

Article

An Education Process Mining Framework: Unveiling Meaningful Information for Understanding Students' Learning Behavior and Improving Teaching Quality

Hameed AlQaheri ^{1,*} and Mrutyunjaya Panda ²

¹ College of Business Administration, Kuwait University, P.O. Box 5486, Safat 13055, Kuwait

² Department of Computer Science and Applications, Utkal University, Odisha 751004, India; mrutyunjaya74@gmail.com

* Correspondence: hameed.alqaheri@ku.edu.kw

Abstract: This paper focuses on the study of automated process discovery using the Inductive visual Miner (IvM) and Directly Follows visual Miner (DFvM) algorithms to produce a valid process model for educational process mining in order to understand and predict the learning behavior of students. These models were evaluated on the publicly available xAPI (Experience API or Experience Application Programming Interface) dataset, which is an education dataset intended for tracking students' classroom activities, participation in online communities, and performance. Experimental results with several performance measures show the effectiveness of the developed process models in helping experts to better understand students' learning behavioral patterns.

Keywords: educational process mining (EPM); model discovery algorithms; inductive visual miner; directly follows visual miner; learning performance prediction

Citation: AlQaheri, H.; Panda, M. An Education Process Mining Framework: Unveiling Meaningful Information for Understanding Students' Learning Behavior and Improving Teaching Quality. *Information* **2022**, *13*, 29. <https://doi.org/10.3390/info13010029>

Academic Editors: Tianchong Wang and Beng Soo Ong

Received: 10 November 2021

Accepted: 7 January 2022

Published: 10 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Process mining, a subset of data mining, is a tool where students' collected activity logs can be used to discover non-trivial learning process information and create process models in terms of data flow diagrams, Petri nets, etc. Educational process mining (EPM) is a relatively new research technique used in educational data mining (EDM), in which the process plays a central role in the discovery, analysis, and visualization of students' learning compared to just obtaining interesting predictive results from large volumes of educational data [1–3]. The main objective of process mining techniques are to extract an unambiguous process model from event logs and then bridge the gap between traditional simulated model-based process analysis and data-oriented analysis techniques, such as machine learning and data mining. Furthermore, EPM uses end-to-end processes rather than local patterns to extract knowledge from event logs recorded using various ICT tools, such as online learning management systems (LMS) and MOOCs (massive open online courses) [4].

During the COVID-19 pandemic, online teaching and learning have become inevitable for students and teachers. This has created the possibility for students to learn from anywhere at any time, but the effectiveness of the online teaching and learning process has not been fully explored, with students facing struggles during online learning [5]. To improve productivity in online learning, traditional classroom learning along with students' learning styles and behaviors should be taken into consideration to achieve effective personalized learning [6].

Although many researchers have tried to explore educational data mining, very few studies have considered the entire process and its possible variants when analyzing students' learning processes. This has led to the evolution of educational process mining

(EPM), where process mining techniques are used in an online learning environment to discover learning insights from the extracted event log or audit trail data [7]. In EPM, online student learning behavior is recorded in such a way that an event is represented in an ordered manner with a process instance (or a case), an activity, a time stamp, and an originator for initiating the event activities.

There are significant differences between process mining and data mining based on the process-realized characteristics of event log data. Data mining uses the currently available data for analysis, whereas process mining looks at how these data were created and how they fit in a process.

Process discovery is somewhat different from the numerical computation of averages or sums over a set of values with slice, dice, and drilling operations, which are quite common in data warehouse applications. In contrast, process mining deals with semantically different time events, such as years or semesters, organizational entities, such as students or faculty members, dimensions of the process, such as male or female and group or individual assignments as sub-processes.

EPM is used to provide visual representations of the entire online educational process scenario in terms of process flow using Petri nets, causal nets, etc., along with the relationships among the class and attributes [8].

