$ and $x_{b_x} \in E$ and $y_{b_y} \in E$. A sequential rule is defined as an ordered association between two sets of events, denoted as $X \Rightarrow Y$, where X stands for antecedent set of events, while Y denotes consequent set of events [123]. Sequential Rule $X \Rightarrow Y$ means, that every event in X must occur before every event in Y. If there is a trace ψ_i characterized by this property, one says that sequential rule $X \Rightarrow Y$ is supported by the trace ψ_i. This can be formulated mathematically as follows: trace ψ_i supports the sequential rule $X \Rightarrow Y$, if $X, Y \subseteq \psi_i$, such that $X \cap Y = \emptyset$, $X, Y \neq \emptyset$ [149] and the condition $\max\limits_{(x,\tau_i)\in X}(\tau_i) < \min\limits_{(y,\tau_i)\in Y}(\tau_i)$ holds, indicating that the latest event in X must occur earlier than the earliest event in Y.

Based on the timestamps, the elapsed time between the antecedent set of events and consequent set of events can be calculated with the next formula:

$$\Delta t_i(X \Rightarrow Y) = \max_{(y,\tau_i)\in Y}(\tau_i) - \max_{(x,\tau_i)\in X}(\tau_i) \tag{5.16}$$

To understand how a trace supports a sequential rule and the concept of elapsed time, consider Figure 5.4. In this figure, letters represent events, and numbers denote timestamps. For instance, the sequential rule $\{A, B, C\} \Rightarrow \{D, E\}$ implies that if events A, B, and C occur in any order, events D and E will appear in a later timestamp, also in any order. Take Trace 1 as an example: here, the latest event of the antecedent part occurs at timestamp 12, while the earliest event of the consequent part occurs at timestamp 15. Consequently, the condition $\max\limits_{(x,\tau_i)\in X}(\tau_i) < \min\limits_{(y,\tau_i)\in Y}(\tau_i)$ holds, indicating the sequential rule $\{A, B, C\} \Rightarrow \{D, E\}$ is supported by Trace 1. The elapsed time of Trace 1 respecting sequential rule $\{A, B, C\} \Rightarrow \{D, E\}$ can be calculated as follows: $\Delta t_1(\{A, B, C\} \Rightarrow \{D, E\}) = 19 - 12 = 7$. Following this logic, Trace 2 also supports the sequential rule $\{A, B, C\} \Rightarrow \{D, E\}$, while Trace 3 does not.

The support represents the proportion of traces in which the sequential rule is observed, providing insight into the prevalence of the rule. The support can be expressed by the next equation [150]:

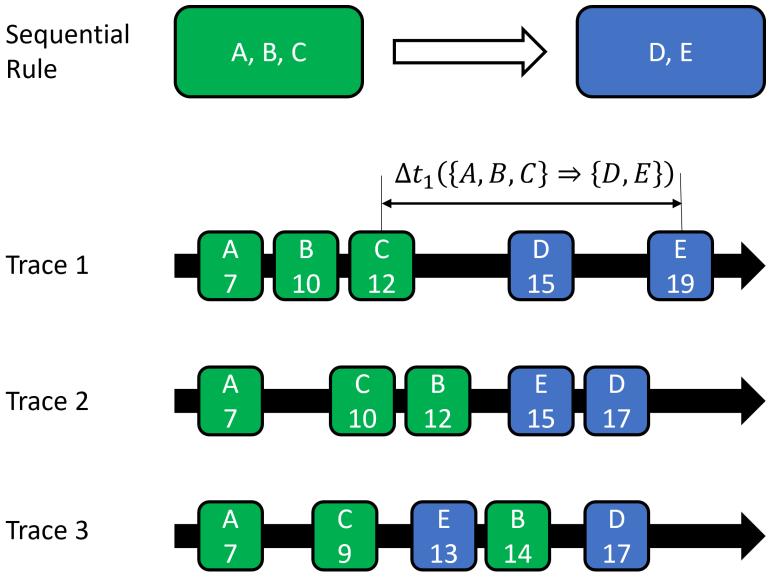

FIGURE 5.4: An example of the definition of support and elapsed time. The sequential rule $\{A, B, C\} \Rightarrow \{D, E\}$ means that if A, B and C occurs in any order, then D and E will appear in a later timestamp also in any order. Trace 1 and 2 support the Sequential rule $\{A, B, C\} \Rightarrow \{D, E\}$, but Trace 3 not

$$supp(X \Rightarrow Y) = \frac{\left|\{\xi \in \Psi : X, Y \subseteq \xi, \max_{(x,\tau_i)\in X}(\tau_i) < \min_{(y,\tau_i)\in Y}(\tau_i)\}\right|}{N} \tag{5.17}$$

On the other hand, there is another metric called confidence, which measures the conditional probability of the consequent event set given the occurrence of the antecedent event set. The confidence can be expressed by the next equation [150]:

$$conf(X \Rightarrow Y) = \frac{supp(X \Rightarrow Y)}{supp(X)} \tag{5.18}$$

A high confidence value indicates a strong likelihood of the consequent events following the antecedent events within traces. The strength of a sequential rule is measured using two main metrics: support and confidence. A sequential rule is considered relevant if both support and confidences exceed a minimum threshold such as $supp(X \Rightarrow Y) \geq minsupp$ and $conf(X \Rightarrow Y) \geq minconf$, respectively [124]. The next section introduces how the temporal relation can be determined based on survival analysis.

5.3.2 Kaplan-Meier empirical survival function-based analysis of the sequential rules to determine time-dependent confidences

This study considers the support time-dependent $supp(X \Rightarrow Y, t)$, as well as the confidence $conf(X \Rightarrow Y, t)$. The main contribution is that the time-dependent support is determined by the multiplication of two terms: (I.) the time-independent support and (II.) the baseline probability distribution function of the elapsed time between the antecedent set of events and consequent set of events. The confidence can be derived from the support. The formulas can be expressed based on Equations 5.17 and 5.18, respectively, by the next equations:

$$supp(X \Rightarrow Y, t) = \big(1 - S(X \Rightarrow Y, t)\big) supp(X \Rightarrow Y) \tag{5.19}$$

$$\begin{aligned} conf(X \Rightarrow Y, t) &= \frac{supp(X \Rightarrow Y, t)}{supp(X)} \\ &= \frac{\big(1 - S(X \Rightarrow Y, t)\big) supp(X \Rightarrow Y)}{supp(X)} \\ &= \big(1 - S(X \Rightarrow Y, t)\big) conf(X \Rightarrow Y) \end{aligned} \tag{5.20}$$

where $S(X \Rightarrow Y, t)$ stands for the survival function of the elapsed time between the antecedent set of events and consequent set of events. In this study, the survival function is determined by using the Kaplan-Meier estimator of the survival analysis. This method provides an empirical survival function causing a discrete stepped shape. The calculation can be executed through the next equation [35]:

$$S(X \Rightarrow Y, t) = \prod_{f:\Delta t_f \leq \Delta t} \left(1 - \frac{m_f}{N_f}\right) \tag{5.21}$$

where N_f represents the number of events that have not occurred at time difference Δt_f, while m_f is the number of events that occurred between periods of time differences Δt_{f-1} and Δt_f. It is important to acknowledge that within a dataset, there may exist multiple relevant sequential rules that adhere to these criteria.

5.4 Application for the analysis of patient pathways in hospitals with frequent sequence mining-based survival analysis method

This section describes how the proposed method can be practically applied to analyze medical data from a hospital information system. The method focuses on the diagnosed disorders (event) of the patients and utilizes the date of diagnosis (timestamp) for each patient (trace). The primary objective is to identify frequent sequences of disorders, providing hypotheses for the possible chains of disorders. The association rules are formed for the antecedent part, which is represented as current diseases of the patient. The objective of the analysis is to identify relevant consequent parts of the antecedent events, which correspond to possible future events. Additionally, the elapsed time until the occurrence of the consequent event can be extracted from the database, and its empirical distribution can be estimated using the Kaplan-Meier estimator. The confidence function is calculated from the distribution function and the confidence of the given frequent sequence.

The method was developed in a MATLAB environment. The dataset used for the results is confidential for privacy reasons. However, the MATLAB code is available on GitHub, and a presentation is provided using synthesized data to illustrate the application of the methodology. The files can be reached via: `https://github.com/CsalodiR/Time_dependent_sequential_pattern`.

The description of the data is presented in Section 5.4.1, the frequent sequence mining and the formation of association rules are introduced in Section 5.4.2, the analysis of time dependence is presented in Section 5.4.3, and finally a brief discussion is presented in Section 5.4.4.

5.4.1 Description of the data

The case study utilizes deidentified electronic health records from a Hungarian hospital from January 1, 2006, to November 25, 2019. The analysis includes patients whose initial data were recorded between January 1, 2006, and December 31, 2018. The medical conditions of the patients are represented using the International Classification of Diseases (ICD-10), and their gender was also recorded as

an explanatory variable. The deidentification of the patients was ensured through using a simple sequence number that cannot be decrypted or linked to any personal information. Notably, the database does not contain any censored samples, meaning that follow-up information is complete for all patients.

The input sequences for the analysis were constructed based on the chronological order of ICD-10 codes assigned to each patient. Prior to conducting the analysis, several data pre-processing steps were performed. Firstly, although a disorder can be recorded multiple times, only the initial occurrence was considered to prevent duplicated sequence items. This decision was made to align with the nature of the case study and the research objectives. However, the proposed method can accommodate multiple occurrences as well. Secondly, to reduce potential data bias, patients who were examined only once were excluded from the investigation. As a result, the study included disease sequences from a total of 13,299 patients. Both the number of records per day and the number of examined patients per day exhibited a uniform distribution.

5.4.2 Frequent sequence mining and association rules

In this study, frequent sequences were explored to analyze typical sequences of patient diseases. Although various algorithms can be used for mining frequent sequences, this study considered the CM-SPAM algorithm was applied, which identifies all frequent sequences faster than the usual SPAM algorithm [120].

The objective was to identify all possible frequent sequences of diseases with a minimum support threshold of 0.05%. Using this specific hyperparameter configuration, the algorithm identified 145,285 frequent sequences. Although the selected minimum support threshold may seem low, even with a higher value of 0.25%, the maximum length of the frequent sequences remained narrow. Considering that there are a total of 1730 unique ICD-10 codes, patient disease patterns tend to be highly individualized, rendering long sequences unable to become frequent. However, employing a lower minimum support value widens the range of items considered as frequent sequences, thereby enhancing the robustness of the analysis and increasing the complexity. These symptoms are indicative of the necessity of employing a sequential rule mining algorithm. The distribution of sequence lengths is depicted in Figure 5.5.

