Decision Analytics Journal 11 (2024) 100452

Contents lists available at ScienceDirect

Decision Analytics Journal

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

A hybrid machine learning with process analytics for predicting customer experience in online insurance services industry

Fatemeh Akhavan, Erfan Hassannayebi *

Department of Industrial Engineering, Sharif University of Technology, Tehran, Iran

ARTICLE INFO

Keywords:
Process intelligence
Business process mining
Predictive analytics
Customer experience
Online insurance

ABSTRACT

It is essential to innovate and improve service levels by predicting possible process outcomes due to the growth of online service providers. This study estimates the customer satisfaction level based on customer experience analysis. In doing so, we answer recent calls for research about a more thorough exploration of customer behavior using predictive process monitoring techniques. In particular, a hybrid framework of supervised/unsupervised machine learning methods is proposed to predict the outcomes of customers' experiences while dealing with the problem of high intra-class variance. This problem occurs due to the large dispersion of traces identified in the customer journeys. In this regard, customer journeys are first matched with the event log format aiming to implement a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering technique based on the similarity between the customer journeys. After summarizing the journeys by removing low-value activities, the multi-class decision tree classification method is applied, and the level of customer satisfaction is predicted. Due to the imbalanced nature of the data, the oversampling for imbalanced classification is applied to achieve good results in accuracy indicators such as recall, precision, and F1-score. Finally, the proposed approach has been evaluated on a real-life event log, BPI Challenge 2016, to investigate unsatisfied customers. The results of the machine learning models on the test data show a high degree of accuracy in predicting customer dissatisfaction.

1. Introduction

Process intelligence is a powerful technique in the sub-disciplines of both Artificial intelligence (AI) and business process management (BPM) [1]. It allows gaining insights into business process performance and behavior using historical records from event logs. Process mining is an essential category of process intelligence that targets existing processes in an organization and provides solutions to subject matter experts for modeling, documenting, and collaborating to re-engineer an organization's operational processes [2,3].

In customer services, process intelligence can be used to monitor, analyze, and optimize all customer interactions including phone calls, email, chatbots, and social media. Process intelligence can provide insights into customer behavior, preferences, and pain points. The benefits of using process intelligence in customer services include the ability to improve customer satisfaction, optimize business processes, better allocate resources, and reduce churn. In this context, customer journey mapping is a traditional technique used to understand the end-to-end customer experience and identify any pain points or areas for improvement. On the other hand, customer journey analytics refers to the science of analyzing customer behavior to measure its impact on business outcomes based on real data. The entire sequence of activities that customers go through when interacting with a business is known as the customer journey [4].

Nowadays, many fields and types of analysis apply the customer journeys. One of the applications of customer journey analytics is the extraction of journey maps using data mining approaches [5]. With the increase of online service providers, customer behavior has become more complex and thus the extracted journey map will be incomprehensible. To solve this problem, the issue of abstracting customer journey maps has been considered one of the main aspects of customer journey analysis [6]. On the other hand, recently, to improve the key performance indicators (KPIs) related to businesses, recommender systems have been widely used. Personalized recommendations, which can be related to various decision-making processes [7,8], such as what things to buy, are provided by recommender systems [9], which use customer behavior as input. In this regard, some of the recent articles have focused on improving recommender systems using customer journey analysis methods [4,10,11]. Marketing professionals can better understand what customers want and how to engage with them by mapping the consumer journey [12]. Also, due to the attraction of customers from different communication channels, customer journey analysis methods are used to improve the attraction and effectiveness of advertising strategies [13].

* Corresponding author.
E-mail addresses: FatemehAkhavan1375@gmail.com (F. Akhavan), hassannayebi@sharif.edu (E. Hassannayebi).

https://doi.org/10.1016/j.dajour.2024.100452
Received 8 June 2023; Received in revised form 24 March 2024; Accepted 25 March 2024
Available online 27 March 2024

Customer satisfaction analysis plays a pivotal role in enhancing the quality of services provided by industries. In today's digital age, where billions of individuals are using online platforms, each piece of information carries emotions and sentiments. Whether it is a positive or negative experience during customer journeys. In this situation, companies face fierce competition and must continually strive to attract and retain customers. To achieve this, understanding customer needs and levels of satisfaction during their journeys is paramount. Traditional methods of collecting feedback, such as paper forms or online surveys, have limitations, including low response rates and potential biases. In contrast, customer journeys do not have these limitations [14].

The insurance service industry has always been one of the most important B2C (business-to-consumer) industries in the world. Businesses in this field are in constant and continuous communication with customers. The complexity of customer behavior has made the way to interact with them in this industry become a debatable issue. Businesses are trying to create a competitive advantage so that they can create a good experience for their customers. In this regard, analyzing customers and providing predictions of their behavior has become one of the favorite topics [15].

The insurance industry faces various challenges that can be addressed through innovative methods. Here is a list of issues in the insurance sector along with suggested methods based on the provided research [16]:

- Fraud Detection: Detecting fraudulent claims is a significant challenge for insurance companies.
- Claim Analysis: Distinguishing between genuine and fraudulent claims is crucial for efficient operations.
- Data Management: Handling and analyzing vast amounts of data efficiently is essential for decision-making.
- Customer Profiling: Understanding client behavior and preferences is key to offering tailored insurance solutions.
- Data Utilization: Leveraging the full potential of available data for improved decision-making and operational efficiency.

Suggested Methods:

- Machine Learning Algorithms: Utilize ML algorithms for fraud detection, and claim analysis to enhance accuracy and efficiency.
- Data Mining Techniques: Employ data mining for fraud detection, and customer profiling to extract valuable insights.
- Exploratory Data Analysis (EDA): Conduct EDA to identify meaningful factors for claim filing and acceptance, aiding in decision-making.
- Feature Selection: Implement feature selection techniques to reduce data dimensionality and improve analysis results.
- Predictive Analytics: Use predictive analytics for claims processing, underwriting analysis, and customer behavior insights.
- Enhanced Data Utilization: Increase data utilization through ML to automate processes, reduce claim handling costs, and improve customer satisfaction.

By addressing these issues with the suggested methods, insurance companies can enhance their operations, improve decision-making, and better serve their clients in a rapidly evolving industry.

Process mining can be a powerful approach for exploring customer journeys, as it allows for the visualization and analysis of the actual customer behavior and interactions with the system. The value-creating applications of process mining for organizations are process visualization, identifying and monitoring key performance indicators of processes, and making decisions and corrective actions based on them. Value means efficiency and improvement of current processes, financial and non-financial benefits such as customer satisfaction [17]. Here are some ways that process mining can play a role in exploring customer journeys:

Visualizing the customer journey: Process mining can provide a visual representation of the customer journey, highlighting the different touchpoints and how customers navigate through the system. This can help to identify bottlenecks and areas for improvement.

Identifying customer behavior patterns: With process mining, it is possible to analyze the behavior of individual customers or groups of customers, such as identifying which pages they visit most often or how they interact with the website. This can help to understand the customer journey more deeply and identify opportunities for improvement.

Measuring the efficiency of the customer journey: Process mining can help quantify the efficiency of a given customer journey, such as the time it takes for customers to complete certain tasks or the number of steps they must go through to complete a transaction. This can help to streamline the customer journey and improve its overall effectiveness.

Analyzing the impact of changes: By analyzing the customer journey before and after changes are made, process mining can help to determine the effectiveness of those changes and identify further areas for improvement.

Besides the above-mentioned use cases of customer journey analytics, novel applications of process mining techniques are helping decision-makers to go beyond the traditional process discovery and involving the machine learning algorithm to enhance the user experience further. For example, customer journey prediction is the process of analyzing customer behavior and data to predict the potential path that a customer is likely to take in their interaction with a brand or company. This is especially relevant in the context of e-commerce and online businesses such as insurance services, where customers interact with companies through multiple channels such as social media, email, and websites. In this context, predictive modeling refers to building predictive models based on data analysis to identify potential customer journeys and make predictions about future behavior. This paper, therefore, focuses on the following research questions:

– **RQ1:** According to previous studies, how can supervised and unsupervised machine learning algorithms be used to analyze customer journeys and their patterns to predict the behavior of website users?

– **RQ2:** According to the history of previous users' behavior, how process mining and predictive process monitoring algorithms can be used to predict whether or not users will submit a complaint?

It is crucial to highlight the challenges that necessitated the innovative solutions presented. The challenges faced in predictive business process monitoring, particularly in the realm of customer journey analysis, are multifaceted. One significant challenge addressed by this study is the high intra-class variance, stemming from the substantial dispersion of traces identified in customer journeys. This variance poses a fundamental obstacle to accurate outcome prediction and customer satisfaction assessment. Additionally, the complexity of customer behavior in online service environments, characterized by diverse communication channels and intricate interaction patterns, presents a challenge in extracting meaningful insights from customer journey data.

Moreover, the imbalanced nature of the data, where certain outcomes are underrepresented, further complicates the predictive modeling process. These challenges underscore the necessity for a comprehensive and innovative approach that combines supervised and unsupervised machine learning methods to overcome the limitations of traditional predictive models and enhance the accuracy and effectiveness of outcome predictions in customer experience analysis. By addressing these challenges head-on, the research contributions outlined in the paper aim to bridge critical gaps in existing methodologies and offer a proactive and comprehensive framework for improving customer satisfaction analysis and predictive process monitoring in the online service industry.

In summary, the research contributions are fourfold: first, a hybrid machine learning method extended existing studies by proposing a novel predictive model to fill the existing research gaps and overcome the problem of high intra-class variance. This framework considers

the similarity between journeys and simultaneously employs decision tree classification as supervised and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering as unsupervised learning methods. Second, the analysis proposes an essential contribution to moving from reactive actions to proactive behavior. By predicting the outcome of a running case and identifying journeys that cause the failure, organizations can plan for corrective actions and improve the rate of succession. Third, this study extends the scope of the traditional outcome prediction with binary labels towards improving the classification method by considering a range of outcomes, e.g., 0–5 be the range of customer satisfaction level. Thus, a multi-class decision tree as a classification method is employed, and customer satisfaction level is predicted. Fourth, combining statistical feature selection methods reduces the created clusters' dimensions and increases the prediction accuracy. The research contributions, both in terms of application and modeling framework, have been refactored to directly connect to the potential reader, the distinguishing aspects and distinctiveness of the proposed process mining method.

The remainder of this paper is organized as follows: Section 3 discusses the literature review on process mining and predictive process monitoring, positioning the study within the scope of the application of predictive process monitoring methods in predicting the measures of processes. Some important concepts used in the definition of the problem are described in Section 2. The problem is described in Section 4 and the solution method is presented in Section 5, followed by a summary of the outcomes and research findings. The case study will be presented in Section 6. At last, Section 7 draws findings and directions for future work.