Many studies on EPM use alpha miners as the basic process discovery algorithm to address educational pitfalls. However, alpha miners do not take frequency into account in their discovery process; at the same time, they have severe limitations in dealing with noisy event logs that attract further developments [4,9,10]. In contrast, heuristic miners can deal with noisy and infrequent event logs, detect short loops, and allow the skipping of single activities, which makes it a popular approach in process mining applications [11]; however, it is still unable to make sound models. Following these pitfalls of process mining approaches, the motivations of this research are as outlined below.

Motivations

The following questions motivated the authors of this paper to carry out further research in this emerging area.

M1: What are the relevant methodologies adopted by researchers to address the issues of student learning analytics using process mining?

M2: What are the most common process mining tools and techniques employed in educational process mining for the automatic discovery of process models from collected event logs?

M3: How can EPM add insights to improve students' learning management systems and

M4: What are the challenges that need to be explored to make this emerging topic useful for the overall development of the academic system?

Research Questions:

This paper has the following research questions (RQs):

RQ 1: How the large volume of educational system data is exploited by instructors and administrators to understand students' learning habits, the factors influencing their learning habit and academic performance?

RQ 2: How automatic process mining algorithms can help to discover simplified educational process models and to extract more knowledge about the student learning behavioral properties?

Research Objectives (RO):

The objectives of this research are as follows:

RO 1: To compare and study the students learning behavior using correlation amongst several features available in the learning analytics dataset.

RO 2: To explore discovered process models with various modeling languages.

RO 3: To apply process mining using an inductive visual miner (IvM) and directly follows visual miner (DFvM) to discuss the learning behavior of the students.

The remainder of this paper is organized as follows. Section 2 discusses related work on process mining and learning analytics with their potential applications. Section 3 introduces the stepwise procedures followed in developing an educational process mining model, followed by the details of the dataset used for the experimental analysis in Section 4. Section 5 presents the proposed methodologies adopted in the implementation of the educational process, and Section 6 discusses the outcomes of the experimentation and analysis. Finally, the paper concludes with a scope for future research in Section 7.

2. Related Work

Process mining is considered to be an emerging technique used to bridge the gap between data science and process science. In this way, process models can be discovered, conformance checking can be performed, deviations or bottleneck situations can be analyzed, and improvements can be suggested [12]. Martin et al. [13] and Bogarín et al. [9] applied process mining in healthcare and education, respectively, to highlight the domain-specific usability from unstructured event log data.

Recently, event log data were used to create a domain-specific model with a supported theoretical framework to construct value-added service processes [13], along with an in-depth analysis of students' online learning behavior through massive open online courses (MOOCs) [14] and a behavioral analysis of the reasons behind the students dropping out [15].

Research has been conducted to predict students' academic performance using enrolled students' demographic data, data on their parents' education, their marks secured in their earlier classes, etc. In addition, decision tree methods [16] and clustering techniques [17] have been used to predict students' profiling and drop out behavior; however, the challenges of these traditional data mining techniques lie in dealing with sequences of events that contain dynamic behavior.

Process mining is considered to be able to bridge the gap between data science and process science, perform knowledge extraction from information systems' event logs, discover process models, perform conformance checking, understand bottlenecks, and suggest further improvements [4]. In 2004, Luan [18] proposed the use of process mining techniques in higher education institutions in order to better understand students' desires to join a particular course, provide them with proctorial assistance in completing their degree, obtain a placement, and determine how more alumni play a role in supporting their alma mater in both academic and financial pledges, to name a few.

Werner et al. [19] investigated the feasibility of embedding process mining with contemporary audits and demonstrated their effectiveness in terms of reliability and robustness in comparison with manual financial audit statements.

In [20], Hamdan et al. discussed how the COVID-19 pandemic has impacted students worldwide through the temporary shutdown of schools. Additionally, they discussed how online teaching has posed challenges both to teachers as well as to students regarding effective teaching–learning processes and has assisted students in obtaining a high-quality education. This has opened up opportunities for students to receive self-regulating learning [21] and enabled them to gain insights into effective learning strategies while learning a subject of interest [22].