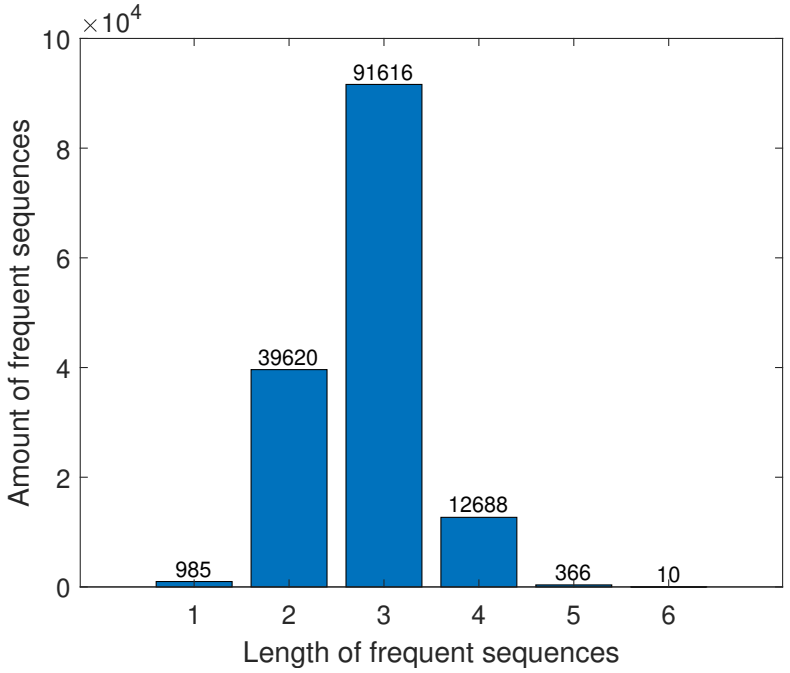

FIGURE 5.5: Histogram of the length of frequent sequences. Disease patterns of patients tend to be highly individualized, rendering long sequences unable to become frequent. These symptoms are indicative of the necessity of employing a sequential rule mining algorithm.

The frequent sequences can be broken down into antecedent and consequent parts by constructing association rules. The events in the antecedent and consequent parts are considered previously diagnosed and possible future disorders, respectively. The goal is to identify all frequent sequences characterized by the given antecedent parts for alerting patients about potential future disorders. Suppose a patient is suffering from hypertension (*I10*). Hypertension can be considered as an antecedent event. In that case, the goal is to identify $\alpha = \langle I10 \rightarrow \alpha^2 \rangle$ frequent sequences, where α^2 gives the consequent part of the sequences. Association rules are established to select all frequent sequences that commence with *I10*. These frequent sequences can be analyzed based on their confidence values to reveal the most likely continuations of the disease of *I10*. A total of 10,456 such frequent sequences are identified. The lengths of these frequent sequences with the antecedent part *I10* are visualized in Figure 5.6 and the top ten possible frequent sequences in this scenario are summarized in Table 5.3.

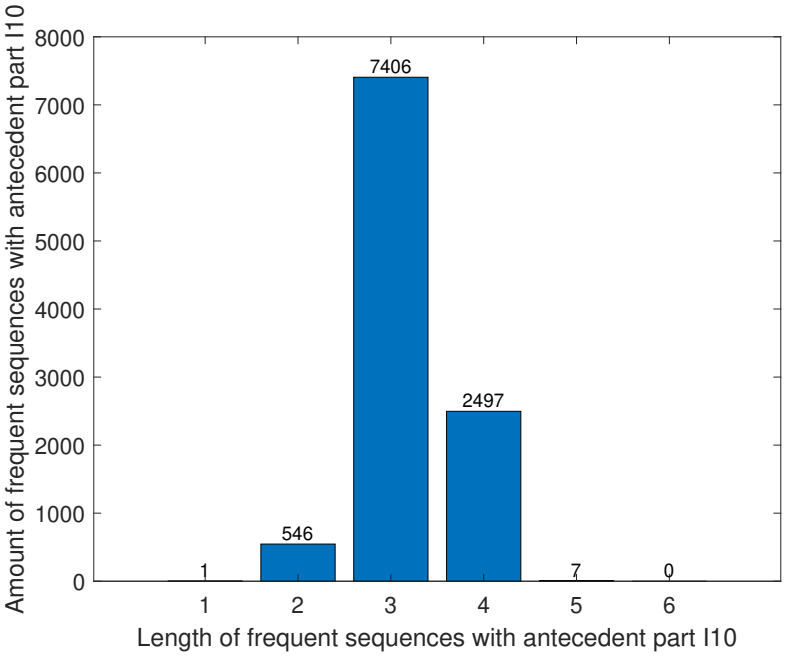

FIGURE 5.6: Histogram of the length of the frequent sequences with antecedent part *I10*. This histogram has the same tendency as in case of all frequent sequences.

5.4.3 Analyzing the time dependence of the rule confidences

The elapsed time between two disorders can significantly influence the probability of appearance of a disease. Therefore, understanding its temporal distribution can add valuable information to the prediction. The proposed post-processing methodology can determine how the probability of developing a disorder changes over time. This confidence function can be determined for each rule based on Equation 5.10. For instance, consider the prediction presented in Section 5.4.2 related to the antecedent event of hypertension ($I10$). A very potential outcome is the development of Type 2 diabetes mellitus ($E11$) [151].

The elapsed times between $I10$ and $E11$ $\left(\Delta t_i(I10 \rightarrow E11)\right)$ can be calculated for each patient based on the timestamps. The survival function of the elapsed times $\left(S(I10 \rightarrow E11, t)\right)$ can be determined by the Kaplan-Meier method. The confidence function $\left(conf(I10 \rightarrow E11, t)\right)$ can be computed by multiplying the distribution function by the confidence of $\langle I10 \rightarrow E11 \rangle$. The confidence interval for this scenario can also be obtained using the proposed bootstrapping method. In this study, bootstrapping was performed 150 times, with 3000 randomly selected patients in each iteration. The confidence level ($\hat{\alpha}$) was set at 5%. The confidence function for $\langle I10 \rightarrow E11 \rangle$ can be seen in Figure 5.7.

TABLE 5.3: The ten most frequent consequent disease patterns following hypertension (*I10*). Only one consecutive event is found in the top ten predicted disorders.

Frequent sequence	English name for the consequent part	Confidence
I10 → *J18*	Pneumonia, unspecified organism	0.2024
I10 → *Z00*	Encounted during a general examination without complaint, suspected or reported diagnosis	0.1853
I10 → *R10*	Abdominal and pelvic pain	0.1820
I10 → *E78*	Disorders of lipoprotein metabolism and other lipidemias	0.1594
I10 → *I25*	Chronic ischemic heart disease	0.1560
I10 → *D64*	Other anemias	0.1471
I10 → *I50*	Heart failure	0.1453
I10 → *E11*	Type 2 diabetes mellitus	0.1305
I10 → *W01*	Fall on same level from slipping, tripping and stumbling	0.1271
I10 → *I20*	Angina pectoris	0.1201

The explanatory variables divide the population into subgroups to analyze them separately. In this study, the gender of the patients was considered, as diseases may have different effect on men or women. The algorithm was executed for the two subgroups, and the resultant confidence functions can be seen in Figure 5.8. Although the curves suggest a notable difference between the two subgroups, the log-rank test did not reveal a significant difference between males and females in case of $\langle I10 \rightarrow E11 \rangle$.

The proposed post-processing method can be applied to generate the confidence functions for longer sequences of disorders. In such a case, the aim is to show the confidence functions of a sequence for all its continuations. Suppose a patient who already suffers from hypertension (*I10*) as an antecedent event, as presented in Section 5.4.2. The goal is to determine the confidence functions of the consequent sequence $\langle E11 \rightarrow E78 \rightarrow E14 \rangle$, where *E78* stands for lipoprotein metabolism and other lipidemias and *E14* denotes unspecified diabetes mellitus. In this case, the frequent sequences $\langle I10 \Rightarrow E11 \rangle$, $\langle I10 \rightarrow E11 \Rightarrow E78 \rangle$, and $\langle I10 \rightarrow E11 \rightarrow E78 \Rightarrow E14 \rangle$ should be examined. Support and confidence values of the given frequent sequences are summarized in Table 5.4. The confidence function for the exact times of occurrence of continuations is shown in Figure 5.9.

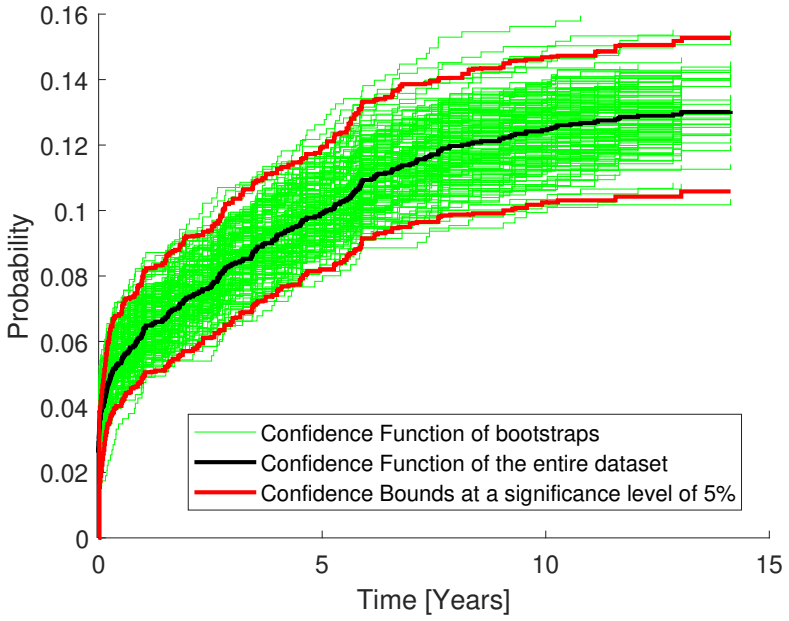

FIGURE 5.7: The confidence function for the case $\langle I10 \rightarrow E11 \rangle$ and the confidence bounds by applying bootstrapping method. Bootstrapping was executed 150 times and 3000 patients were randomly selected in every iteration. The confidence level ($\hat{\alpha}$) was set at 5%. This function increases rapidly to near zero Δt, because multiple disorders were diagnosed in one investigation.