2. Literature review

This section provides background on process mining and predictive process monitoring. Research on predictive process monitoring and machine learning methods has recently acknowledged a noteworthy extent of consideration by process analysts and business owners. According to a systematic review conducted in 2020 [18], the trend of publishing articles on predictive process monitoring from the first decade of the 21st century has grown. In the continuation of this section, previous research on process mining and predictive process monitoring are reviewed.

The concept of predictive process monitoring was introduced in 2011 by van der Aalst et al. [19]. They constructed an annotated transition system (ATS) and used it for prediction purposes, focused on completion time prediction, and implemented the proposed approach on the event log of a municipality in the Netherlands. Studies on the use of predictive process monitoring are categorized by the type of prediction task. Continuing, we will review the literature of relevant studies based on these categories.

2.1. Next-event prediction models

Nowadays, predicting the next step of a customer journey by increasing the understanding of customer behavior is one of the factors influencing the success of organizations [11]. Organizations can use this information to develop a recommender system and improve customer experience. Researchers looked into the use of neural networks in business process monitoring due to the popularity of Deep Learning. Most state-of-the-art next activity/event(s) prediction approaches use Long Short Term Memory (LSTM) cells at the prediction phase. Deep learning methods often use a one-hot encoding feature vector as input. Pasquadibisceglie et al. [20] proposed a novel predicting approach based on one-hot encoding and LSTM as a learning algorithm. Jalayer et al. [21] similarly predicted the next activity, with the difference that by using the SoftMax function, calculated the importance of the activities and selected the most important features to enter the learning algorithm. One of the challenges of predicting the next activity/event is recurring prediction. That means the following steps can be predicted sequentially by using one predicted step. In this regard, Pauwels et al. [22] introduce a basic neural network method that can incrementally predict the next activity of sequences.

2.2. Outcome-oriented predictive process monitoring

The outcome-oriented prediction tasks such as predicting the failure of a process [23], purchase or not, etc., can give an insight into the business and help the business owners in the decision-making process. In recent years, different approaches have been proposed to predict the outcome of a business process. Some studies, i.e. Kim et al. [24], Elkhawaga et al. [25], Gusmao et al. [26], Lee et al. [27], have focused on selecting appropriate features to predict the outcome with higher accuracy. In this regard, a resource-aware feature selection method that identifies features related to resources was presented by Kim et al. [24]. Also, Elkhawaga et al. [25] extracted the appropriate features by assigning importance to the activities with the implementation of Shapley Additive Explanations (SHAP) and Partial Dependence Plot (PDP) methods. Gusmao et al. [26] addressed the application of this method and they analyzed the customer journeys by detecting fraud in the energy industry. They used Pareto analysis in the feature selection phase. Lee et al. [27] used Repeated Incremental Pruning to Produce Error Reduction (RIPPER) in the training phase and they could improve prediction performance. Francescomarino et al. [28] used the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Model-Based Clustering (MBASED) after using frequency-based encoding to predict the outcome of the tumor detection process. Since data privacy against potential attacks is important and vital in the healthcare industry, Rafiei et al. [29] also investigated the challenges of protecting patients' privacy protection techniques, focusing on group-based techniques in this industry. Francescomarino et al. [30] also employed the Canopy algorithm as an incremental-based clustering method after using index-based encoding. They implemented Hoeffding Tree (HT) and Adaptive Hoeffding Tree (AT) techniques to improve the prediction results. Moreover, Wang et al. [31] proposed LSTM-based approaches to overcome the time challenge in training an incremental model. Pasquadibisceglie et al. [32] focused on the encoding phase of a prediction process and trained a Convolutional Neural Network (CNN) model to predict the outcome of a running case. In the following, Weinzierl et al. [33] also identified temporary solutions (workarounds) according to the registered incidents. In this regard, Boolean encoding and CNN classification methods were used.

2.3. Remaining and completion time prediction

One of the indicators for measuring the quality of service providers such as banks, information technology companies, etc. is the waiting time in processes such as loan application and incident and problem troubleshooting [34]. So, in addition to outcome-oriented approaches, time-related prediction tasks have many applications in identifying deviations, such as preventing breaches of service level agreements which can increase customer satisfaction, identifying delays and the bottlenecks that cause delays, etc. The most famous and widely used methods of predicting process execution time are summarized in a systematic review published by Márquez-Chamorro et al. [12]. These methods fall into two groups: statistical techniques and machine learning methods. Polato et al. [35] used regression models to predict the remaining time of a running case by capturing the stability of the primary process. On the other hand, Gunnarsson et al. [36] implemented LSTM as a machine learning method for predicting the completion time of a luggage handling process in an airport. Some studies used window-based encoding for predicting the time-related indicators [36,37]. In this method, the size of the window is specified and the number of events to be encoded and entered into the algorithm is reduced. One of the applications of predicting the completion time of processes is to prevent the breach

of service level agreement (SLA) in the information technology (IT) industries. In this context, Mehdiyev et al. [38] predicted an SLA breach in the Volvo Group using a K-means clustering method. They extracted the effective features by implementing a deep Neural Network (DNN) and Surrogate Decision Trees (SDT). Also, Costache et al. [39] Also, used the STEP method for the encoding phase and predicted the workload by focusing on the intra-case features.

In the systematic review conducted by Maita et al. [40], prediction has been considered one of the main topics of mining in the field of process mining, in which traditional techniques based on process graphs have been of more interest than machine learning methods.

The following table (cf. Table 1) summarizes the studies in predictive process monitoring literature. The prediction tasks have been performed by using classification methods such as decision trees, random forests, and neural networks. As can be seen, many of these researches answer the question, what is the outcome of a process? What is the next activity/event? When is the completion time of a running case?

As highlighted in recent studies, most of them have used only supervised learning methods to make predictions. Since real-world processes are not fully structured, mining a process model that covers all possible traces will result in a complex and spaghetti-like process model. Then the implementation of the prediction model will have lots of errors and be useless. One of the proposed solutions to solve this problem is the use of clustering methods on different process paths [41].

Unsupervised learning methods such as clustering methods are important because different processes are extracted due to the various traces in an event log Some of these execution paths are more similar and located in a cluster, and classification models fit into each cluster [4].

Due to the significant diversity in the journey traveled by the customers, hybrid supervised/unsupervised machine learning methods are used to deal with the problem of high intra-class variance [42]. The variation between multiple traces of a label is the intra-class variance that defines the performance of a model. The low value of the intra-class variance shows the repeatability of the test which means the closeness between the results of successive tests [43]. Furthermore, almost none of the previous studies have paid attention to the usefulness of the attributes used to develop such models which determine the accuracy levels of prediction models [44]. Data mining procedures remove extraneous attributes since some provide little (or no) information and may even overshadow significant ones. In addition to the limited consideration of outcome prediction, existing studies almost exclusively focus solely on binary classification and neglect other possible and relevant outcomes. This is a relevant problem when the outcome of a process, i.e., customer satisfaction, is defined as a numerical interval.

To address the mentioned research gaps, the current research study has proposed a prediction approach that uses both supervised and unsupervised learning methods. In this regard, DBSCAN clustering is implemented to group similar traces, then the attribute selection technique is conducted and activities with no added value are removed. DBSCAN has several advantages over other clustering algorithms. It can handle clusters of arbitrary shape and can identify clusters of varying densities. Additionally, it does not require the number of clusters to be specified beforehand, as the algorithm can discover the number of clusters automatically. Also, a zero-to-five numerical interval is considered as the outcome, and a multi-class classification is implemented to predict the outcomes of customer journeys. The next section explains the preliminaries of the proposed solution method for customer journey analytics.

3. Preliminaries

This section discusses the preliminaries and fundamental concepts of the research problem. The customer journeys and customer experience are basic components of any e-commerce business environment. Process mining techniques can add significant value for analyzing the existing and future interaction patterns of customers. In this study, the exploration of the journeys is handled by user objectives to predict a specific outcome of the process at different levels of granularity of the event log.

3.1. Customer journey

A customer journey is a sequence of interactions between a customer and a service provider that calls them “touch points”. This journey tells the story of the customer’s experience, from the first encounter until a long-term relationship is formed. Customer journeys include the following elements [57] (cf. Table 2).

3.2. Event log (ϵ)

An event log is a set of data stored in the information systems based on the observed behaviors of different cases [58]. An event log captures data related to each event or activity, including the time of occurrence, the type of event, and any associated data or details [59].

Event logs are inputs of process mining algorithms. Event logs include the following components [20,24,60] (cf. Table 3). In customer services, an event log is a record of all customer interactions and related activities that occur during the execution of customer service processes. It captures data related to each interaction, including the time of occurrence, customer details, type of interaction, and any associated data or details. An event log of customer services can help organizations analyze customer interactions, identify customer needs, and optimize customer service processes.

3.3. Predictive business process monitoring

When executing a redesigned business process, the new process may not meet expectations. For example, unforeseen exceptions may occur that cause the processing time of some activities to be much longer than expected. These cases increase user dissatisfaction. The first step to addressing these issues, preventing and resolving them, is to understand what happens in reality. Process monitoring means using the data obtained from the execution of a business process, to extract insights about the actual performance of the process and its compliance with norms, policies, or regulations. The data resulting from the implementation of business processes is generally in the form of a set of event logs. Business process monitoring methods use incoming event logs and generate some artifacts to help process actors, analysts, process owners, and other managers gain a picture of process performance at various levels. Predictive business process monitoring means predicting the continuation of the running cases based on the models extracted from the historical event logs. This prediction includes various tasks such as predicting the next activity, the future path, the remaining cycle time, the outcomes of processes, etc. [61]. In the following, some functions that are commonly used in predictive process monitoring are explained.

Encoding function (f: $\epsilon \rightarrow$ X): This function receives an event log (ϵ) as input and provides a vector (X) that can be entered into the classification function. The following methods used in the literature for the encoding part:

- **One-hot:** one-hot encoding is one of the popular encoding techniques that has been applied to categorical features without any kind of order or relationship. It is a useful technique that enables machines to learn from datasets with categorical variables, helping to develop predictive models. It involves the conversion of categorical values into binary vectors, where each vector indicates the presence or absence of a particular category. In this method, each of the unique values of the mentioned variable is converted into a binary variable. According to the value of the variable, the numbers 0 or 1 are considered for each of the binary variables [62].

Table 1
A taxonomy of the existing literature on business process monitoring.