Buij et al. [23] proposed a genetic algorithm-inspired flexible evolutionary tree model for process discovery that may improve educational processes, including curriculum design, software-assisted learning, professional training, and MOOC.

Kas et al. [24] demonstrated the use of the AutoML technique in a real-world case of process mining and advocated its application using sequential data in future work.

Azeta et al. [25] developed a process mining framework to study the virtual learning behavior of students in order to show the disparity between students who passed or failed a particular course using inductive and fuzzy miners with various fitness level values.

In [26], Omori et al. presented an in-depth analysis of several process tools and types of software available, along with their pros and cons and suitable applications for testing concept drift detection situations using an event log collected from a process control system-based industrial application.

K. Okoye et al. [27] experimented with using the LAEPI (learning analytics educational process innovation) model in an educational dataset to understand the learning activities of university students using an IvM (inductive visual miner) to understand the potential bottlenecks or deviations in students' final exam grades and current grades.

Kurniati [28] highlighted the significance of the process mining approach to healthcare using the electronic health records (HERs) of the patients and discussed the need to improve the data quality, along with the application of several other efficient process mining algorithms to investigate the effectiveness of the process mining approach in EHR systems.

Various applications of process mining have been suggested by the authors of [29] to understand students' learning behaviors based on student activity event logs and traces collected from an MOOC platform. These authors divided the students into separate groups in order to improve their analysis using a fuzzy miner process mining technique to visualize and understand the real behavior of students during learning.

The authors of [30] experimented with using a process mining layer based on the data set extracted from the Moodle LMS. The main objective was to identify learners at risk of dropping out and students with the potential to fail at an early stage, as well as to significantly improve learners' academic outcomes. The authors used process mining software with the PM4Py libraries for their experiments, along with event log clustering and process discovery.

3. Educational Process Mining

The concept of process mining techniques is based on the innovative work of Dutch scientist Wil van der Aalst, who used process discovery and advanced learning analytics. This concept most recently surfaced as a disparate area that ransacks electronic processes into fine granules to revamp both human and computerized components.

Process mining is an inquisitive technique that uses data from information systems to gain unprejudiced discernment and discover invisible problems. Its prime objective is to explore event log data in a consequential way in order to improve processes, provide intuition, endorse actions, discover bottlenecks, and take appropriate actions to mitigate them. In process mining, the sequential recoding of event logs is performed, where each event mentions an activity in a well-defined step in a process that is coupled to an explicit case of a process instance. There is also supplementary information, which includes the timestamp of the activity or event and resources (person or device) used for starting the event [31].

The evolution of the digital world and the emergence of the internet of things (IoT) have enabled us to record a huge volume of data with a high variability in order to discover events and extract knowledge. Furthermore, as real-time data analysis is now ubiquitous, the need to gain a deep understanding of business processes along with their impacts on social security has become of utmost importance. This means that the process mining approach leads to the provision of effective solutions in real-time big data environments by optimizing the data through event logs and recorded information.

There are three key process mining capabilities: (i) automated process discovery, where issues concerning the bottlenecks, acquiescence, and slackness of automated process models are detected; (ii) conformance checking, where the designed processes are compared with the actual ones to prioritize the issues and perform root cause analysis;

and (iii) process enhancement, which increases process automation through some optimization.

Business processes were difficult to understand during the pre-digital revolution age where data were gathered manually, causing the collection process to be very time consuming, limiting instinctive knowledge, and introducing human bias. As such, current process mining approaches help us to optimize existing processes.

While process mining warrants the analysis of back-end application event log files, process discovery based on the power of artificial intelligence and machine learning allows us to perceive and document millions of human–machine interactions and generate real-time inputs. As such, many industries, including Vodafone, KPMG, and Walmart, to name a few, are able to discover deviations and construct better and more productive processes after the use of process mining.