TABLE 5.4: Support and Confidence values of rule $\langle I10 \rightarrow E11 \rightarrow E78 \Rightarrow E14 \rangle$ at the given continuations

Frequent sequence	**Support**	**Confidence**	**Cumulative confidence**
I10	2698	1	1
$I10 \Rightarrow E11$	352	0.1305	0.1305
$I10 \rightarrow E11 \Rightarrow E78$	72	0.2045	0.0267
$I10 \rightarrow E11 \rightarrow E78 \Rightarrow E14$	11	0.1528	0.0041

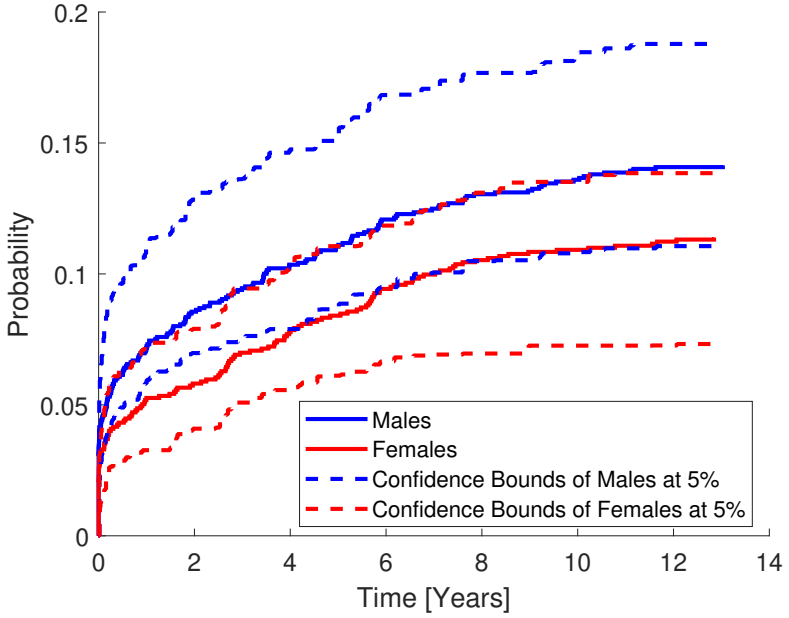

FIGURE 5.8: The confidence function of $\langle I10 \rightarrow E11 \rangle$ case separately for male and female genders with confidence bounds. Curves raise the possibility of a significant difference between the two subgroups. However, the confidence functions overlap their bounds.

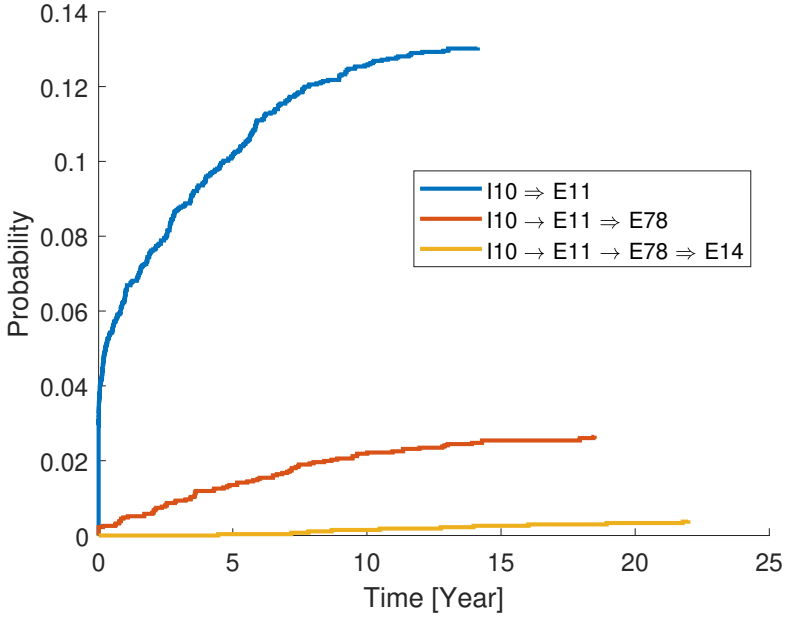

FIGURE 5.9: The confidence function for absolute occurrence times. It can be seen, that *E14* can occur after 4 years but also with minimal risks, if the frequent sequence $\alpha = \langle I10 \rightarrow E11 \rightarrow E78 \rightarrow E14 \rangle$ is considered.

5.4.4 Discussion

The proposed algorithm which integrates frequent sequence mining with survival analysis was applied to identify typical sequences of disorders. Data preprocessing played a crucial role in obtaining meaningful and unbiased results. By using ICD-10 codes as events, the case study demonstrated its predictive potential. Given the large number of individual diagnoses, a low minimum support threshold was chosen to capture diverse frequent sequences. With a minimum support of 0.05%, the resulting frequent sequences ranged in length from 1 to 6. The longest sequence, consisting of 3 events, included 91,616 patients, representing 63% of all frequent sequences discovered. These phenomenons suggest the need to use a sequential rule mining algorithm. In parallel with the execution, an analysis of time and space complexities was conducted, with a defined hyperparameter set and execution circumstances. The memory usage of MATLAB over the execution time can be observed in Figure 5.10, which also highlights the algorithmic steps explained in Section 5.2.7.

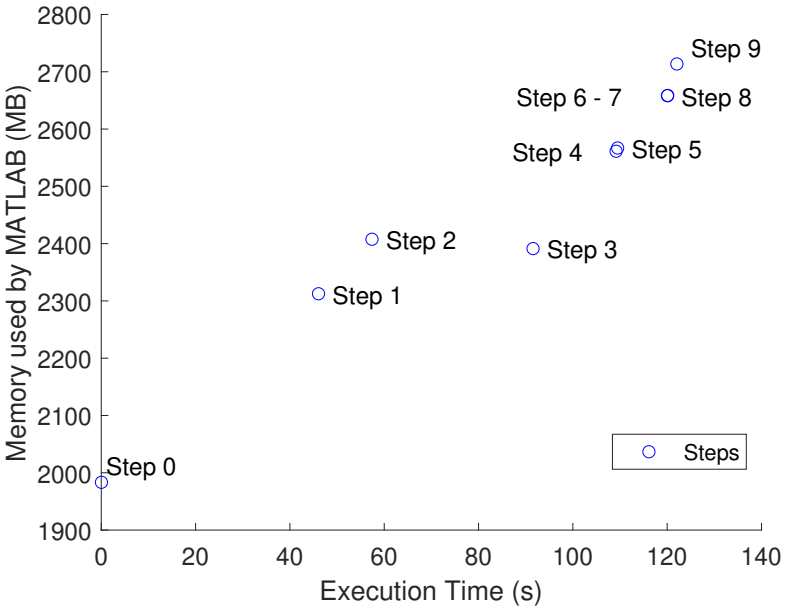

FIGURE 5.10: The memory usage of MATLAB over the execution time, highlighting the algorithmic steps explained in Section 5.2.7.

The analysis primarily focused on hypertension, a prevalent health issue among the Hungarian population [152]. The findings revealed a wide range of potential disorders associated with hypertension. Some of these associations indicated pathophysiological relationships, suggesting that certain events could contribute to the development of hypertension, such as diabetes [151], anemia [153], or other

cardiovascular diseases [154]. Hypertension can also increase the risk of pneumonia [155]. However, not all continuations implies pathophysiological links as in the case of falling on the same level from slipping, tripping, and stumbling *W01*, which also had high confidence value.

The study also examined type 2 diabetes mellitus, which is a frequent comorbidity of hypertension. The confidence function indicated a significant increase in the initial stage. This finding can be attributed to the fact that hypertension and diabetes screenings are often less intensive than recommended [156], leading to simultaneous detection of multiple diseases. The phenomenon of multimorbidity also plays a role, where the older an individual is, the more diseases they tend to have. In addition, both conditions can be attributed to a common factor such as obesity, which indicates a manifestation of metabolic disease.

The confidence function implied that the probability of getting diabetes in five years, given that the patient already has hypertension, is approximately 10%. Analysis of explanatory variables is also possible, but the statistical log-rank test should be used to prove the significant difference between the subgroups. The result of the test is $p = 0.6810$, which implies that there is no significant difference between men and women.

The study also presented multiple occurrences associated with diabetes, including lipoprotein metabolism disorders and other lipidemias, as well as unspecified diabetes mellitus. The chain of disorders analyzed was $\langle I10 \rightarrow E11 \rightarrow E78 \Rightarrow E14 \rangle$. The confidence functions indicate that the probability of having lipoprotein metabolism disorders and other lipidemias after diabetes within five years given that the patient has already hypertension is approximately 1.5%. Furthermore, the probability of developing unspecified diabetes mellitus after lipoprotein metabolism disorders and other lipidemias after diabetes within five years, since the patient has already hypertension, is approximately 0.05%. These findings underscore the complex nature of diabetes and its potential connections to various metabolic disorders.

The integrated frequent sequence mining and survival analysis methodology serves as a valuable tool for extracting comprehensive insights into a given process. The advantage of this approach lies in the specific functions of its components, each of which contributes differently to the output information. Frequent sequence mining captures the chronological order of typical events and the relative frequency of

event chains. Based on the frequent sequences, association rules can be established to effectively highlight antecedent and consequent events. These sets of events can be considered as current and future occurrences, and the probability of their appearance can also be calculated based on the rule support data. This value reveals the probability of a consequent event occurring, given its antecedent events have occurred.

The approach enables the precise estimation of the probability of an individual developing a particular disease when they already have a pre-existing condition. The probabilities accumulate as illustrated in Table 5.4. The probability of acquiring condition $E11$ is 0.1305, while the likelihood of sequentially acquiring $E11$ and then $E78$ is 0.0267. Acquiring $E11$, $E78$, and $E14$ in succession has a probability of 0.0041, given that $I10$ has already occurred. The confidence values do not change over time. However, there are cases, where the probability of consequent event occurring is low after one day but high after five years. The method posits that this temporal probability function approaches a limit comes from the rule confidence and the dynamics can be derived by survival analysis. Omitting any of the components of the method results in a loss of their respective contributions. Although, the dynamic probability of event continuations can also be calculated using the Kaplan-Meier estimator alone, the contribution of the frequent sequence mining lies in its ability to assist the user selecting critical events of interest, especially when dealing with a large number of unique events. Consequently, this method empowers the estimation of treatments and interventions aimed at mitigating adverse outcomes.

The proposed methodology can be developed into a decision support system, emphasizing predictions for a given antecedent disorder. However, its implementation should involve additional factors describing the condition of patients, such as laboratory data. Managing a substantial volume of unique events can make each trace highly individual. Furthermore, the distribution of laboratory and diagnostic events may be inappropriate, resulting in outcomes that predominantly reflect laboratory data. With 1540 unique ICD-10 codes for 148,954 records and 178 unique laboratory codes for 145,919 records, addressing this challenge requires an alternative approach capable of handling numerous unique, poorly distributed events, known as sequential rule mining, as presented in the next section.

5.5 Application for the analysis of patient pathways in hospitals with sequential rule mining-based survival analysis method

The motivation behind the development of this method stemmed from a medical application, wherein the objective is to predict future diseases based on data of Hospital Information Systems (HIS). The method centers its focus on diagnosed disorders of patients (events). This analysis hinges on the date of diagnosis (timestamp) for case history of each patient. The principal objective is to ascertain sequential rule within the occurrences of these health conditions (trace), thereby contributing insights into potential future developments. Medical diseases of patients are also codified using the International Classification of Diseases (ICD-10). Categorized laboratory events are also integrated, offering indications of diseases. These categories are stratified into four levels, denoting high, extremely high, low, and extremely low variances. Subsection 5.5.1 presents, how the model can be trained, while Subsection 5.5.2 introduces, how the trained data can be used to predict future disorders for patients.