Article	Method				Data usage	Application/sector
	Encoding	Clustering	Feature selection	Classification		
Deng et al. [45]	LSTMED	–	FDR	LSTM	Tennessee Eastman (TE) process dataset	Time monitoring
Galanti et al. [46]	Not mentioned	–	–	Catboost	Italian utility provider company	Next activity prediction
Delias et al. [47]	Not mentioned	–	–	LR	Claims management process dataset	Outcome prediction
Mehdiyev et al. [48]	Not mentioned	–	SHAP	QRF	Manufacturing Execution Systems	Run-time prediction
Bozorgi et al. [49]	One-hot	–	–	Not mentioned	BPIC16, 17, 19, 20	cycle time and outcome prediction
Kim et al. [24]	index-based	–	Resource-aware	GB, RF	BPIC11, 12, 15	Outcome prediction
Amponsah et al. [50]	Not mentioned	–	–	DT	NHIS claims process	Fraud detection
Lee et al. [27]	Index-based, One-hot	prefix-length bucketing		RIPPER, XGB, RF	BPIC11,12,15,17	Outcome prediction
Weinzierl et al. [33]	Boolean	–	–	CNN	BPIC12,13,19,20	Workaround detection
Jalayer et al. [21]	One-hot		SoftMax	LSTM	Helpdesk, BPIC12,15,17	Next activity prediction
Elkhawaga et al. [25]	aggregation-based, index-based	–	SHAP, PDP	XGB, LR	Sepsis1, 2, 3, BPIC17, Traffic fines, Hos_billing1, 2	Outcome prediction
Pasquadibisceglie et al. [20]	One-hot	–	–	LSTM	BPIC12, 13, 17, 20, CoSeLoG	Next activity prediction
Mehdiyev et al. [38]	autoencoder	K-means	DNN	SDT	Volvo IT Belgium's incident management	SLA breach Prediction
Costache et al. [39]	STEP	–	Intra-case	MLR, 3-layer CNN	BPIC17	Workload prediction
Wang et al. [31]	Frequency-based	–	–	LSTM	BPIC12, 17, Sepsis cases, Production, Road Traffic Fines and Hospital Billing	Outcome prediction
Pauwels et al. [22]	One-hot	–	–	SDL, DBN, LSTM, CNN	Helpdesk, BPIC11, 12, 15	Next activity prediction
Pasquadibisceglie [32]	Image encoding		Autoencoder, Mutual info approach	CNN	SEPSIS, BPIC11, 12, Production	Outcome prediction
Mello et al. [23]	Boolean, Frequency-based	–	–	DT, RF, GB	IT department of a Brazilian organization	Predicting the failures
Taymouri et al. [51]	One-hot		GAN's discriminator	GNA	Helpdesk, BPIC12, 17	Next activity prediction
Xu et al. [52]	Frequency-based	–	–	DT, RF	Electronic Medical Record	Outcome prediction
Gusmao et al. [26]	Not mentioned	–	Pareto analysis	LR	CPFL Energia	Fraud detection
Gunnarsson et al. [36]	Window-based	–	–	LSTM	An airport's baggage system	Completion time prediction
Rizzi et al. [53]	Frequency-based, Simple-index, Complex-index	–	–	RF	BPIC11, Claim Management log	Outcome prediction
Pauwels et al. [54]	One-hot	–	–	DBN	BPIC12, 15, 18,	Next events prediction
Polato et al. [35]	One-hot	–	–	LR, CR, DATS	BPIC12, Help desk log, Road traffic log	Remaining time prediction
Francescomarino et al. [30]	Index-based	Canopy		HT, AT	BPIC11, 15, drift1, 2	Outcome prediction
Marquez et al. [37]	Window-based			DT, RT, ANN, SVM, EA	BPIC13, Incident management log	Run-time prediction
Tax et al. [55]	One-hot	–	–	LSTM	Helpdesk, BPIC12	The remaining time, the next event, and its timestamp

(*continued on next page*)

Table 1 (*continued*).

Article	Method				Data usage	Application/sector
	Encoding	Clustering	Feature selection	Classification		
Francescomarino et al. [28]	Frequency-based	DBSCAN, MBASED	–	DT, RF	BPIC11	Tumor diagnostic
Dadashnia et al. [56]	Boolean	–	–	LSTM	BPIC16	Next activity prediction
This study	Frequency-based	DBSCAN	Hybrid of variance and covariance analysis	DT	BPIC16	Outcome prediction

Table 2
Customer journey elements.

Title	Definition
Customer	A client receiving a service.
Journey	A common route was taken by a customer.
Touchpoint	An interaction between a customer and a service provider.
Timeline	The duration between the first and last touchpoints in a journey.

When one-hot encoding is to be used, the desired feature must be selected first. For example, when an activity is chosen for encoding, all distinct activities will be considered as columns. Table 4 is an example of an event log that contains three main columns: Case ID, Activity, and Timestamp. After implementing one-hot encoding method, Table 5 is obtained.

• **Frequency/Aggregation-based:** In this method, the output matrix includes unique Case IDs as rows and unique activities as columns. The internal cells are the frequency of each activity in each case journey [58,63]. The frequency-based encoding is shown in the following example (cf. Table 6, 7):

• **Boolean:** This method is similar to the previous method, with the difference that instead of the frequency of each unique activity in the journey of each case ID, the fulfillment of the activity or the non-fulfillment is shown with the numbers 1 and 0 [63] (For example cf. Table 8, 9).

• **Index-based:** In the index-based encoding method, the columns represent activities and features related to them [63]. Below is an example to explain this method, Act means activity and TS means time stamp (cf. Table 10, 11).

Classification function (y: X → Y): This function receives output from the encoding function (X) as input and delivers its class label (Y) as output. It should be noted that in the available studies, different supervised learning methods such as Long Short-Term Memory (LSTM), Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Dynamic Bayesian Networks (DBN), Single Dense Layer (SDL), Support Vector Machine (SVM), Random Forests (RF), Decision Trees (DT), Logistic Regression (LR), Multiple Linear Regression (MLR), Contextual Regression (CR), Data-aware Transition System (DATS), GBoost, XGBoost have been used to predict the desired task.

4. Problem scope and definition

Nowadays, extracting knowledge from historical data has been noticed. Process mining is a new concept for discovering process models from event logs and can be used as a suitable method for extracting knowledge from customer journeys. By understanding the behavior of customers, business owners can imagine themselves in the position of their customers and, as a result, have a more accurate prediction of their needs and desires.

Customer interactions with service providers can be considered as sequences of events. Customers follow certain paths to accomplish a particular outcome [11]. With the expansion of businesses that provide online services, the journey of customers takes place on the websites. In this research, the application of process mining in analyzing the journeys of a website's customers is investigated. By using predictive business process monitoring, an attempt has been made to predict the output of the process and the journeys taken by customers and to answer the research questions.

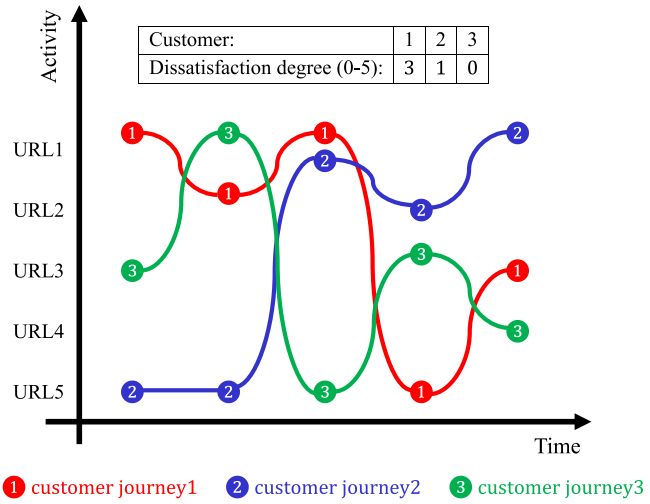

Fig. 1. An overview of the process outcome prediction.

One of the limitations that affect the problem of customer journey analysis is that users have different behaviors when visiting a website. There are many reasons why users have different behaviors when visiting a website. Some of the common factors include personal preferences, demographics, experience, the layout and navigation of a website, and accessibility. Because of this phenomenon, there is a lot of variation in the journeys taken within the identified classes, and a lot of dispersion in the class causes the durability and repeatability of the learning model to be less.

Websites consist of several pages that users visit and perform actions according to their needs. So at first, it is necessary to match the customer journeys, website visit data, and event log format (cf. Tables 12 to 14).

Customer journey heterogeneity refers to variations in the path that customers take when interacting with a business or brand. The problem with customer journey heterogeneity is that it can make it difficult for businesses to create a consistent and cohesive experience for their customers. When customers interact with a business, they may do so through a variety of touchpoints, such as a website, social media, email, phone, or in-person interactions. Each touchpoint can influence the customer journey, and customers can have vastly different experiences based on their preferred touchpoints, their needs, and their behaviors.

To deal with huge variability in process instances, in this research, an attempt has been made to predict the degree of dissatisfaction by analyzing the customer's journey on online sites and to improve the customer's user experience by using the obtained results. In this regard, according to the registration of complaints by customers, the level of dissatisfaction is measured in a range of 0 to 5 (cf. Fig. 1).

Table 3
Event logs: definitions, notations, and structures.

Title		Symbol	Definition
Event logs Columns	Activity	*A*	A step in a process. An event log typically has one activity column.
	Case ID	*ID*	A unique phrase or number for each instance of the process.
	Timestamp	*TS*	The time of occurrence of a particular event
	Case/event attributes	*D*	Attributes that change along the path of an event, such as the resource, are the event attributes, else, gender and age are the case attributes. An event log typically has several cases and event attributes.
Event logs Rows	Event	*e*	A change in the state of a process, such as a start or completion of an activity, an input or output message, a timeout violation, etc. $e = (A, ID, TS, D_1, \ldots, D_m) \quad m \geq 0$
Trace		σ	The sequence of events $\sigma = [e_1, \ldots, e_i] \vee i \in [1,n]$
Variant		*S*	A set of all unique traces in the event log.

Table 4
An event log.

ID	Act	TS
1	A	09:20
2	B	09:19
2	C	09:20
3	C	09:25

Table 5
The event log encoded by the one-hot encoding method.

ID	A	B	C
1	1	0	0
2	0	1	1
3	0	0	1

Table 6
An event log.

ID	Act	TS
1	A	09:20
2	B	09:19
2	A	09:20
2	B	9:26
1	B	09:25

Table 7
The event log encoded by the frequency-based encoding method.

ID	A	B
1	1	1
2	1	2

Table 8
An event log.