Figure 1 depicts a general educational process mining framework. Figure 1 shows the educational process, leverages the learning environment, and stores it in a database to enable the analysis and visualization of students' learning behavior by applying process mining algorithms. Investigating the learning environment in connection with the societal context enables us to describe the learning activities implemented and use information and communication technology (ICT) to assess the strategies adopted for effective and engaging learning along with student characteristics, motivating students to explore innovative ideas and engage in creative thinking to solve complex problems and discover knowledge rather than just memorizing information. The details of this teaching–learning process, as implemented in a rich learning environment, are stored in a database in order to enable teachers and administrators to gain further insight into the development of the education system being employed in their institution. The raw data recorded about the students' behavior are transformed into an XES (extensible event stream) event log that can be used by process mining tools and techniques for process model discovery, visualization, conformance, extension, and enhancement. The process model generated will present the results of the process mining activities and be visualized in the form of a Petri net, causal net, business process model and notation (BPMN), business process tree, or unified modeling language (UML) activity diagram. Finally, the instructor can analyze the results by detecting bottlenecks or deviations found in the student learning management process. Feedback is obtained in the academic realm in order to make further improvements.

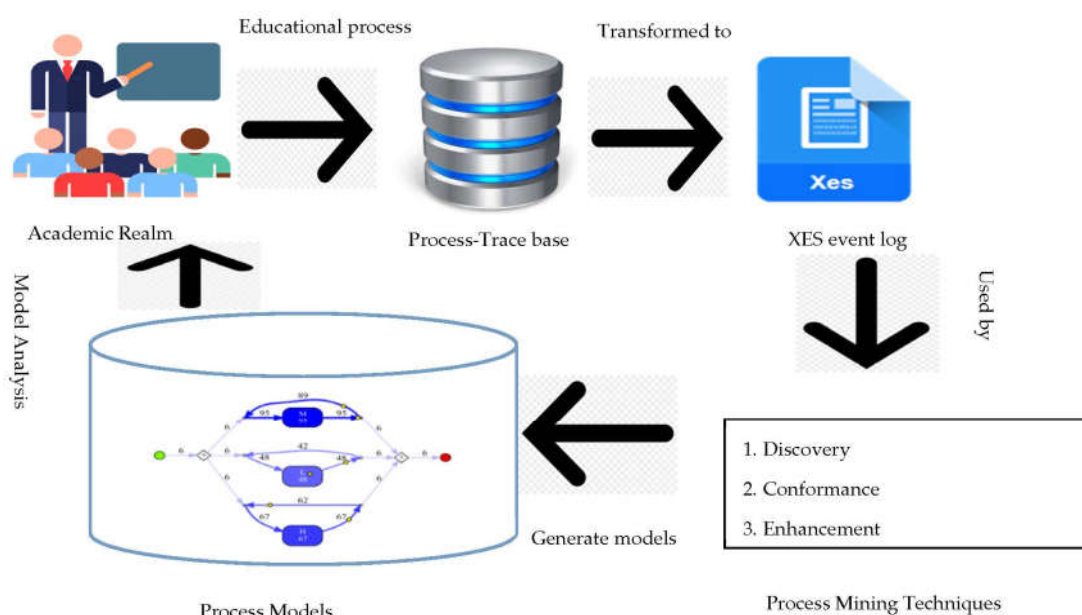


Figure 1. Education process mining framework.

4. Dataset Used

The quality of educational process mining is dependent on the data archive and the student behavioral features recorded. In this research, the xAPI data set [32] and an e-learning management system called Kalboard 360, which makes use of the Experience API Web service (xAPI) with avant-garde technology where students and parents have access to the resources through the internet, are used. These data were gathered through the use of xAPI, a learning activity tracker tool, as a part of the training and learning architecture [33] in order to monitor the students' learning progress and behaviors, such as raising their hands, participating in quizzes, reading articles, and utilizing online resources. This educational dataset enables institute administrators to effectively monitor students' academic activities, allowing parents to be involved in students' progress and recording their learning experiences.