5.5.1 Model training

An additional option was implemented to selectively choose relevant diagnoses and laboratory events based on expert knowledge. Among the 1719 distinct events, 120 unique events were selected based on their frequency and importance. The records of 8524 patients were deidentified using numerical trace IDs. A distinctive feature in comparison to the other scenario in the case study is that unique events may appear multiple times in the dataset. This was implemented to filter out irrelevant disorder continuations, such as high blood pressure causing leg fractures.

In medical scenarios, the events within a sequence may not necessarily unfold consecutively; rather, this occurs only when their timestamps are closely aligned. In such cases, events are deemed to occur simultaneously if they transpire within a narrow time interval. This consideration of events occurring within a confined temporal window adds an additional layer of intricacy to the analysis. Since multiple

disorders can be recorded for a patient at a given examination period, it is important to handle them simultaneously to avoid misconceptions that may arise. To address this, any events occurring in two weeks are treated together, e.g. when event $a_{i,1A}$ and $a_{i,1B}$ are occurred closely to each other, the resulting trace is represented as : $\psi_i = \langle\{(a_{i,1A}, \tau_{i,1}), \ldots, (a_{i,1B}, \tau_{i,1})\} \rightarrow \ldots \rightarrow \{(a_{i,qA}, \tau_{i,q}), \ldots, (a_{i,qB}, \tau_{i,q})\}\rangle$.

The study employed the Rule Growth Algorithm to mine sequential rules [123]. The goal was to discover meaningful patterns by establishing a minimum support threshold of 1% and a minimum confidence threshold of 10%. The choice of a relatively low support threshold was twofold: not only to spotlight infrequently occurring disorders, but also to ensure that even less common occurrences were brought to light. Furthermore, the decision to set the confidence threshold at 10% stems from the notion that rules meeting this criterion are more likely to possess significance. However, it is important to emphasize that the revealed sequential rules still required further screening and elimination. The reason for this filtration process lies in the medical context: predicting abnormal deviations in laboratory results is not pertinent to the objectives of predicting diagnoses. Through this hyperparameter configuration and filtering condition, the algorithm successfully revealed a total of 940 sequential rules. Ordering the confidences descending, the top ten relevant association rules can be seen in Table 5.5. The rule $\{E11, E78, I20\} \Rightarrow \{I10\}$ means that if a patient has $E11$, $E78$ and $I20$, then $I10$ will appear with the probability of 0.7769.

TABLE 5.5: The top ten relevant sequential rules, their supports and confidences in descending order based on confidence values.

Antecedent disease	**Consequent disease**	**Support**	**Confidence**
{E78, I20, I25}	{I10}	0.02	0.793
{E11, E78, I20}	{I10}	0.012	0.777
{E78, I10, I21}	{I25}	0.01	0.768
{E78, I20}	{I10}	0.025	0.755
{I20, Z95}	{I10}	0.012	0.752
{E78, I21}	{I25}	0.01	0.748
{E11, E78, I25}	{I10}	0.011	0.742
{E78, I25}	{I10}	0.026	0.739
{E78, Z95}	{I10}	0.01	0.738
{E11, I20}	{I10}	0.018	0.737
{I20, I25, Z95}	{I10}	0.01	0.73
{E78, I10, Z95}	{I25}	0.01	0.725
{E11, I20, I25}	{I10}	0.013	0.718
{E78, M16}	{I10}	0.012	0.713

In the field of medicine, understanding the probability of the occurrence of a disorder over time, say after five years, holds immense significance [157]. This insight equips healthcare professionals with the foresight to anticipate potential developments, enabling early interventions and informed patient care. To provide the time-dependent confidence function, the distribution function of the elapsed times and the rule confidence are multiplied. The resultant function shows the probability of a given disorder appearing, given that particular disorders have already been identified. The bootstrapping method is applied to provide the confidence function with the following parameters: the number of bootstraps was set to 200 and the number of samples to 3000 and the confidence level $\hat{\alpha}$ was set at 5%. The sequential rule $\{E78\} \Rightarrow \{I25\}$ is selected to present the results, which is the development of "chronic ischaemic heart diseases" from "disorders of lipoprotein metabolism and other lipidaemias". The support of the rule is 0.0285, while the confidence is 0.3. The confidence function with the confidence bounds and the individual bootstraps can be seen in Figure 5.11.

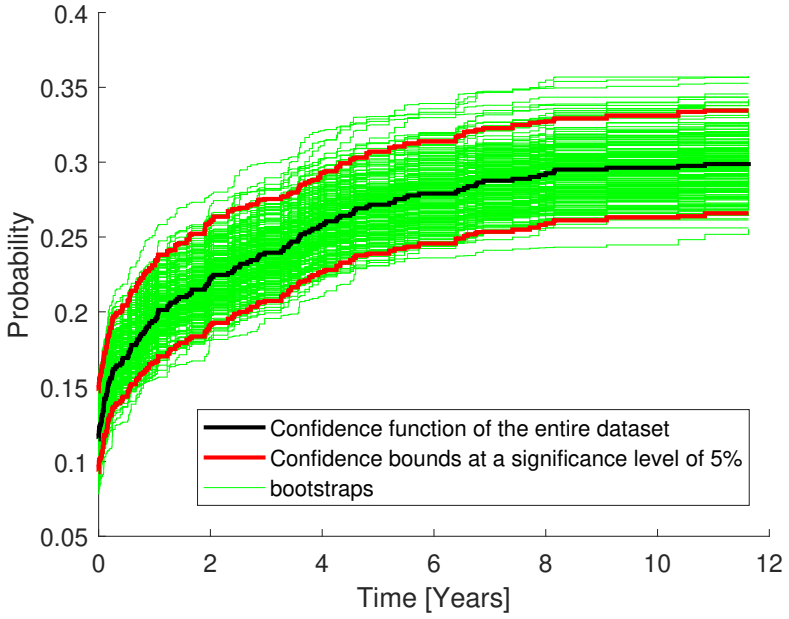

FIGURE 5.11: The confidence function of sequential rule $E78 \Rightarrow I25$ with the confidence bounds and individual bootstraps. The limit of the function is the rule confidence value. The initial probability is above 0.1 due to the identification of more disorders at the same time.

The confidence function indicates that the initial probability value is above 0.1, which means that in a significant proportion of the cases (almost 40%) these two disorders are identified at the same time. This founding can be explained by the fact, that Hungarians visit doctors rarely or only if they have significant symptoms. The confidence function also indicates that the overall chances of getting $I25$ is

limited to the rule confidence value 0.3. One can also conclude, that the chances of getting $I25$ in five years is about 0.27.

5.5.2 Model prediction

The extracted sequential rules along with their associated confidence functions have the potential to anticipate additional disorders within a patient. Consider a scenario where a patient is already diagnosed with $E11$ and $I10$, denoting "type 2 diabetes mellitus" and "hypertonia," respectively. The algorithm identifies all sequential rules where either $E11$ or $I10$, or both, are present as antecedent components. It is important to match the individual events or their combinations as well, as there could be potential future disorders stemming from only this single actual disease or specific combinations. Identifying instances where both disorders align may lead to a loss in predictive accuracy. The consequent part of these rules serves as the prediction. Matching the predicted events with the derived rules can cause redundancy in the prediction process. There may be instances where pre-existing disorders are predicted due to matching individual events or their particular combinations. Additionally, situations can arise where the consequent segment remains consistent, but the antecedent segments differ in terms of subset relations. In such cases, the most extensive antecedent segment takes precedence for prediction. The predictions of antecedent events $E11$ and $I10$ can be seen in Table 5.6. The temporal aspect is introduced by highlighting the probability of disease occurrence within a five-year span.

Table 5.6: The resultant predictions and supplementary data of antecedent events $E11$ and $I10$. The most relevant predicted disorder is $E78$ (Disorders of lipoprotein metabolism and other lipidaemias) with the probability of 0.2799.

Source	**Predicted disease**	**Support**	**Confidence**	**Confidence after five years**
E11, I10	E78	0.021	0.2799	0.2388
E11, I10	J18	0.02	0.266	0.225
E11, I10	I25	0.02	0.264	0.255
E11, I10	I50	0.018	0.237	0.208
E11, I10	R10	0.015	0.195	0.182
E11, I10	I20	0.014	0.19	0.17
E11, I10	E14	0.014	0.186	0.152
E11, I10	E10	0.012	0.162	0.136
E11, I10	I70	0.012	0.156	0.136
E11, I10	I25, I50	0.011	0.149	0.113

The results in the table encapsulates a significant aspect of medical comorbidity assessment. It highlights a compelling phenomenon wherein the simultaneous coexistence of two distinct disorders, each demonstrating comorbidity with a common third disorder, is shown to elicit a heightened level of confidence in predictive outcomes. This finding underscores the notion that the presence of these two disorders together carries a predictive weight greater than their individual contributions. These results are suitable for doctors for predicting diseases and decision support. The related web service is currently under development.

5.6 Summary of the chapter

This study introduced a method that integrates survival analysis and sequential pattern mining algorithms to enhance the information of frequent sequences by incorporating complete temporal dynamics. The study highlighted that traditional time-stamped frequent sequence mining approaches often faced challenges related to the loss of temporal information. However, the proposed method effectively addressed this limitation. Integration of survival analysis resulted in a more comprehensive and accurate representation of temporal patterns within frequent sequences. The chapter considered two different approach for sequential pattern mining: the frequent sequence and sequential rule mining.

In the first step, the raw data set was converted into input sequences. Sequential pattern mining algorithms were then applied to identify typical patterns in the database. The method further divided the mined rules into two subsequences: rule-antecedent and rule-consequent. This division facilitated the identification of future event occurrences, with rule-antecedent representing past events and rule-consequent representing future events. Association rules were then used to determine the relevant event continuations with the selected antecedent events, and the goal was to find their consequent events with high confidence value. These rules were ranked according to their confidence values, emphasizing the possible occurrence of events.

The time dependence of the events was established by considering the times of occurrences. The Kaplan-Meier estimator was used to determine the survival function, which described the elapsed time between two events. Subsequently, the confidence function was computed by multiplying the confidence of the sequential

patterns with the distribution function of the elapsed time between two events. This confidence function indicates the probability of the next event occurring, given that the relevant antecedent event has occurred. Furthermore, the method can be extended by incorporating explanatory variables, which allow for the analysis of subgroups. The confidence intervals of the confidence functions were also determined using the bootstrapping method, which helped assess significant differences between two groups. The presence of a significant difference was also confirmed through a statistical test known as the log-rank test.

The algorithm were presented based on frequent sequence mining, that explored the order of events. However, sequential rule mining was demonstrated as an alternative approach when handling a substantial volume of unique events that are poorly distributed. The method identified potential itemset of antecedent events and their corresponding itemset of consequent events, illuminating the associations between these occurrences. In this case, sets of antecedent events are explored that may occur in any order before another set of consequent events that also occurred in any order. Consequently, the permutation of potential outcomes becomes more limited. The confidence value of the rule indicates the probability of the occurrence of consequent events, given that antecedent events have already occurred.