ID	Act	TS
1	A	09:20
2	B	09:19
2	A	09:20
2	B	9:26

Table 9
The event log encoded by the boolean encoding method.

ID	A	B
1	1	0
2	1	1

Table 10
An event log.

ID	Act	TS
1	A	09:20
2	B	09:19
2	A	09:20
1	B	09:25

Table 11
The event log encoded by the index-based encoding method.

ID	Act1	TS1	Act2	TS2
1	A	09:20	B	09:25
2	B	09:19	A	09:20

Table 12
A website visit dataset example [4].

User ID	URL	Time
1	.com/search	11-09-2022:20.08
2	.com/product	12-09-2022:16.11
2	.com/details	12-09-2022:21.46

Table 13
An example of an event log [4].

Case ID	Activity	Time
1	a	11-09-2022:20.08
2	b	12-09-2022:16.11
2	c	12-09-2022:21.46

Table 14
Matching event log and website visit dataset formats.

Event log	Website visit data	Customer Journey
Case ID	User ID	Customer
Activity	URL	Touchpoint
Timestamp	Time	Timeline

5. Process mining methodology

The foundation of the prediction approach and the outline of the tasks performed to accomplish the objectives of this research are explained in this section. Fig. 2 provides an overview of the hybrid supervised/unsupervised process mining approach which includes several computing modules, i.e., event log preprocessing, data transformation, trace clustering, feature engineering, and finally classification.

This approach can combine the strengths of both techniques to achieve more accurate and efficient results. Supervised process mining involves using labeled data to train a machine-learning model to recognize patterns and make predictions about future events. In contrast, unsupervised process mining involves analyzing data without prior knowledge or labeling, to discover hidden patterns and insights.

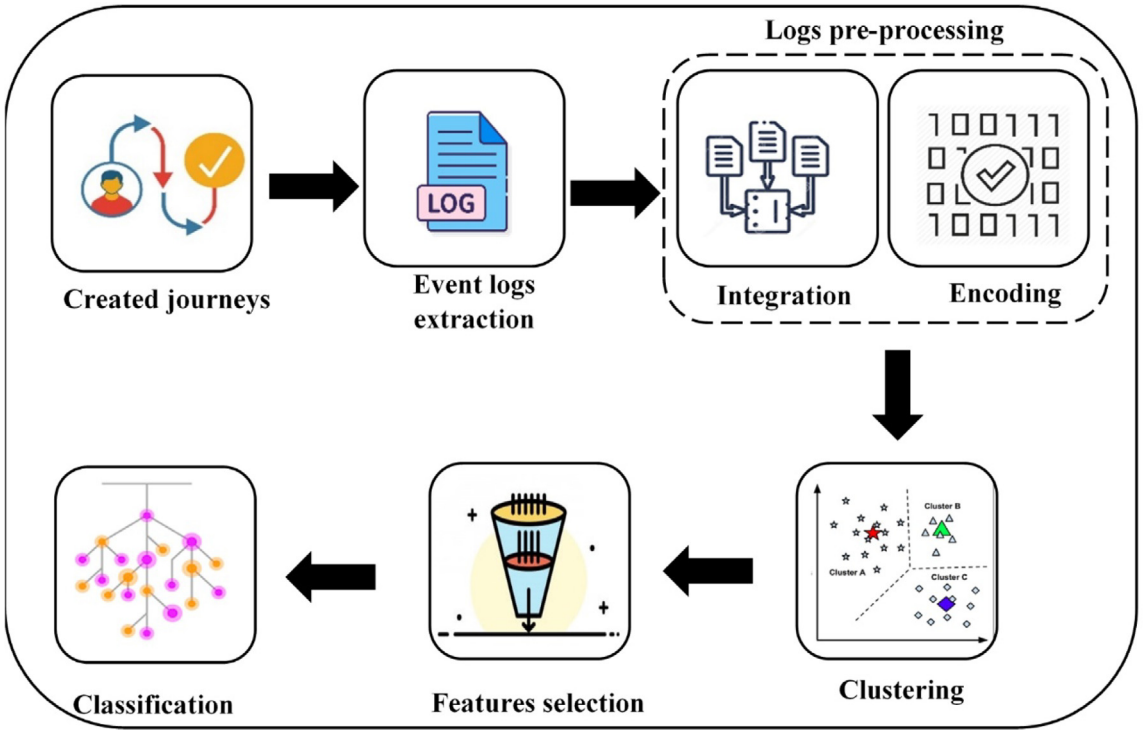

Fig. 2. An overview of proposed process mining approach.

In the first step, customer journeys are created by users' behavior and formed event logs. The event logs extracted from information systems are encoded using the Frequency-based method and transformed into the appropriate format for entering the learning algorithm. Next, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is used as the clustering method based on the similarities between the journeys. Also, silhouette index analysis has been used to adjust the parameters to find the appropriate values of the input parameters to the clustering algorithm and create the appropriate clusters. Analysis based on variance and covariance is performed to select the most effective attributes of each cluster. Finally, using supervised learning methods (multi-class decision trees), the output of customer journeys is predicted.

In the present setting, the input consists of an event log representing the traces and the attributes of the matching event sets. The output is a set of learned decision trees for each corresponding customer journey cluster. The pseudo-code of the proposed process mining approach is defined as follows (cf. Tables 15): (see Fig. 3).

The following notations to describe the approach for building a prediction model were required:

Encoding function (ENCODING (ε)**):** the function takes the event log as an input and returns an encoded matrix. In this study, the frequency-based encoding method is applied.

Clustering function (EXTRACT CLUSTER (trace $_{\text{encoded}}$, clustering parameters)**):** this function takes an encoded matrix and clustering parameters as inputs and returns clusters as outputs. In this study, The DBSCAN (Density-based spatial clustering of applications with noise) clustering method has been implemented. DBSCAN is a popular unsupervised clustering algorithm that is used to group data points that are close to each other in a given dataset. DBSCAN operates by grouping data points that are close enough to each other and have enough other data points in their vicinity. The algorithm defines two parameters, epsilon, and minimum points, to determine what counts as a cluster. Epsilon is the distance between two points below which they are considered neighbors, while the minimum points parameter determines the minimum number of points needed to form a cluster. Points that do not belong to any cluster are considered outliers or noise. It uses epsilon as the maximum radius of each cluster and the minimum number of cases in a cluster; then, based on similarities between journeys, it returns clusters that instants in each one have similar journeys. The basic steps of the DBSCAN algorithm are:

1. Choose a random unvisited data point.
2. Retrieve all its neighboring points within a distance of epsilon.
3. If the number of neighboring points is greater than or equal to the minimum points parameter, then a new cluster is formed.
4. If the number of neighboring points is less than the minimum points parameter, the point is marked as noise.
5. Repeat the process until all data points have been visited.

Each parameter of the function can take various values that change the result in clustering. In this study, to find the most suitable clustering, the analysis of the silhouette score has been used. The silhouette score takes different values for different values of the epsilon and the minimum number of samples, but the closer this score is to one, the more favorable it is. According to the pseudo-code below, by calculating this index for different values of the epsilon and the minimum number of samples, an attempt has been made to determine the best values of the parameters (cf. Tables 16) (see Fig. 4).

Feature selection functions (drop_event_with_zero_frequency(cluster) & drop_event_with_low_variance(cluster without event with frequency 0) & drop_correlated_event (cluster without event with low variance)**):** the functions take the clusters from the clustering functions as inputs, remove inefficient, redundant and correlated attributes (activity) and returns lower dimension matrixes. In this study, some statistical analyses (variance and covariance-based feature selection) according to the pseudo-code below have been implemented (cf. Tables 17): (see Fig. 5).

Classification function (BUILD TREE (Feature selected Cluster, Event)**):** Finally, the label is predicted by implementing the selected learning algorithm on the clusters. In this study, the decision trees method is used. In this regard, 70% of the event log was separated for training and 30% for testing. Then, the model was evaluated, and

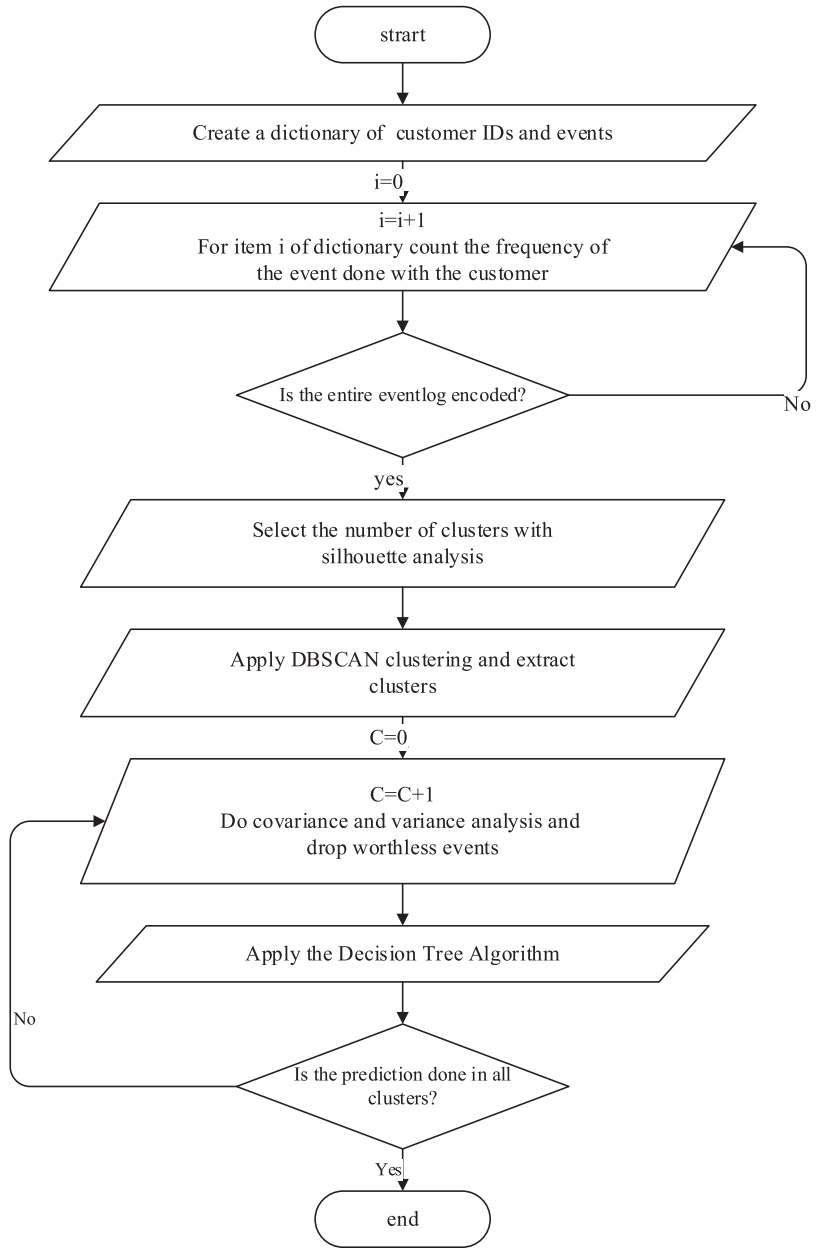

Fig. 3. Flowchart of the prediction model.

the following indicators based on the confusion matrix were calculated [23] (cf. Fig. 6):

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

6. Case study: online insurance services

In this section, the proposed predictive process monitoring approach was applied to a real event log. For this purpose, An event log that was released in the sixth International Business Process Intelligence Challenge (BPIC'16) is used [64].