The xAPI student learning analytics dataset consists of 480 student records with 16 features contained in each. These 16 features (nominal as well as discrete) can be compartmentalized into 3 major groups: (1) population tally in terms of gender and nationality; (2) scholarly background, such as education level, grade, and section; and (3) behavior—e.g., the frequency of students raising their hands during learning, the participation of parents in the survey, the use of e-learning resources by students, and parents' satisfaction with the school. The details are listed in Table 1.

Table 1. xAPI students learning analytics data set.

Feature Number	Feature Characteristics		Type
1	Gender	Male, Female	nominal
2	Nationality	Kuwait, Lebanon, Egypt, Saudi Arabia, USA, Jordan, Venezuela, Iran, Tunis, Morocco, Syria, Palestine, Iraq, Libya	nominal
3	Place of birth	Lower level, Middle School, High School	nominal
4	Educational stage	12 grades ranging from G-01 to G-12	nominal
5	Grade level	Section-A, B, C	nominal
6	Section ID	English, Spanish, French, Arabic, IT, Math, Chemistry, Biology, Science, History, Quran, Geology	nominal
7	Topic	First/Second	nominal
8	Semester	Mother/Father	nominal
9	Parent responsible for the student	(0, 1, 2, ... 100)	discrete
10	Number of times the student raises their hand during class	0, 1, 2, ... 100	discrete
11	Visited resources	0, 1, 2, ... 100	discrete
12	Viewing announcements: the number of times the student checks the new announcements	0, 1, 2, ... 100	discrete
13	Discussion groups: the number of times the student participates in a discussion	0, 1, 2, ... 100	discrete
14	Parents Answering Survey	Yes/No	nominal
15	Parent satisfaction with the school	Yes/No	nominal
16	Days the student is absent	Above 7/under 7	nominal

It is customary to say here that students' academic excellence largely depends on their learning behavior and their motivation to actively participate in learning processes, which represents their academic engagement. The successful implementation of a learning management system (LMS) can only be achieved when the students' overall characteristics are improved in terms of quality.

From the literature, it is observed that several aspects can effect students' learning performance, such as: (1) gender, where their aptitude for learning changes (e.g., male students are more inclined towards e-learning compared to female ones); (2) their parents' education level and their involvement in grooming their child (especially mother)—this also plays a major role in students' class attendance, and higher parental satisfaction will lead to the better academic performance of the student [34]. In [35], the authors use data mining techniques and conclude that parents' involvement in their child's education plays a vital role in students' performance. For example, men were found to be more responsible for their child's academic excellence in Kuwait, while women were found to be more responsible in Jordan.

Considering the above literature review, this research deals with the behavioral and interactional features of students' learning, along with their parents' involvement in the xAPI dataset, aiming to develop a process mining model. Box plot representations of the xAPI dataset are presented in Figures 2–4. These help us to understand the distribution and identify whether there are any outliers present in the dataset.

Box plot for student raised hand during learning

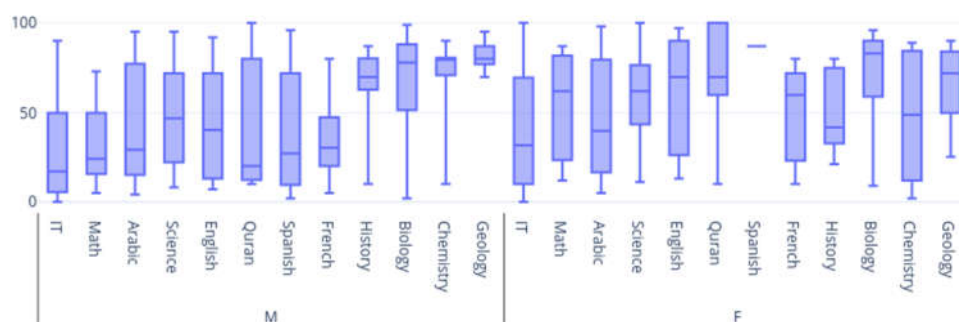


Figure 2. Box plots for students' interactional behavior (raised hands) by gender and topic studied.