The proposed methodology was applied to analyze the medical records of hospital patients, with the aim of predicting diseases that are likely to develop in the future based on previously known disorders. It should be noted that the methodology in the current form cannot identify causal relationships. In this case study, the focus was on predicting future disorders for patients who already suffer from hypertension. The time dependence of Type 2 diabetes mellitus was also determined, where a causal relationship has already been concluded based on medical researches. The gender of the patients was established as a categorical variable to find a significant difference between the development characteristics of men and women.The methodology was demonstrated for predicting multiple future disorders in hypertension patients, showcasing its potential utility.

Chapter 6

Conclusions

The motivation behind this thesis was introduced in Chapter 1. Survival analysis is a statistical methodology and serves as a pivotal tool across diverse fields for analyzing the time to event data. However, for a more comprehensive analysis aimed at delving into the root causes of inefficient operations, relying solely on this method proves to be a limited approach. Therefore, survival analysis should be applied in an integrated framework with machine learning algorithms. The thesis presented three algorithms that integrated survival analysis with machine learning techniques. Section 2 described the foundational concepts of survival analysis that were frequently applied in the description of proposed methods.

The first algorithm presents an integrated survival analysis and expectation - maximization - based clustering framework, that was presented in Chapter 3. This method identified clusters based on the similarity of survival times and explanatory variables. An expectation maximization estimated the parameters of Weibull distribution that represented the survival time, while simultaneously estimated the parameters of multivariate Gaussian distribution that represented the explanatory variables. The cluster memberships were represented using Takagi-Sugeno fuzzy rules, offering a framework to determine the operating domain of continuous variables. The method demonstrated versatility and is applied for categorizing continuous variables. The determination of the number of clusters was achieved through the use of the Akaike Information Criterion. The effectiveness of the proposed algorithm was demonstrated across diverse case studies, where student dropout rate, remaining useful life of Li-ion batteries, survival chances for patients

with prostate cancer and mortality rate per 100K population of countries related to the COVID-19 were estimated.

The second algorithm introduces an integrated survival analysis and frequent itemset-based association rule mining method that was presented in Chapter 4. This method estimated probability of competing risks that was determined at specific time instances using frequent itemset-based association rules. The proposed method operated within a discrete time domain. The survival was characterized by categorical explanatory variables at every time instance. The time-dependent categorical variables were conceptualized as triggering events, acting as precursors, initiating consequent events that signify the competing outcomes of the survival process. This approach identified relevant triggering events, that lead to a competing risk. The approach also segmented the dataset for subjects characterized by a sequence of frequent itemsets consisting of specific combinations of triggering events. The estimation of cumulative incidence function was based on the resultant rule supports and confidences. When focusing solely on consequence events, the baseline cumulative incidence function can be estimated. However, the cumulative incidence function can be delineated by sequence of specific triggering events. The effectiveness of the proposed algorithm was demonstrated by estimating the student dropout rate based on uncompleted subject patterns.

The third algorithm demonstrates an integrated survival analysis and sequential pattern mining framework, that was presented in Chapter 5. The study highlighted that traditional time-stamped frequent sequence mining approaches often faced challenges related to the loss of temporal information. Integration of survival analysis resulted in a more comprehensive and accurate representation of temporal patterns within frequent sequences. The time dependence of the events was established by considering the times of occurrences. The Kaplan-Meier estimator was used to determine the survival function, which described the elapsed time between two events. The algorithm were presented based on frequent sequence mining, that explored the order of events. However, sequential rule mining was demonstrated as an alternative approach when handling a substantial volume of unique events that are poorly distributed. The proposed methodology was applied to analyze the medical records of hospital patients, with the aim of predicting diseases that are likely to develop in the future based on previously known disorders.

Appendix A

Appendix

A.1 Sample curriculum of the chemical engineering study

Table A.1: Summary of the identifiers, names used in the study and the number of recommended semesters of subjects according to the sample curriculum

Subject ID	**Subject name**	**Recommended semester**
1	Material science	1
2	Introduction to chemical engineering	4
3	Biochemistry	3
4	Electronics	4
5	Electronics laboratory practice	4
6	Process design I.	4
7	Physics I.	1
8	Physics (problem solving practice)	1
9	Physics II.	2
10	Physics lab. Pract.	2
11	Physical chemistry I.	2
12	Physical chemistry II.	3
13	Laboratory practice in physical chemistry	3

Subject ID	**Subject name**	**Recommended semester**
14	Problem solving practice in physical chemistry	3
15	Process control	4
16	Machine elements and presentation	1
17	Process dynamics and control	4
18	Introduction to law	4
19	Corrosion Basics	4
20	Comprehensive exam in chemistry	5
21	Chemical analysis	3
22	Chemical analysis laboratory practice	4
23	Economics	1
24	Mathematical analysis I.	1
25	Mathematical analysis I. Practice	1
26	Mathematical analysis II.	2
27	Mathematical analysis I. Practice	2
28	Quality assurance	2
29	Industrial quality management	6
30	Effective technical communication	6
31	Effective technical communication practice	6
32	IT tools for effective technical communication	6
33	Engineering thermodynamics	3
34	Technical thermodynamics	3
35	Flow and heat engineering machines (lab. pract)	4
36	Technical fluid mechanics	3
37	Basic energetics for unit operations	2
38	Unit operations A	4
39	Unit operations B	4
40	Numerical mathematics	2
41	Statistics	2

Subject ID	**Subject name**	**Recommended semester**
42	Basics of radiation	1
43	Organic chemistry I.	2
44	Organic chemistry II.	3
45	Laboratory practice on organic chemistry	4
46	Computer science for engineers I.	1
47	Modeling of chemical processes	5
48	Modeling of chemical processes (laboratory practice)	5
49	Design of technological systems	6
50	Design project I.	6
51	Design project II.	7
52	Transportphenomena	3
53	Chemical process engineering laboratory practice	5
54	Chemical Engineering BSc Field Practice	7
55	Chemical process safety	6
56	Selected chemical technologies	5
57	Selected chemical technologies (laboratory practice)	5
58	Process design II.	5
59	Process design III.	6
60	General and inorganic chemistry	1
61	Problem solving in general and inorganic chemistry I.	1
62	Problem solving in general and inorganic chemistry II.	2
63	Laboratory practice in general and inorganic chemistry	2
64	Hydrocarbons and petrochemical technologies	5

List of notations

Chapter 2. ***Formalization of survival analysis***

$h(t)$ - hazard function

$\lambda(t)$ - cumulative hazard function

$S(t)$ - survival function

$F(t)$ - cumulative distribution function

$f(t)$ - probability density function

$P(T > t)$ - probability that a subject will survive beyond time instance t

N_f - number of event of interest that have not been occurred or censored at time instance t_f

m_f - number of event of interest occurred between periods of time instance t_{f-1} and t_f.

$\lambda_0(t)$ - baseline cumulative hazard function

N_Z - number of explanatory variables

$\boldsymbol{\gamma}$ - set of Cox regression model parameters

$\mathbf{Z}$ - vector of the relevant explanatory variables.

$\mathcal{L}(\boldsymbol{\Theta})$ - Likelihood function of Θ parameter set

Chapter 3. ***Integrated survival analysis and expectation-maximization-based clustering: a collection of case studies***

M - number of clusters

$p(j)$ - unconditioned cluster probability of jth cluster

$p(t|j)$ - probability that the survival event occurs at time instance t, given that the sample belongs to cluster j

θ_j - the jth scale parameter of Weibull distribution of survival times

β_j - the jth shape parameter of Weibull distribution of survival times

$\mathbf{x}^d$ - the vector of discrete variables

N_d - number of discrete variables

$\mathbf{x}^c$ - the vector of continuous variables

N_c - number of continuous variables

$p(t, \mathbf{x})$ - the probability of features and survival times

$p(\mathbf{x}^c|j)$ - continuous feature probability

$\mathbf{v}_j$ - the center of the j-th Gaussian distribution of continuous explanatory variables

$\mathbf{F}_j$ - the covariance matrix of the j-th Gaussian distribution of continuous explanatory variables

$p(\mathbf{x}^d|j)$ - discrete feature probability

$A_{j,i}(\mathbf{x^c})$ - Gaussian membership function of i explanatory variables related to the jth cluster

σ - the variance of Gaussian function

N - the total population of the dataset

$p(j|t, \mathbf{x}_n^c, \mathbf{x}_n^d)$ - cluster membership value

AIC - Akaike information criterion

BIC - bayesian information criterion

K - total number of estimated model parameters

Chapter 4. ***Integrated survival analysis and frequent itemset-based association rule mining: a course failure-based prediction of student dropout***

$h_{CS}^c(t_k)$ - hazard function of the cth competing event based on cause specific hazard model

$CIF_{CS}^c(t_k)$ - cumulative incidence function of the cth competing event based on cause specific hazard model

ϕ_i - modified survival time based on subdistribution hazard model $h_{SD}^c(t_k)$ - hazard function of the cth competing event based on subdistribution hazard model

$CIF_{SD}^c(t_k)$ - cumulative incidence function of the cth competing event based on

subdistribution hazard model

N_f^* - the corrected number of cases according to subdistribution hazard model at time instance f

E - set of unique events

M - number of unique events

e_M - Mth unique item

N_k - number of cases in the k time instance

$\mathcal{A}_i(k)$ - set of events of the ith case in the k time instance

$a_{i,j}(k)$ - the occurrence of the jth event in the ith case at the kth time instant

$\mathcal{A}(k)$ - collection of cases in the k time instance

$X_p(k)$ - pth frequent itemset in the k time instance

$X(k)$ - collection of frequent itemset in the k time instance

$supp\big(X_p(k)\big)$ - support of frequent itemset $X_p(k)$

minsupp - minimum support hyperparameter of the algorithm

$X_{p*}(k)$ - set of triggering events of frequent itemset $X_p(k)$

$e_c(k)$ - set of consequent events of frequent itemset $X_p(k)$

L_{X_p} - the length of the association rule $X_p(k)$

$conf\big(X_{p*}(k) \Rightarrow e_c(k)\big)$ - confidence of rule $\big(X_{p*}(k) \Rightarrow e_c(k)\big)$: probability that $e_c(k)$ occurs given that $X_{p*}(k)$ occurs at the same time instance

$\mathcal{D}$ - segmented dataset

$\hat{\omega}_j$ - the semester in which a student is expected to complete the jth subject according to the sample curriculum

$\omega_{i,j}$ - the semester in which the first successful completion of the j subject is recorded for the ith student

Chapter 5. ***Integrated survival analysis and sequential pattern mining: a healthcare application***