The BPI Challenge event logs provide valuable advantages for research in process mining and business process management. These logs offer real-world data, covering diverse business processes from different industries [65]. Since the RQs focus on customer journeys and complaints registration on an online site, a dataset extracted from the online site was required. The published event log in the BPIC'16 constitutes genuine and standardized data. This data represents the authentic behavior of users, rendering the results applicable in real-world scenarios [49].

The event log is related to the behavior of users in eight months in an online provider of employment and insurance services (UVM) in the Netherlands. The discovered process map of the activity event log for clicks logged in is presented in Fig. 7.

Several data sources are used to collect the required information.

(1) Customer click data from the site www.werk.nl

(2) Message data, showing when applicants contacted the agency through a digital channel,

(3) Call data from the call center, presenting when applicants contacted the call center by cell phone,

(4) Complaint data showing when applicants complained.

Table 15
Pseudocode for the prediction model.

Algorithm1. A process mining approach to analyze the customer journey

```
Input : ε, event log data set
Output : Decision trees, a set of decision trees for each cluster
trace encoded = ENCODING (ε)
ID = set of unique Case ID in ε
Event= set of unique Events in ε
Initialize distance = empty matrix |ID|*|ID|
distance ← Euclidean distance (trace encoded)
Function TUNING (distance)
        Return eps, MinPts
Function EXTRACT CLUSTER (trace encoded, eps, MinPts)
        Return Clusters
cluster=1
Do
        Function drop_event_with_zero_frequency (cluster)
        Return cluster without event with frequency 0
        Function drop_event_with_low_variance (cluster without event with frequency 0)
        Return cluster without event with low variance
        Function drop_correlated_event (cluster without event with low variance)
        Return cluster without correlated event
        Feature selected Clusters.Append(cluster without correlated event)
        cluster= cluster +1
While cluster <=|Clusters|
Feature selected Cluster = 1
Do
        Function BUILD TREE (Feature selected Cluster, Event)
        Return decision tree
        Decision trees.Append(decision tree)
        Feature selected Cluster = Feature selected Cluster +1
While Feature selected Cluster<=|Feature selected Clusters|
```

Table 16
Pseudocode for parameters tuning of the clustering phase.

Algorithm1.1. DBSCAN parameters tuning

```
Input : trace encoded
Output : eps, optimum maximum radius for each cluster
         MinPts, the optimum minimum number of points in each cluster
Initialize distance = empty matrix |ID|*|ID|
distance ← Euclidean distance (trace encoded)
Function TUNING (distance)
    For i=0 : maximum (distance)
        For j = 1 : |ID|
            Calculate silhouette score
        End For
    End For
    Identify the best silhouette score
Return eps, MinPts
```

In this context, we intend to gain knowledge about the different behavioral patterns of applicants and to determine how different channels are being used.

The data set covers the following main objects and their corresponding attributes and events:

– Customer - client of a Dutch public agency for handling unemployment benefits which include several attributes such as customerID, age category, and gender

– Session - browser-session identifier of a user browsing the website of the agency

– IP - IP address of an applicant browsing the website of the agency

– Office_U - user involved in an activity handling an applicant interaction

– Office_W - worker involved in an activity handling an applicant interaction

– Complaint - a complaint document handed in by an applicant

– ComplaintDossier - a collection of complaints by the same applicant [66]

The event log, as mentioned above, includes the behavior of 27,412 users in CSV format. In the eight months of data collection, 815 activities (site page visits) were executed by them. In total 26,477 process variants (S) are detected in this event log (cf. Table 18).

The goal is to predict user satisfaction using extracted customer journeys from the event log. In this event log, there is the activity of registered complaints by users. Thus, an attempt has been made to predict the level of user satisfaction by using predictive process monitoring and multi-class classification (cf. Fig. 1).

Encoding: The mentioned event log (ε) contains 7,364,684 events (rows), which after entering the encoding function (ENCODING (ε)), a matrix (trace $_{encoded}$) with dimensions of 27412*815 is formed. To evaluate the proposed model, first, the usual approach of predictive process monitoring has been implemented on this event log. In fact, without clustering, the decision tree algorithm has been implemented. By performing the classification after encoding, the following results are obtained (cf. Table 19). As it is shown, the false positive error (FP) for classes 1 to 5 is significant and this indicates the problem of intra-class variance. To solve this problem, a hybrid of supervised and unsupervised machine learning methods is implemented.

Clustering:

After performing the encoding step, clustering is handled based on the similarity between the customer journeys. In this regard, The DBSCAN clustering method is applied. It is necessary to set the appropriate input clustering parameters, including the radius and the

Table 17
Pseudocode of the feature selection phase.

Algorithm1.2. Feature selection in cluster

```
Input : Clusters, a set of extracted clusters from the DBSCAN algorithm
Output : Feature selected Clusters
cluster=1
Event set= set of unique Events in the cluster
Do
    Function drop_event_with_zero_frequency(cluster)
        For each event in the Event set
            if the frequency of the event in all ID=0
                Drop event from the cluster
            End if
        End For
    Return cluster without event with frequency 0
    Event= set of unique Events in the cluster without event with frequency 0
    Function drop_event_with_low_variance(cluster without event with frequency 0)
        Variance=empty matrix |1|*|Event|
        For each event_i in the Event set
            Variance_1i ← calculate the variance event
        End For
        For each event_i in the Event set
            if Variance_1i<=Q1 of variance
                Drop event_i from the cluster
            End if
        End For
    Return cluster without event with low variance
    Event= set of unique Events in the cluster without events with low variance
    Function drop_correlated_event (cluster without event with low variance)
        Covariance=empty matrix |Event|*|Event|
        For each event_i in the Event set
            For each event_j in the Event set
                Covariance_ij ← calculate covariance event i and event j
            End For
        End For
        For each event_i in the Event set
            For each event_j in the Event set
                if covariance_ij<=Q1 of covariance or covariance_ij>=Q3 of covariance
                    Drop event_ij from the cluster
                End if
            End For
        End For
    Return cluster without correlated event
    Feature selected Clusters.Append(cluster without correlated event)
    cluster=cluster+1
While cluster<=|Clusters|
```

Table 18
Descriptive statistics of the BPI Challenge 2016 dataset.

Indicator	Trace length ($\|\sigma\|$)	Activity frequency	Euclidean distance
Min.	1	1	0
Max.	9,719	1,748,353	6,339.06
Avg.	268.6	903.6	122.62

Table 19
Confusion matrix before applying the clustering technique.

Class	Confusion matrix			
	TP	FP	FN	TN
0	8085	0	100	3
1	1	126	0	8071
2	0	41	0	8198
3	0	4	0	8218
4	0	3	0	8222
5	0	0	0	8223

minimum number of samples in the clusters. To adjust these parameters, the Silhouette-score analysis was used and the following results were obtained.

In the first step, the minimum and maximum possible radius are determined by calculating the Euclidean distance between different journeys. The below chart shows the distribution of computed distances between different customer journeys. As can be seen, the radius can be checked from 0 to 6,339.06, and the minimum number of samples from 0 to 27,412.

This analysis is done to find the best combination of the radius and the minimum number of samples in the clusters to maximize the Silhouette-score value. As reported in Table 13, the silhouette score and the number of clusters are obtained by considering the radius and minimum number of points. Finally, ten combinations are found, which are reported in Table 13. The best available combination is the radius of 247 and the minimum sample of 4, which results in a Silhouette score of 0.9045 and 3 clusters (cf. Table 20).

Feature selection: In this step, the effective attributes in each cluster have been identified and selected by calculating the variance of each attribute and the covariance between them. In the first step,

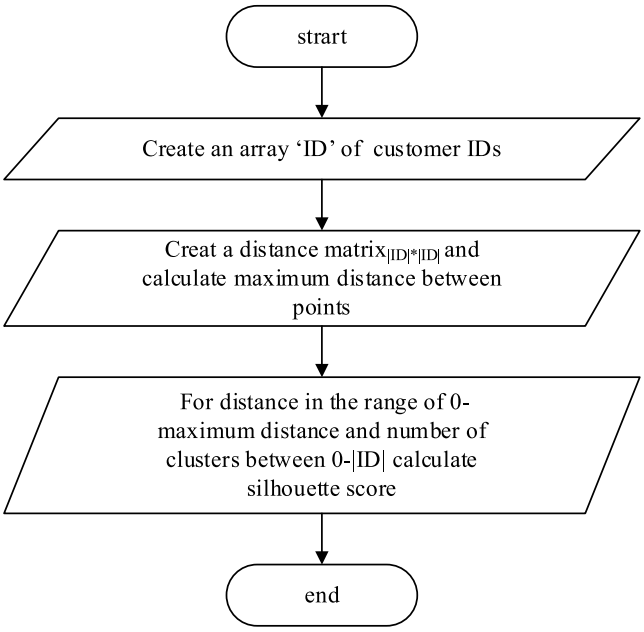

Fig. 4. Flowchart of DBSCAN parameters tuning.

Table 20
Silhouette-score analysis.

Radius	#Min-points	#Clusters	#Noise	Silhouette-score
247	4	3	97	0.9045
229	3	3	106	0.9029
156	6	2	252	0.8585
249	5	2	99	0.8575
249	4	4	90	0.8572
249	3	4	89	0.8516
130	5	2	413	0.8259
130	4	3	401	0.8151
200	2	7	123	0.7847
201	2	7	123	0.7847
249	1	88	0	0.6719

attributes with little variance are removed, and attributes with a very high positive or negative correlation are removed among the remaining attributes (cf. Table 21).

It should be noted that the meaning of the feature in the matrix encoded is the columns of this matrix, which are the URLs of different site pages.

In this table, in the first column, the number of activities observed in each cluster is listed, after calculating the variance of each activity (see Table 17), the number of activities that have been identified as having little variance should be removed. In the next step, the number of activities that are correlated with each other is given in the fourth column, and after removing all of them, the number of valuable and remaining activities is specified in the last column.

Classification: at the final step, for predicting the level of dissatisfaction of users, a scale of 0 to 5 was considered, and the following results were obtained using the multi-class decision tree algorithm in each cluster. For this purpose, the event of 'Complaint' is considered as a label in this dataset, and the frequency of doing this activity will indicate the degree of dissatisfaction.

To evaluate the proposed learning model, the data in each cluster has been split into training and testing data. Finally, the accuracy, recall, and F1-score were measured by creating confusion matrixes, and the following results were obtained (cf. Table 22). In clusters 2 and 3, only classes 0 and 1 were present in the training data; therefore, only these classes were predicted in the test data.

In this table, after implementing the classification algorithm in each cluster, the values of the confusion matrix were calculated for each class. After calculating these values, the accuracy indices including accuracy, recall, precision, and F1-score were obtained.

As it is known, due to the unbalanced data in cluster C1 (distribution of classes: {0:27081},{1:180},{2:35},{3:6},{4:4},{5:1}), the model is biased to class 0. After classifying the data, the accuracy index in this class is very high, and the indices in other classes are not suitable; for this reason, over-sampling calcification for the imbalanced data technique has been implemented [67]. It can improve the model, and the results of the following table have been obtained (cf. Table 23).

Similar to the previous table in this table, after implementing the classification algorithm using the oversampling technique in each cluster, the values of the confusion matrix were calculated for each class. After calculating these values, the accuracy indices including accuracy, recall, precision, and F1-score were obtained.

In the research conducted by Dadashnia et al. [56], a model based on the LSTM method was implemented on the BPIC16 dataset. In comparing the accuracy, their model stands at 64%, this paper with the proposed method achieving a significantly higher accuracy of 98%, it is evident that the suggested approach outperforms the current standard by a substantial margin. This notable difference underscores the superior predictive capabilities and effectiveness of the new method in customer journey analytics. By leveraging a hybrid of supervised and unsupervised learning techniques, the proposed model not only demonstrates a remarkable accuracy rate but also addresses the challenge of high intra-class variance, a common issue in process monitoring. The results clearly indicate that the new approach is not only proficient in making accurate predictions but also excels in providing valuable insights into customer satisfaction levels. This substantial improvement in accuracy highlights the potential of the proposed method to revolutionize predictive process monitoring in online services, paving the way for enhanced customer experiences and improved decision-making for service providers.

7. Conclusion and policy implications

The benefits of customer journey prediction are numerous, including higher customer loyalty, increased customer satisfaction, and greater revenue growth. By predicting customer behavior, companies can more effectively engage with customers, anticipate their needs, and provide personalized experiences that are likely to drive customer retention and business growth. Process mining can provide valuable insights for exploring and predicting customer journeys by visualizing customer behavior, identifying areas for improvement, and measuring the effectiveness of the customer journey. One of the important applications of process mining is improving process performance by using predictive business process monitoring that helps to gain insightful knowledge about the features of processes. This study employs a novel hybrid computing method of supervised/unsupervised machine learning techniques for predicting the outcomes of process instances under the high degree of internal heterogeneity in process variants.

This paper underscores the importance of customer satisfaction analysis in online services. By harnessing the power of machine learning and customer journey analysis, it empowers service providers to gain deep insights into user experience, identify areas for improvement, and ultimately enhance the overall experience [14]. This research holds the potential to guide online service providers towards more informed decision-making and, consequently, greater customer satisfaction and loyalty.

The study aims to fill the existing research gap by answering recent calls for research about a more thorough exploration of customer

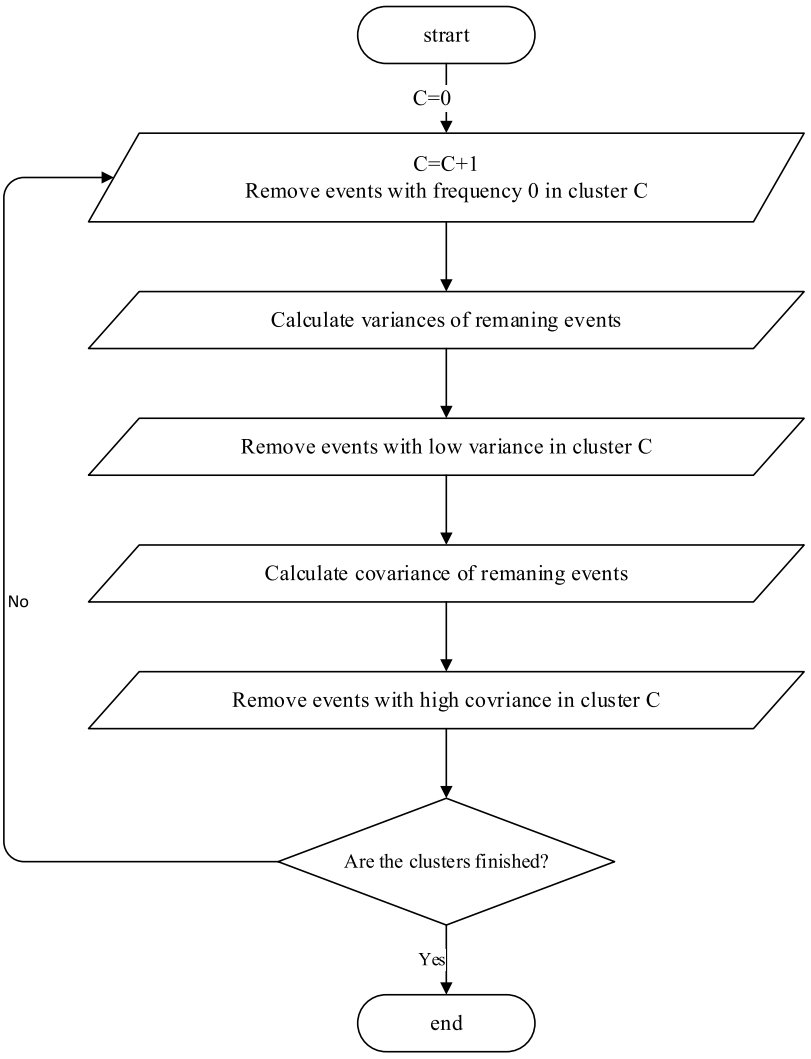

Fig. 5. Flowchart of feature selectin phase.

Table 21
Feature selection in the clusters.

Clusters	#total attr.	#attr. with low var.	#correlated attr.	#remaining attr.
C1	807	215	114	478
C2	69	21	10	38
C3	90	29	30	31

behavior by using predictive process monitoring techniques. The authors use process mining to explore customer journeys, identify customer behavior patterns, and measure the efficiency of the customer journey. The study focuses on two research questions: (1) how can supervised and unsupervised machine learning algorithms be used to analyze customer journeys and their patterns to predict the behavior of website users? and (2) how can process mining and predictive process monitoring algorithms be used to predict whether or not users will submit a complaint? The study's contributions include proposing a novel predictive model to fill the existing research gaps, moving from reactive actions to proactive behavior, improving the classification method by considering a range of outcomes, reducing the created clusters' dimensions, and increasing the prediction accuracy.

The purpose of the proposed approach is to predict the outcome of journeys taken by customers. In this regard, by Considering the journey traveled by customers as a sequence of events, the proposed predictive process monitoring approach has been used to predict the outcome of the customer journey. First, the journeys taken by customers were clustered. And then, low-value activities in each cluster were removed. To label the outcomes of journeys, a multi-class classification method was used. Using the multi-class classification on the BPI Challenge 2016 dataset, the level of customer dissatisfaction and complaint registration by them has been predicted.

According to earlier research in the field of prediction business process monitoring, machine learning methods have been widely used to make various prediction tasks. The researchers aim to enhance service

Table 22
Classification results.

Clusters	Class	Confusion matrix				Accuracy	Recall	Precision	F1-score
		TP	FP	FN	TN				
	0	8023	62	108	0	0.9792	0.9867	0.9923	0.895
	1	0	92	50	8051	0.9826	0	0	–
C1	2	0	14	9	8170	0.9971	0	0	–
	3	0	1	1	8191	0.9997	0	0	–
	4	0	1	1	8191	0.9997	0	0	–
	5	0	0	1	8192	0.9998	0	–	–
C2	0	1	0	0	0	1	1	1	1
C3	1	1	0	0	0	1	1	1	1

Table 23
Classification results after applying the over-sampling technique.

Clusters	Class	Confusion matrix				Accuracy	Recall	Precision	F1-score
		TP	FP	FN	TN				
	0	7835	0	174	40 737	0.9964	0.9782	1	0.9890
	1	8084	126	0	40 536	0.9974	1	0.9846	0.9922
C1	2	8080	41	0	40 625	0.9991	1	0.9949	0.9974
	3	8189	4	0	40 553	0.9999	1	0.9995	0.9997
	4	8171	3	0	40 572	0.9999	1	0.9996	0.9998
	5	8213	0	0	40 533	1	1	1	1
C2	0	1	0	0	0	1	1	1	1
C3	1	1	0	0	0	1	1	1	1

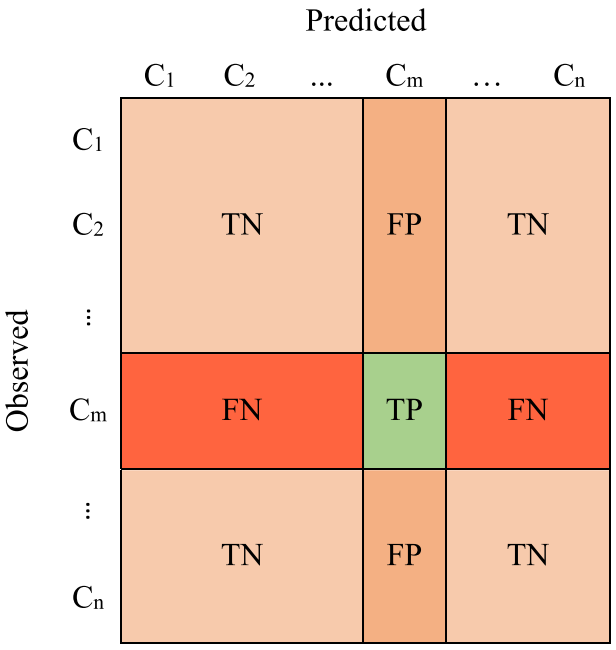

Fig. 6. Confusion Matrix for multi-class classification.

quality and innovation by anticipating potential outcomes in processes. In this study, a real-life event log is used to evaluate the proposed approach. The research finding shows an accuracy of 0.99 in predicting customer satisfaction. In predicting outcomes of processes, researchers have favored the supervised learning method. However, as noted in the literature review section, one challenge in analyzing customer journeys and predicting customer behavior is the high variability in extracted journeys taken by customers (referred to as the high intra-class variance problem). Due to the dispersion in the traces identified in the customer journey analysis, this article proposes using a combination of supervised and unsupervised learning methods to address the issue of high intra-class variance. The experiment demonstrates that our strategy outperforms well. The model is useful for making predictions, but it also has value on its own because it can show the level of consumer satisfaction.