E - set of unique events

M - number of unique events

e_M - Mth unique item

Z_i - explanatory variables of ith trace

z_{i,N_z} - N_zth explanatory variable of ith trace

ψ_i - sequence of ith trace

$a_{i,j}$ - jth event of the ith trace

$\tau_{i,j}$ - timestamp of jth event of the ith trace

Ψ - collection of sequences

N - total number of traces

α - a frequent sequence

α_l - lth event of frequent sequence α

supp(α) - support of sequence α

minsupp - minimum support hyperparameter of the algorithm

$\alpha^l \rightarrow \alpha^b$ - representation of frequent sequence α as antecedent part α^l and consequent part α^b

$\alpha^l \rightarrow \alpha_{l+1}$ - representation of frequent sequence α as antecedent part α^l and consequent event α_{l+1}. The superscript letter indicates a subsequence of α, which may contain multiple event transitions, while the subscript letters only denote one event

conf($\alpha^l \rightarrow \alpha^b$) - confidence of rule $\alpha^l \rightarrow \alpha^b$: probability that α^l continues with α^b

conf(α) - confidence of sequence α

$\alpha(\tau_i)$ - the ordered list of frequent event-timestamp duplet of the frequent sequence α in the ith trace

$(\alpha_b, \tau_{i,b})$ - the bth frequent event-timestamp duplet in the ith trace

$\Delta t_i(\alpha^l \rightarrow \alpha_{l+1})$ - duration from α_l to α_{l+1} in the case of the ith supporting trace.

supp($\alpha^l \rightarrow \alpha_{l+1}, t$) - time-dependent support function of rule $\alpha^l \rightarrow \alpha_{l+1}$

conf($\alpha^l \rightarrow \alpha_{l+1}, t$) - time dependent confidence function of rule $\alpha^l \rightarrow \alpha_{l+1}$

N_B - sample of bootstraps

M_B - number of bootstraps

$\hat{\alpha}$ - significance level

N_α - number of frequent sequences

$(N_{\alpha^l \rightarrow \alpha^b})$ - the quantity of discovered association rules

(L_τ) - the average length of traces

(L_α) - average length of frequent sequences

($L_{S,\alpha^l \to \alpha^b}$) average support of association rules

X - antecedent set of events of sequential rule $X \Rightarrow Y$

Y - consequent set of events of sequential rule $X \Rightarrow Y$

Bibliography

[1] Xian Liu. *Survival analysis: models and applications*. John Wiley & Sons, 2012.

[2] F. Emmert-Streib, and Matthias Dehmer. Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, 1(3):1013–1038, 2019.

[3] Amy Wachholtz and Gerardo Gonzalez. Co-morbid pain and opioid addiction: long term effect of opioid maintenance on acute pain. *Drug and alcohol dependence*, 145:143–149, 2014.

[4] Juan C Juajibioy et al. Study of university dropout reason based on survival model. *Open Journal of Statistics*, 6(5):908–916, 2016.

[5] Ahmed Ragab, Mohamed-Salah Ouali, Soumaya Yacout, and Hany Osman. Remaining useful life prediction using prognostic methodology based on logical analysis of data and Kaplan–Meier estimation. *Journal of Intelligent Manufacturing*, 27(5):943–958, 2016.

[6] Siddhartha Asthana, Pushpendra Singh, and Parul Gupta. Survival analysis: Objective assessment of wait time in hci. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 367–376, 2015.

[7] Philip Hougaard. *Multivariate Interval-Censored Survival Data: Parametric, Semi-parametric and Non-parametric Models*, pages 9–21. Springer International Publishing, Cham, 2014.

[8] Omid Alavi, Kasra Mohammadi, and Ali Mostafaeipour. Evaluating the suitability of wind speed probability distribution models: A case of study of east and southeast parts of iran. *Energy Conversion and Management*, 119:101–108, 2016.

[9] In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100, 2003.

[10] Samar Abd ElHafeez, Graziella D'Arrigo, Daniela Leonardis, Maria Fusaro, Giovanni Tripepi, and Stefanos Roumeliotis. Methods to analyze time-to-event data: the cox regression analysis. *Oxidative Medicine and Cellular Longevity*, 2021:1–6, 2021.

[11] Shankar Prinja, Nidhi Gupta, and Ramesh Verma. Censoring in clinical trials: review of survival analysis techniques. *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine*, 35(2):217, 2010.

[12] Arnoud J Templeton, Olga Ace, Eitan Amir, Francisco Vera-Badillo, Alberto Ocana, Gregory R Pond, and Ian F Tannock. Influence of censoring on conclusions of trials for women with metastatic breast cancer. *European journal of cancer*, 51(6):721–724, 2015.

[13] Sarah Lacny, Todd Wilson, Fiona Clement, Derek J Roberts, Peter Faris, William A Ghali, and Deborah A Marshall. Kaplan–meier survival analysis overestimates cumulative incidence of health-related events in competing risk settings: a meta-analysis. *Journal of clinical epidemiology*, 93:25–35, 2018.

[14] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9(1):381–386, 2020.

[15] Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding, and Chin-Teng Lin. A review of clustering techniques and developments. *Neurocomputing*, 267:664–681, 2017.

[16] Philippe Fournier-Viger, Wensheng Gan, Youxi Wu, Mourad Nouioua, Wei Song, Tin Truong, and Hai Duong. Pattern mining: Current challenges and opportunities. In *International Conference on Database Systems for Advanced Applications*, pages 34–49. Springer, 2022.

[17] Maximilian Pichler and Florian Hartig. Machine learning and deep learning—a review for ecologists. *Methods in Ecology and Evolution*, 14(4):994–1016, 2023.

[18] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021.

[19] Tamas Ruppert, Robert Csalodi, and Janos Abonyi. Estimation of machine setup and changeover times by survival analysis. *Computers & Industrial Engineering*, 153:107026, 2021.

[20] Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.

[21] Rui Xiao, Tarek Zayed, Mohamed A Meguid, and Laxmi Sushama. Improving failure modeling for gas transmission pipelines: A survival analysis and machine learning integrated approach. *Reliability Engineering & System Safety*, 241:109672, 2024.

[22] Borja del Pozo Cruz, Duncan E McGregor, Jesús del Pozo Cruz, Matthew P Buman, Javier Palarea-Albaladejo, Rosa M Alfonso-Rosa, and Sebastien FM Chastin. Integrating sleep, physical activity, and diet quality to estimate all-cause mortality risk: a combined compositional clustering and survival analysis of the national health and nutrition examination survey 2005–2006 cycle. *American Journal of Epidemiology*, 189(10):1057–1064, 2020.

[23] Xinliang Zhu, Jiawen Yao, and Junzhou Huang. Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 544–547. IEEE, 2016.

[24] Róbert Csalódi, Zsolt Bagyura, and János Abonyi. Mixture of survival analysis models-cluster-weighted weibull distributions. *IEEE Access*, 9:152288–152299, 2021.

[25] Róbert Csalódi, Zoltán Birkner, and János Abonyi. Learning interpretable mixture of weibull distributions—exploratory analysis of how economic development influences the incidence of covid-19 deaths. *Data*, 6(12):125, 2021.

[26] Róbert Csalódi and János Abonyi. Integrated survival analysis and frequent pattern mining for course failure-based prediction of student dropout. *Mathematics*, 9(5):463, 2021.

[27] Róbert Csalódi, Zsolt Bagyura, and János Abonyi. Time-dependent sequential association rule-based survival analysis: A healthcare application. *MethodsX*, page 102535, 2024.

[28] Róbert Csalódi, Zsolt Bagyura, Ágnes Vathy-Fogarassy, and János Abonyi. Time-dependent frequent sequence mining-based survival analysis. *Knowledge-Based Systems*, page 111885, 2024.

[29] Usha S Govindarajulu and Ralph B D'Agostino Sr. Review of current advances in survival analysis and frailty models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(6):e1504, 2020.

[30] Peter C Austin, Aurélien Latouche, and Jason P Fine. A review of the use of time-varying covariates in the fine-gray subdistribution hazard competing risk regression model. *Statistics in medicine*, 39(2):103–113, 2020.

[31] Shixin Huang, Xiaoyu Wan, Hang Qiu, Laquan Li, and Haiyan Yu. Constrained optimization for stratified treatment rules with multiple responses of survival data. *Information Sciences*, 596:343–361, 2022.

[32] Despina Koletsi and Nikolaos Pandis. Survival analysis, part 2: Kaplan-meier method and the log-rank test. *American journal of orthodontics and dentofacial orthopedics*, 152(4):569–571, 2017.

[33] Torben Martinussen. Causality and the cox regression model. *Annual Review of Statistics and Its Application*, 9:249–259, 2022.

[34] Enrico Colosimo, Fla´ vio Ferreira, Maristela Oliveira, and Cleide Sousa. Empirical comparisons between kaplan-meier and nelson-aalen survival function estimators. *Journal of Statistical Computation and Simulation*, 72(4):299–308, 2002.

[35] David G Kleinbaum and Mitchel Klein. *Survival analysis*. Springer, 2010.

[36] Terry M Therneau and Patricia M Grambsch. The cox model. In *Modeling survival data: extending the Cox model*, pages 39–77. Springer, 2000.

[37] Jose A Carta, Penelope Ramirez, and Sergio Velazquez. A review of wind speed probability distributions used in wind energy analysis: Case studies in the canary islands. *Renewable and sustainable energy reviews*, 13(5):933–955, 2009.

[38] Xiao-Hua Pan, Qi-Quan Xiong, and Zhi-Jun Wu. New method for obtaining the homogeneity index m of weibull distribution using peak and crack damage strains. *Int. J. Geomech*, 18(6):04018034, 2018.

[39] Achraf Bennis, Sandrine Mouysset, and Mathieu Serrurier. Estimation of conditional mixture weibull distribution with right censored data using neural network for time-to-event analysis. *Advances in Knowledge Discovery and Data Mining*, 12084:687, 2020.

[40] Katerina Langova et al. Survival analysis for clinical studies. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub*, 152(2):303–307, 2008.

[41] Emad E Elmahdy. Modelling reliability data with finite weibull or lognormal mixture distributions. *Appl Math Inform Sci*, 11(4):1081–1089, 2017.

[42] Yimin Chen, Jialing Huang, Xianying He, Yongxiang Gao, Gehendra Mahara, Zhuochen Lin, and Jinxin Zhang. A novel approach to determine two optimal cut-points of a continuous predictor with a u-shaped relationship

to hazard ratio in survival data: simulation and application. *BMC Medical Research Methodology*, 19(1):1–12, 2019.

[43] Patrick Royston, Douglas G Altman, and Willi Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25(1):127–141, 2006.

[44] Walter Bouwmeester, Nicolaas PA Zuithoff, Susan Mallett, Mirjam I Geerlings, Yvonne Vergouwe, Ewout W Steyerberg, Douglas G Altman, and Karel GM Moons. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med*, 9(5):e1001221, 2012.

[45] Sameer Sundrani and James Lu. Computing the hazard ratios associated with explanatory variables using machine learning models of survival data. *JCO Clinical Cancer Informatics*, 5:364–378, 2021.