In future work, we intend to work more on overcoming the imbalance condition in an event log of the customer journeys. In this regard, we will use reinforcement learning, etc. as a classification method. Additional directions for further work include the extension of the proposed approach to predict the continuation of the customer journey sequence by using recursive prediction methods. Finally, we want to look into how the proposed approach may be used to predict the following activities, their timestamp, and the cycle time needed to complete the process.

Reproducibility

The algorithm described in Section 5 is implemented using Python. The source code and supplementary material are publicly available at https://github.com/FtemehAkhavan/Predictive-Process-Monitoring-for-Predicting-Customer-Experience/tree/main#readme . This repository contains the code, input data, and configuration files necessary to reproduce the results.

Declaration of competing interest

We wish to confirm that there are no known conflicts of interest associated with this publication. There has been no significant financial support for this work that could have influenced its outcome.

Data availability

Data will be made available on request.

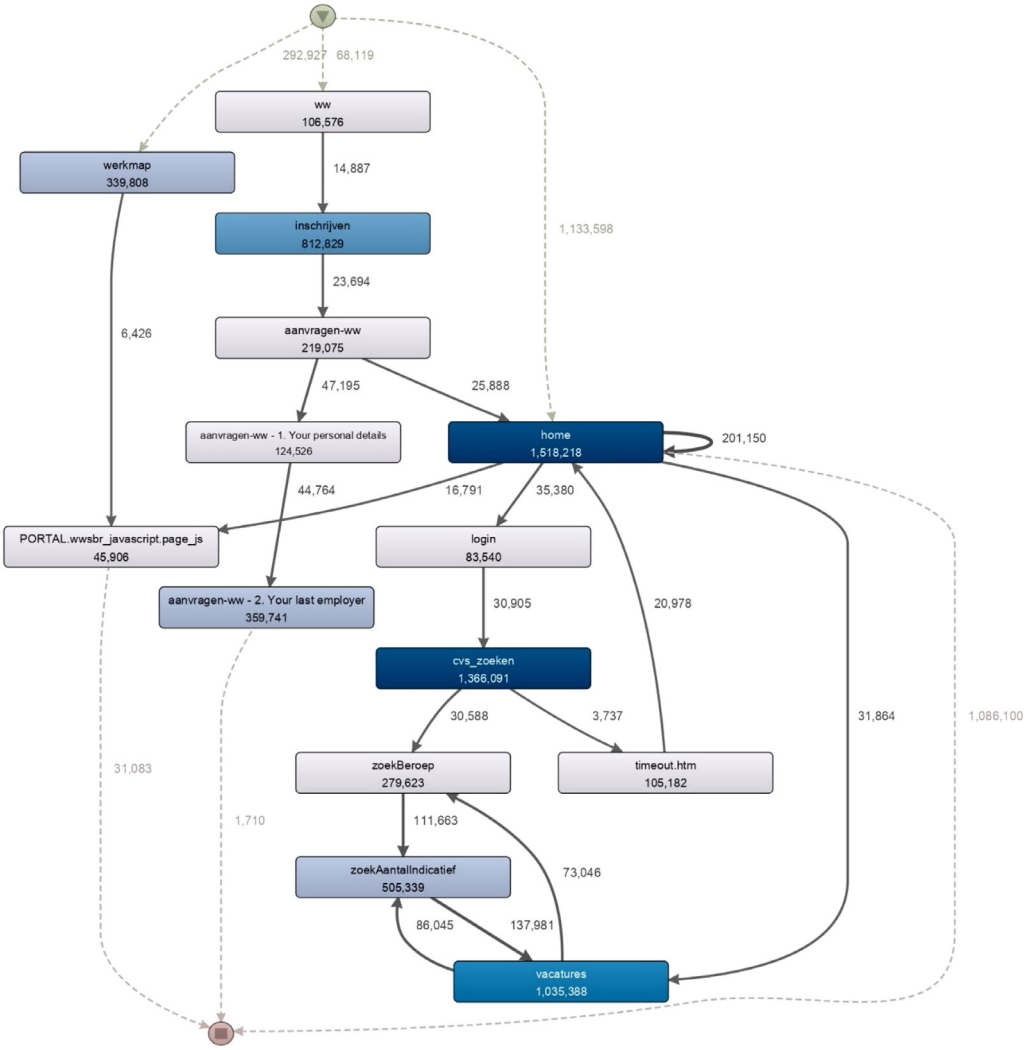

Fig. 7. The discovered process map of the activity event log for clicks logged in.

References

[1] B. Ramos Gutiérrez, A.M. Reina Quintero, L. Parody, M.T. Gómez López, When business processes meet complex events in logistics: A systematic mapping study, Comput. Ind. 144 (2023) http://dx.doi.org/10.1016/j.compind.2022.103788.

[2] W. Van Der Aalst, Process mining: Overview and opportunities, ACM Trans. Manag. Inf. Syst. 3 (2) (2012) http://dx.doi.org/10.1145/2229156.2229157.

[3] S.J.J. Leemans, S.J. van Zelst, X. Lu, Partial-order-based process mining: a survey and outlook, Knowl. Inf. Syst. 65 (1) (2023) http://dx.doi.org/10.1007/s10115-022-01777-3.

[4] A. Terragni, M. Hassani, Analyzing Customer Journey with Process Mining: From Discovery to Recommendations, 2018, http://dx.doi.org/10.1109/FiCloud.2018.00040.

[5] G. Bernard, P. Andritsos, Discovering customer journeys from evidence: A genetic approach inspired by process mining, in: Lecture Notes in Business Information Processing, vol. 350, 2019, http://dx.doi.org/10.1007/978-3-030-21297-1_4.

[6] G. Bernard, P. Andritsos, CJM-ab : Abstracting customer journey maps using process mining, in: International Conference on Advanced Information Systems Engineering, Vol. 1, 2018, pp. 49–56, http://dx.doi.org/10.1007/978-3-319-92901-9.

[7] M. Yari Eili, J. Rezaeenour, An approach based on process mining to assess the quarantine strategies' effect in reducing the COVID-19 spread, Libr. Hi Tech. 41 (1) (2023) http://dx.doi.org/10.1108/LHT-01-2022-0062.

[8] S.H. Hosseinizadeh Mazloumi, A. Moini, M. Agha Mohammad Ali Kermani, Designing synchronizer module in CMMS software based on lean smart maintenance and process mining, J. Qual. Maint. Eng. 29 (2) (2023) http://dx.doi.org/10.1108/JQME-10-2021-0077.

[9] N. Verma, J. Singh, A comprehensive review from sequential association computing to Hadoop-MapReduce parallel computing in a retail scenario, J. Manag. Anal. 4 (4) (2017) http://dx.doi.org/10.1080/23270012.2017.1373261.

[10] A. Terragni, M. Hassani, Optimizing customer journey using process mining and sequence-aware recommendation, in: Proceedings of the ACM Symposium on Applied Computing, Vol. Part F147772, 2019, http://dx.doi.org/10.1145/3297280.3297288.

[11] J. Goossens, T. Demewez, M. Hassani, Effective steering of customer journey via order-aware recommendation, in: IEEE Int. Conf. Data Min. Work. ICDMW, 2018-Novem, 2019, pp. 828–837, http://dx.doi.org/10.1109/ICDMW.2018.00123.

[12] M.D. Vollrath, S.G. Villegas, Avoiding digital marketing analytics myopia: revisiting the customer decision journey as a strategic marketing framework, J. Mark. Anal. 10 (2) (2022) http://dx.doi.org/10.1057/s41270-020-00098-0.

[13] M. Cordewener, Customer journey identification through temporal patterns and Markov clustering, 2016.

[14] S. Kumar, M. Zymbler, A machine learning approach to analyze customer satisfaction from airline tweets, J. Big Data 6 (1) (2019) http://dx.doi.org/10.1186/s40537-019-0224-1.

[15] M.R. Islam others, Discovering dynamic adverse behavior of policyholders in the life insurance industry, Technol. Forecast. Soc. Change 163 (2021) http://dx.doi.org/10.1016/j.techfore.2020.120486.

[16] S. Rawat, A. Rawat, D. Kumar, A.S. Sabitha, Application of machine learning and data visualization techniques for decision support in the insurance sector, Int. J. Inf. Manag. Data Insights 1 (2) (2021) http://dx.doi.org/10.1016/j.jjimei.2021.100012.

[17] P. Badakhshan, B. Wurm, T. Grisold, J. Geyer-Klingeberg, J. Mendling, J. vom Brocke, Creating business value with process mining, 2022.

[18] F. Spree, Predictive process monitoring : A use-case-driven literature review, in: EMISA Forum, 2020.

[19] W.M.P. Van Der Aalst, M.H. Schonenberg, M. Song, Time prediction based on process mining, Inf. Syst. 36 (2) (2011) http://dx.doi.org/10.1016/j.is.2010.09.001.