[46] Ali Shariq Imran, Fisnik Dalipi, and Zenun Kastrati. Predicting student dropout in a MOOC: An evaluation of a deep neural network model. In *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence*, pages 190–195, 2019.

[47] T. A. Johansen, and Robert Babuska. Multiobjective identification of Takagi-Sugeno fuzzy models. *IEEE Transactions on Fuzzy Systems*, 11(6):847–860, 2003.

[48] Paidamoyo Chapfuwa, Chunyuan Li, Nikhil Mehta, Lawrence Carin, and Ricardo Henao. Survival cluster analysis. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 60–68, 2020.

[49] Janos Abonyi, Robert Babuska, and Ferenc Szeifert. Modified Gath-Geva fuzzy clustering for identification of takagi-sugeno fuzzy models. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 32(5):612–621, 2002.

[50] Mehdi Azizmohammad Looha, Elaheh Zarean, Fatemeh Masaebi, Mohamad Amin Pourhoseingholi, and Mohamad Reza Zali. Assessment of prognostic factors in long-term survival of male and female patients with colorectal cancer using non-mixture cure model based on the weibull distribution. *Surgical Oncology*, 38:101562, 2021.

[51] Horst Rinne. *The Weibull distribution: a handbook*. CRC Press, 2008.

[52] Nima Sammaknejad, Yujia Zhao, and Biao Huang. A review of the expectation maximization algorithm in data-driven process identification. *Journal of process control*, 73:123–136, 2019.

[53] Tomohiro Takagi and Michio Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE transactions on systems, man, and cybernetics*, (1):116–132, 1985.

[54] Tomaž Bučar, Marko Nagode, and Matija Fajdiga. Reliability approximation using finite Weibull mixture distributions. *Reliability Engineering & System Safety*, 84(3):241–251, 2004.

[55] K Das. A comparative study of exponential distribution vs weibull distribution in machine reliability analysis in a cms design. *Computers & Industrial Engineering*, 54(1):12–33, 2008.

[56] Gauss M Cordeiro, Edwin MM Ortega, and Artur J Lemonte. The exponential–weibull lifetime distribution. *Journal of Statistical Computation and simulation*, 84(12):2592–2606, 2014.

[57] Di Zhou, Xiao Zhuang, and Hongfu Zuo. A novel three-parameter weibull distribution parameter estimation using chaos simulated annealing particle swarm optimization in civil aircraft risk assessment. *Arabian Journal for Science and Engineering*, 46(9):8311–8328, 2021.

[58] CD Lai, DNP Murthy, and M Xie. Weibull distributions. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(3):282–287, 2011.

[59] Bernd Schoner and Neil Gershenfeld. Cluster-weighted modeling: Probabilistic time series prediction, characterization, and synthesis. In *Nonlinear Dynamics and Statistics*, pages 365–385. Springer, 2001.

[60] Ibrahim A Hameed. Using gaussian membership functions for improving the reliability and robustness of students' evaluation systems. *Expert systems with Applications*, 38(6):7135–7142, 2011.

[61] G. J. McLachlan, and Suren Rathnayake. On the number of components in a Gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(5):341–355, 2014.

[62] Huan Wan, Hui Wang, Bryan Scotney, and Jun Liu. A novel Gaussian mixture model for classification. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3298–3303. IEEE, 2019.

[63] E. Patel and Dharmender Singh Kushwaha. Clustering cloud workloads: K-means vs Gaussian mixture model. *Procedia Computer Science*, 171:158–167, 2020.

[64] Scott I Vrieze. Model selection and psychological theory: a discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). *Psychological Methods*, 17(2):228–243, 2012.

[65] Melania Pintilie. *Competing risks: a practical perspective*, volume 58. John Wiley & Sons, 2006.

[66] Kristen A Severson, Peter M Attia, Norman Jin, Nicholas Perkins, Benben Jiang, Zi Yang, Michael H Chen, Muratahan Aykol, Patrick K Herring, Dimitrios Fraggedakis, et al. Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*, 4(5):383–391, 2019.

[67] D. F. Andrews, and Agnes M Herzberg. *Data: a collection of problems from many fields for the student and research worker*. Springer Science & Business Media, 2012.

[68] Yue Wu, Yicong Zhou, George Saveriades, Sos Agaian, Joseph P Noonan, and Premkumar Natarajan. Local shannon entropy measure with statistical tests for image randomness. *Information Sciences*, 222:323–342, 2013.

[69] John hopkins university, coronavirus resource center: Cases and mortality by country. https://coronavirus.jhu.edu/data/mortality. [Online; accessed 27-Sept-2021].

[70] World data bank. https://data.worldbank.org/. [Online; accessed 27-Sept-2021].

[71] Most obese countries 2021. https://worldpopulationreview.com/country-rankings/most-obese-countries. [Online; accessed 11-Oct-2021].

[72] Roengrudee Patanavanich and Stanton A Glantz. Smoking is associated with covid-19 progression: a meta-analysis. *Nicotine and Tobacco Research*, 22(9):1653–1656, 2020.

[73] Slobodan Peric and Thomas M Stulnig. Diabetes and covid-19. *Wiener Klinische Wochenschrift*, 132(13):356–361, 2020.

[74] Nicola L Bulled and Richard Sosis. Examining the relationship between life expectancy, reproduction, and educational attainment. *Human Nature*, 21(3):269–289, 2010.

[75] Lena K Makaroun, Rachel L Bachrach, and Ann-Marie Rosland. Elder abuse in the time of covid-19—increased risks for older adults and their caregivers. *The American Journal of Geriatric Psychiatry*, 28(8):876, 2020.

[76] NM Katsoulakos, L-MN Misthos, Ilias G Doulos, and VS Kotsios. Environment and development. In *Environment and Development*, pages 499–569. Elsevier, 2016.

[77] Garyfallos Konstantinoudis, Tullia Padellini, James Bennett, Bethan Davies, Majid Ezzati, and Marta Blangiardo. Long-term exposure to air-pollution and covid-19 mortality in england: a hierarchical spatial analysis. *Environment international*, 146:106316, 2021.

[78] Arunava Bhadra, Arindam Mukherjee, and Kabita Sarkar. Impact of population density on covid-19 infected and mortality rate in india. *Modeling Earth Systems and Environment*, 7(1):623–629, 2021.

[79] Preeti Malik, Urvish Patel, Karan Patel, Mehwish Martin, Chail Shah, Deep Mehta, Faizan Ahmad Malik, and Ashish Sharma. Obesity a predictor of outcomes of covid-19 hospitalized patients—a systematic review and meta-analysis. *Journal of medical virology*, 93(2):1188–1193, 2021.

[80] Daniela Calina, Thomas Hartung, Ileana Mardare, Mihaela Mitroi, Konstantinos Poulas, Aristidis Tsatsakis, Ion Rogoveanu, and Anca Oana Docea. Covid-19 pandemic and alcohol consumption: Impacts and interconnections. *Toxicology reports*, 8:529–535, 2021.

[81] Bernhard Haller, Georg Schmidt, and Kurt Ulm. Applying competing risks regression models: an overview. *Lifetime data analysis*, 19:33–58, 2013.

[82] Inmaculada Aban. Time to event analysis in the presence of competing risks. *Journal of Nuclear Cardiology*, 22:466–467, 2015.

[83] Jan Beyersmann and Martin Schumacher. Time-dependent covariates in the proportional subdistribution hazards model for competing risks. *Biostatistics*, 9(4):765–776, 2008.

[84] Peter C Austin and Jason P Fine. Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. *Statistics in medicine*, 36(8):1203–1209, 2017.

[85] Ronald B Geskus. Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring. *Biometrics*, 67(1):39–49, 2011.

[86] Paul C Lambert. The estimation and modeling of cause-specific cumulative incidence functions using time-dependent weights. *The Stata Journal*, 17(1):181–207, 2017.

[87] Takeshi Emura, Jia-Han Shih, Il Do Ha, and Ralf A Wilke. Comparison of the marginal hazard model and the sub-distribution hazard model for competing risks under an assumed copula. *Statistical methods in medical research*, 29(8):2307–2327, 2020.

[88] Thomas A Gerds, Thomas H Scheike, and Per K Andersen. Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in medicine*, 31(29):3921–3930, 2012.

[89] Jason P Fine and Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446):496–509, 1999.

[90] Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

[91] Cristiano Antonelli, Francesco Crespi, and Giuseppe Scellato. Internal and external factors in innovation persistence. *Economics of Innovation and New Technology*, 22(3):256–280, 2013.

[92] Rinku Sutradhar and Peter C Austin. Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios. *Annals of epidemiology*, 28(1):54–57, 2018.

[93] Bryan Lau, Stephen R Cole, and Stephen J Gange. Competing risk regression models for epidemiologic data. *American journal of epidemiology*, 170(2):244–256, 2009.

[94] Moritz Berger, Matthias Schmid, Thomas Welchowski, Steffen Schmitz-Valckenberg, and Jan Beyersmann. Subdistribution hazard models for competing risks in discrete time. *Biostatistics*, 21(3):449–466, 2020.

[95] Mukesh Kumar, AJ Singh, and Disha Handa. Literature survey on educational dropout prediction. *International Journal of Education and Management Engineering*, 7(2):8, 2017.

[96] BOBBY K Simon and ANJANA P Nair. Association rule mining to identify the student dropout in moocs. *Int. Res. J. Eng. Technol.(IRJET)*, 6(01), 2019.

[97] Lovenoor Aulck, Dev Nambi, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. Mining university registrar records to predict first-year undergraduate attrition. *International Educational Data Mining Society*, 2019.

[98] Paul T Von Hippel and Alvaro Hofflinger. The data revolution comes to higher education: Identifying students at risk of dropout in chile. *Journal of Higher Education Policy and Management*, pages 1–22, 2020.

[99] D. Kim, and Seoyong Kim. Sustainable education: analyzing the determinants of university student dropout by nonlinear panel data models. *Sustainability*, 10(4):954, 2018.

[100] Marcell Nagy and Roland Molontay. Predicting dropout in higher education based on secondary school performance. In *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*, pages 000389–000394. IEEE, 2018.

[101] Li Zhang and Huzefa Rangwala. Early identification of at-risk students using iterative logistic regression. In *International Conference on Artificial Intelligence in Education*, pages 613–626. Springer, 2018.

[102] Alana Platt, Onochie Fan-Osuala, and Nicolas Herfel. Understanding and predicting student retention and attrition in it undergraduates. In *Proceedings of the 2019 on Computers and People Research Conference*, pages 135–138, 2019.

[103] Aishwarya Suresh, HS Sushma Rao, and Vinayak Hegde. Academic dashboard—prediction of institutional student dropout numbers using a naïve bayesian algorithm. In *Computing and Network Sustainability*, pages 73–82. Springer, 2017.

[104] Francesca Del Bonifro, Maurizio Gabbrielli, Giuseppe Lisanti, and Stefano Pio Zingaro. Student dropout prediction. In *International Conference on Artificial Intelligence in Education*, pages 129–140. Springer, 2020.

[105] Lijia Chen, Pingping Chen, and Zhijian Lin. Artificial intelligence in education: a review. *Ieee Access*, 8:75264–75278, 2020.

[106] Hui Luan, Peter Geczy, Hollis Lai, Janice Gobert, Stephen JH Yang, Hiroaki Ogata, Jacky Baltes, Rodrigo Guerra, Ping Li, and Chin-Chung Tsai. Challenges and future directions of big data and artificial intelligence in education. *Frontiers in psychology*, 11, 2020.

[107] Neema Mduma, Khamisi Kalegele, and Dina Machuve. A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal*, pages 1–10, 2019.

[108] Oscar Espinoza, Luis Eduardo González, Dante Castillo, and Noel McGinn. Classification of dropouts to improve student re-engagement: The case of chilean secondary opportunity centers. *Urban Education*, 58(9):2177–2208, 2023.

[109] Yujing Chen, Aditya Johri, and Huzefa Rangwala. Running out of stem: a comparative study across stem majors of college students at-risk of dropping out early. In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 270–279, 2018.

[110] Niki Gitinabard, Farzaneh Khoshnevisan, Collin F Lynch, and Elle Yuan Wang. Your actions or your associates? predicting certification and dropout in moocs with behavioral and social features. *arXiv preprint arXiv:1809.00052*, 2018.

[111] Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn Penstein Rosé. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop*, volume 11, page 14, 2013.

[112] Ashish Dutt, Maizatul Akmar Ismail, and Tutut Herawan. A systematic review on educational data mining. *IEEE Access*, 5:15991–16005, 2017.

[113] Vincenzo Coviello and May Boggess. Cumulative incidence estimation in the presence of competing risks. *The Stata Journal*, 4(2):103–112, 2004.

[114] José María Luna, Philippe Fournier-Viger, and Sebastián Ventura. Frequent itemset mining: A 25 years review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1329, 2019.

[115] Tung Kieu, Bay Vo, Tuong Le, Zhi-Hong Deng, and Bac Le. Mining top-k co-occurrence items with sequential pattern. *Expert Systems with Applications*, 85:123–133, 2017.

[116] Philippe Fournier-Viger, Antonio Gomariz, Ted Gueniche, Azadeh Soltani, Cheng-Wei Wu, and Vincent S Tseng. Spmf: a java open-source pattern mining library. *The Journal of Machine Learning Research*, 15(1):3389–3393, 2014.

[117] Standards and guidelines for quality assurance in the european higher education area. https://www.enqa.eu/esg-standards-and-guidelines-for-quality-assurance-in-the-european-higher-education-area/, 2015. [Online; accessed 15-Februar-2021].

[118] Alican Dogan and Derya Birant. Machine learning and data mining in manufacturing. *Expert Systems with Applications*, 166:114060, 2021.

[119] Sunil Kumar and Krishna Kumar Mohbey. A review on big data based parallel and distributed approaches of pattern mining. *Journal of King Saud University-Computer and Information Sciences*, 34(5):1639–1662, 2022.

[120] Philippe Fournier-Viger, Antonio Gomariz, Manuel Campos, and Rincy Thomas. Fast vertical mining of sequential patterns using co-occurrence information. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 40–52. Springer, 2014.

[121] Yan Li, Shuai Zhang, Lei Guo, Jing Liu, Youxi Wu, and Xindong Wu. Netnmsp: Nonoverlapping maximal sequential pattern mining. *Applied Intelligence*, pages 1–24, 2022.

[122] Ling Wang, Jianyao Meng, Peipei Xu, and Kaixiang Peng. Mining temporal association rules with frequent itemsets tree. *Applied Soft Computing*, 62:817–829, 2018.

[123] Philippe Fournier-Viger, Ted Gueniche, and Vincent S Tseng. Using partially-ordered sequential rules to generate more accurate sequence prediction. In *Advanced Data Mining and Applications: 8th International Conference, ADMA 2012, Nanjing, China, December 15-18, 2012. Proceedings 8*, pages 431–442. Springer, 2012.

[124] Thabet Slimani and Amor Lazzez. Efficient analysis of pattern and association rule mining approaches. *arXiv preprint arXiv:1402.2892*, 2014.

[125] David Gotz, Fei Wang, and Adam Perer. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of biomedical informatics*, 48:148–159, 2014.

[126] Mohammed J Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine learning*, 42(1):31–60, 2001.

[127] Heikki Mannila, Hannu Toivonen, and A Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data mining and knowledge discovery*, 1:259–289, 1997.

[128] Seigo Matsuno, Yasuo Uchida, Tsutomu Ito, and Takao Ito. Lifespan of information service firms in japan: a survival analysis. *International Journal of Information Systems and Project Management*, 6(1):61–70, 2018.

[129] Richard A Armstrong. When to use the bonferroni correction. *Ophthalmic and Physiological Optics*, 34(5):502–508, 2014.

[130] Esmaeel Dodangeh, Bahram Choubin, Ahmad Najafi Eigdir, Narjes Nabipour, Mehdi Panahi, Shahaboddin Shamshirband, and Amir Mosavi. Integrated machine learning methods with resampling algorithms for flood susceptibility prediction. *Science of the Total Environment*, 705:135983, 2020.

[131] Marie-Therese Puth, Markus Neuhäuser, and Graeme D Ruxton. On the variety of methods for calculating confidence intervals by bootstrapping. *Journal of Animal Ecology*, 84(4):892–897, 2015.

[132] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage Uday Kiran, Yun Sing Koh, and Rincy Thomas. A survey of sequential pattern mining. *Data Science and Pattern Recognition*, 1(1):54–77, 2017.

[133] Alpa Reshamwala and Neha Mishra. Analysis of sequential pattern mining algorithms. *Int J Sci Eng Res*, 5(2):1034–8, 2014.

[134] Kaustubh Beedkar, Rainer Gemulla, and Wim Martens. A unified framework for frequent sequence mining with subsequence constraints. *ACM Transactions on Database Systems (TODS)*, 44(3):1–42, 2019.

[135] Alberto Oliveira, Ricardo Freitas, Alípio Jorge, Vítor Amorim, Nuno Moniz, Ana CR Paiva, and Paulo J Azevedo. Sequence mining for automatic generation of software tests from gui event traces. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 516–523. Springer, 2020.

[136] Meiling Zhu, Chen Liu, and Yanbo Han. Approach to discovering companion patterns based on traffic data stream. *IET Intelligent Transport Systems*, 12(10):1351–1359, 2018.

[137] Avinash Kadimisetty, C Oswald, and B Sivaselvan. Frequent pattern mining approach to image compression. In *2016 22nd Annual International Conference on Advanced Computing and Communication (ADCOM)*, pages 27–32. IEEE, 2016.

[138] Robert Moskovitch. Multivariate temporal data analysis-a review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(1):e1430, 2022.

[139] Yu Hirate and Hayato Yamana. Sequential pattern mining with time intervals. In *Advances in Knowledge Discovery and Data Mining: 10th Pacific-Asia Conference, PAKDD 2006, Singapore, April 9-12, 2006. Proceedings 10*, pages 775–779. Springer, 2006.

[140] Yu Hirate and Hayato Yamana. Generalized sequential pattern mining with item intervals. *J. Comput.*, 1(3):51–60, 2006.

[141] Joanna Olbryś and Krzysztof Ostrowski. An entropy-based approach to measurement of stock market depth. *Entropy*, 23(5):568, 2021.

[142] Rui Xia, Jie Jiang, and Huihui He. Distantly supervised lifelong learning for large-scale social media sentiment analysis. *IEEE Transactions on Affective Computing*, 8(4):480–491, 2017.

[143] Mohammed Saqr and Sonsoles López-Pernas. The longitudinal trajectories of online engagement over a full program. *Computers & Education*, 175:104325, 2021.

[144] Ishleen Kaur, MN Doja, and Tanvir Ahmad. Time-range based sequential mining for survival prediction in prostate cancer. *Journal of Biomedical Informatics*, 110:103550, 2020.

[145] Girish Keshav Palshikar, Sachin Pawar, and Nitin Ramrakhiyani. Role models: Mining role transitions data in it project management. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 508–517. IEEE, 2016.

[146] Viv Bewick, Liz Cheek, and Jonathan Ball. Statistics review 12: survival analysis. *Critical care*, 8:1–6, 2004.

[147] David Oakes. Biometrika centenary: survival analysis. *Biometrika*, 88(1):99–142, 2001.

[148] Anthony Joe Turkson, Francis Ayiah-Mensah, and Vivian Nimoh. Handling censoring and censored data in survival analysis: a standalone systematic literature review. *International journal of mathematics and mathematical sciences*, 2021:1–16, 2021.

[149] Philippe Fournier-Viger, Roger Nkambou, and Vincent Shin-Mu Tseng. Rulegrowth: mining sequential rules common to several sequences by pattern-growth. In *Proceedings of the 2011 ACM symposium on applied computing*, pages 956–961, 2011.

[150] Huy Quan Vu, Gang Li, Rob Law, and Yanchun Zhang. Travel diaries analysis by sequential rule mining. *Journal of travel research*, 57(3):399–413, 2018.

[151] Bernard MY Cheung and Chao Li. Diabetes and hypertension: is there a common metabolic pathway? *Current atherosclerosis reports*, 14:160–166, 2012.

[152] János Nemcsik, Norbert Habony, György Ábrahám, Csaba Farsang, Attila Simon, Béla Benczúr, Dénes Páll, and Zoltán Járai. First results of the new hungarian hypertension registry: Number of daily measurements and proportion of patients in different blood pressure categories. *Journal of Hypertension*, 39:e398–e399, 2021.

[153] Biju Paul, Neeta C Wilfred, Richard Woodman, and Carmine DePasquale. Prevalence and correlates of anaemia in essential hypertension. *Clinical and experimental Pharmacology and Physiology*, 35(12):1461–1464, 2008.

[154] Marty S Player and Lars E Peterson. Anxiety disorders, hypertension, and cardiovascular risk: a review. *The International Journal of Psychiatry in Medicine*, 41(4):365–377, 2011.

[155] Seyedeh M Zekavat, Michael Honigberg, James P Pirruccello, Puja Kohli, Elizabeth W Karlson, Christopher Newton-Cheh, Hongyu Zhao, and Pradeep Natarajan. Elevated blood pressure increases pneumonia risk: epidemiological association and mendelian randomization in the uk biobank. *Med*, 2(2):137–148, 2021.

[156] János Sándor, Ildikó Tokaji, Nouh Harsha, Magor Papp, Róza Ádány, and Árpád Czifra. Organised and opportunistic prevention in primary health care: estimation of missed opportunities by population based health interview surveys in hungary. *BMC Family Practice*, 21:1–12, 2020.

[157] Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. Mining electronic health records (ehrs) a survey. *ACM Computing Surveys (CSUR)*, 50(6):1–40, 2018.