[20] V. Pasquadibisceglie, A. Appice, G. Castellano, D. Malerba, A multi-view deep learning approach for predictive business process monitoring, IEEE Trans. Serv. Comput. (2021) http://dx.doi.org/10.1109/TSC.2021.3051771.
[21] A. Jalayer, M. Kahani, A. Pourmasoumi, A. Beheshti, HAM-Net: Predictive business process monitoring with a hierarchical attention mechanism, Knowl.-Based Syst. 236 (2022) 107722, http://dx.doi.org/10.1016/j.knosys.2021.107722.
[22] S. Pauwels, T. Calders, Incremental predictive process monitoring: The next activity case, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12875, 2021, http://dx.doi.org/10.1007/978-3-030-85469-0_10, LNCS.
[23] P. Mello, K. Revoredo, F. Santoro, IT incident solving domain experiment on business process failure prediction, J. Inf. Data Manag. 11 (1) (2020) 34–49.
[24] J. Kim, M. Comuzzi, M. Dumas, F.M. Maggi, I. Teinemaa, Encoding resource experience for predictive process monitoring, Decis. Support Syst. 153 (2022) http://dx.doi.org/10.1016/j.dss.2021.113669.
[25] G. Elkhawaga, M. Abuelkheir, M. Reichert, Explainability of predictive process monitoring results: Can you see my data issues? 2022, pp. 1–32, [Online]. Available: http://arxiv.org/abs/2202.08041.
[26] L. Gusmao, H. Helito, T. Anarelli, J.R. Conceicao, T. Ji, G. Barros, A customer journey mapping approach to improve CPFL energia fraud detection predictive models, 2020, http://dx.doi.org/10.1109/TDLA47668.2020.9326214.
[27] S. Lee, M. Comuzzi, N. Kwon, Exploring the suitability of rule-based classification to provide interpretability in outcome-based process predictive monitoring, Algorithms 15 (6) (2022) 187, http://dx.doi.org/10.3390/a15060187.
[28] C. Di Francescomarino, M. Dumas, F.M. Maggi, I. Teinemaa, Clustering-based predictive process monitoring, IEEE Trans. Serv. Comput. 12 (6) (2016) http://dx.doi.org/10.1109/TSC.2016.2645153.
[29] M. Rafiei, W.M.P. van der Aalst, Group-based privacy preservation techniques for process mining, Data Knowl. Eng. 134 (2021) http://dx.doi.org/10.1016/j.datak.2021.101908.
[30] C. Di Francescomarino, C. Ghidini, A.I. Apr, Incremental predictive process monitoring : How to deal with the variability, 2018, arXiv Prepr. arXiv.
[31] J. Wang, Dongjin Yu*, Chengfei Liu, Xiaoxiao Sun, Predicting Outcomes of Business Process Executions Based on LSTM Neural Networks and Attention Mechanism, 2021.
[32] V. Pasquadibisceglie, A. Appice, G. Castellano, D. Malerba, G. Modugno, Orange: Outcome-oriented predictive process monitoring based on image encoding and CNNs, IEEE Access 8 (2020) http://dx.doi.org/10.1109/ACCESS.2020.3029323.
[33] S. Weinzierl, V. Wolf, T. Pauli, D. Beverungen, M. Matzner, Detecting temporal workarounds in business processes–A deep-learning-based method for analysing event log data, J. Bus. Anal. 5 (1) (2022) http://dx.doi.org/10.1080/2573234X.2021.1978337.
[34] R. Šperka, M. Halaška, The performance assessment framework (PPAFR) for RPA implementation in a loan application process using process mining, Inf. Syst. e-Bus. Manag. (2022) http://dx.doi.org/10.1007/s10257-022-00602-2.
[35] M. Polato, A. Sperduti, A. Burattin, M. de Leoni, Time and activity sequence prediction of business process instances, Computing 100 (9) (2018) http://dx.doi.org/10.1007/s00607-018-0593-x.
[36] B.R. Gunnarsson, S.K.L.M. vanden Broucke, J. De Weerdt, Predictive process monitoring in operational logistics: A case study in aviation, in: Lecture Notes in Business Information Processing, vol. 362, 2019, http://dx.doi.org/10.1007/978-3-030-37453-2_21, LNBIP.
[37] A.E. Márquez-Chamorro, M. Resinas, A. Ruiz-Cortés, M. Toro, Run-time prediction of business process indicators using evolutionary decision rules, Expert Syst. Appl. 87 (2017) http://dx.doi.org/10.1016/j.eswa.2017.05.069.
[38] N. Mehdiyev, P. Fettke, Explainable artificial intelligence for process mining: A general overview and application of a novel local explanation approach for predictive process monitoring, Stud. Comput. Intell. 937 (2021).
[39] I. Costache, Dirk. Fahland, A Process-Aware Perspective on the Use of the Performance Spectrum in Predictive Process Monitoring of Business Processes, 2021.
[40] A.R.C. Maita others, A systematic mapping study of process mining, Enterprise Inf. Syst. 12 (5) (2018) http://dx.doi.org/10.1080/17517575.2017.1402371.
[41] F. Folino, G. Greco, A. Guzzo, L. Pontieri, Mining usage scenarios in business processes: Outlier-aware discovery and run-time prediction, Data Knowl. Eng. 70 (12) (2011) http://dx.doi.org/10.1016/j.datak.2011.07.002.
[42] M. Hashemzadeh, B. Adlpour Azar, Retinal blood vessel extraction employing effective image features and combination of supervised and unsupervised machine learning methods, Artif. Intell. Med. 95 (2019) http://dx.doi.org/10.1016/j.artmed.2019.03.001.
[43] D.A. Reid, S. Samangooei, C. Chen, M.S. Nixon, A. Ross, Soft biometrics for surveillance: An overview, in: Handbook of Statistics, Vol. 31, 2013.
[44] C.A.L. Amaral, M. Fantinato, H.A. Reijers, S.M. Peres, Enhancing completion time prediction through attribute selection, in: Lecture Notes in Business Information Processing, vol. 346, 2019, http://dx.doi.org/10.1007/978-3-030-15154-6_1.
[45] W. Deng, Y. Li, K. Huang, D. Wu, C. Yang, W. Gui, LSTMED: An uneven dynamic process monitoring method based on LSTM and Autoencoder neural network, Neural Netw. 158 (2023) http://dx.doi.org/10.1016/j.neunet.2022.11.001.
[46] R. Galanti, M. de Leoni, N. Navarin, A. Marazzi, Object-centric process predictive analytics, Expert Syst. Appl. 213 (2023) http://dx.doi.org/10.1016/j.eswa.2022.119173.
[47] P. Delias, N. Mittas, G. Florou, A doubly robust approach for impact evaluation of interventions for business process improvement based on event logs, Decis. Anal. J. 8 (2023) 100291, http://dx.doi.org/10.1016/j.dajour.2023.100291.
[48] N. Mehdiyev, M. Majlatow, P. Fettke, Quantifying and explaining machine learning uncertainty in predictive process monitoring: an operations research perspective, 2023, pp. 1–43, [Online]. Available: http://arxiv.org/abs/2304.06412.
[49] Z. Dasht Bozorgi, I. Teinemaa, M. Dumas, M. La Rosa, A. Polyvyanyy, Prescriptive process monitoring based on causal effect estimation, Inf. Syst. 116 (2023) http://dx.doi.org/10.1016/j.is.2023.102198.
[50] A.A. Amponsah, A.F. Adekoya, B.A. Weyori, A novel fraud detection and prevention method for healthcare claim processing using machine learning and blockchain technology, Decis. Anal. J. 4 (2022) http://dx.doi.org/10.1016/j.dajour.2022.100122.
[51] F. Taymouri, M. La Rosa, Sarah Erfani, Zahra Dasht Bozorgi, Ilya Verenich, Predictive business process monitoring via generative adversarial nets: The case of next event prediction, in: International Conference on Business Process Management, in: LNCS, Vol. 12168, 2020, pp. 237–256, http://dx.doi.org/10.1007/978-3-030-58666-9_24.
[52] H. Xu, J. Pang, X. Yang, M. Li, D. Zhao, Using predictive process monitoring to assist thrombolytic therapy decision-making for ischemic stroke patients, BMC Med. Inform. Decis. Mak. 20 (2020) http://dx.doi.org/10.1186/s12911-020-1111-6.
[53] W. Rizzi, C. Di Francescomarino, F.M. Maggi, Explainability in predictive process monitoring: When understanding helps improving, in: Lecture Notes in Business Information Processing, vol. 392, 2020, http://dx.doi.org/10.1007/978-3-030-58638-6_9, LNBIP.
[54] S. Pauwels, T. Calders, Bayesian network based predictions of business processes, 2020.
[55] N. Tax, I. Verenich, M. La Rosa, M. Dumas, Predictive business process monitoring with LSTM neural networks, in: International Conference on Advanced Information Systems Engineering, Vol. 3, 2017, pp. 477–492, http://dx.doi.org/10.1007/978-3-319-59536-8.
[56] S. Dadashnia, J. Evermann, P. Fettke, P. Hake, N. Mehdiyev, T. Niesen, Identification of Distinct Usage Patterns and Prediction of Customer Behavior, 2016.
[57] G. Bernard, P. Andritsos, A process mining based model for customer journey mapping, in: CEUR Workshop Proceedings, Vol. 1848, 2017, pp. 49–56.
[58] I. Teinemaa, M. Dumas, M. La Rosa, F.M. Maggi, Outcome-oriented predictive process monitoring: Review and benchmark, ACM Trans. Knowl. Discov. Data 13 (2) (2019) http://dx.doi.org/10.1145/3301300.
[59] U. Singh, A. Muzaffar, R. Vyas, O.P. Vyas, Improving event log quality using autoencoders and performing quantitative analysis with conformance checking, 2023, http://dx.doi.org/10.1109/Confluence56041.2023.10048805.
[60] A. Senderovich, C. Di Francescomarino, F.M. Maggi, From knowledge-driven to data-driven inter-case feature encoding in predictive process monitoring, Inf. Syst. 84 (2019) http://dx.doi.org/10.1016/j.is.2019.01.007.
[61] N. Tax, I. Verenich, M. La Rosa, M. Dumas, Predictive business process monitoring with LSTM neural networks, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10253, 2017, http://dx.doi.org/10.1007/978-3-319-59536-8_30, LNCS.
[62] C. Seger, An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, Degree Proj. Technol. (2018).
[63] A. Leontjeva, R. Conforti, C. Di Francescomarino, M. Dumas, F.M. Maggi, Complex symbolic sequence encodings for predictive monitoring of business processes, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9253, 2015, http://dx.doi.org/10.1007/978-3-319-23063-4_21.
[64] M. Dees, B. van Dongen, BPI challenge 2016, in: 4TU, Centre for Research Data, 2016, Dataset, https://www.win.tue.nl/bpi/doku.php?id=2016:challenge.
[65] L. Blevi, L. Delporte, J. Robbrecht, Process mining on the loan application process of a Dutch Financial Institute BPI Challenge 2017, 2017, pp. 1–33, [Online]. Available: https://home.kpmg.com/be/en/home/insights/2017/09/process-mining.html.
[66] Y. Wang, G. Zacharewicz, M.K. Traore, D. Chen, A tool for mining discrete event simulation model, 2017, http://dx.doi.org/10.1109/WSC.2017.8248027.
[67] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Progr. Artif. Intell. 5 (4) (2016) http://dx.doi.org/10.1007/s13748-016-0094-0.