Box plot for student Absent days during learning and parents school Satisfaction

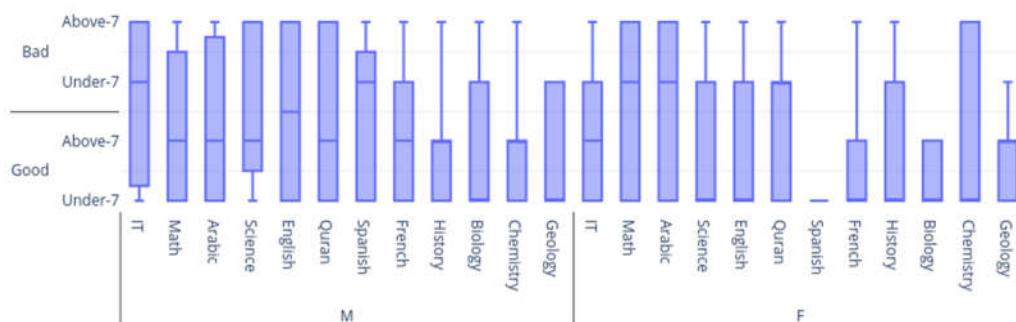


Figure 3. Students' absence from class based on their gender, the topic studied, and parental involvement.

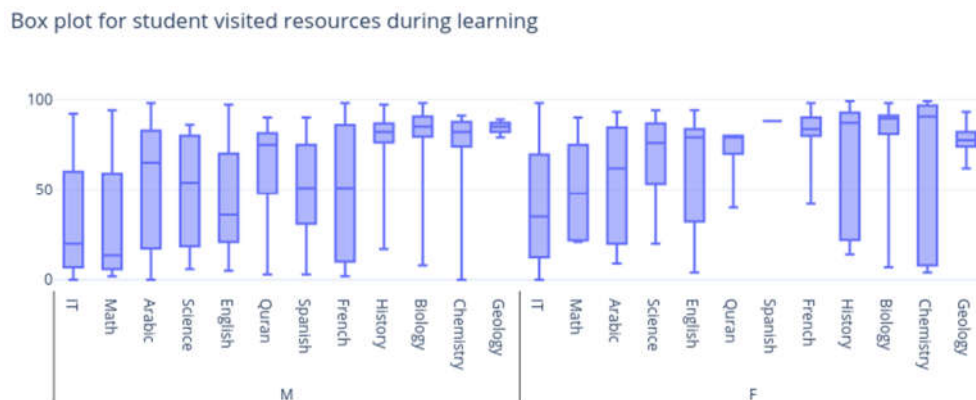


Figure 4. Exploratory data analysis for students' engagement in e-learning based on gender and the topic studied.

5. Implementation Approaches

The successful implementation of the process mining approach is based on gathering the correct event logs and selecting the most appropriate process mining methods. To this end, one can either opt for the direct application of process mining algorithms to the event log collected and analyze the results, which requires some forethought concerning the concepts, techniques, algorithms, and tools used or the specific platform-based implementation to be performed. This does almost everything for the user, starting from the creation of an event log to the process implementation.

Process mining techniques:

This section presents some of the popular process mining techniques available in the literature. In this study, we applied an Inductive visual Miner (IvM) and Directly Follows visual Miner (DFvM), an extension to IvM for process mining. A general process discovery scenario is one where event logs are used by the process discovery model, and the output is shown in terms of a Petri net.

Alpha Miner:

An alpha miner (or α -algorithm) [36] is one of the most widely used process mining techniques for restoring causality from a timeline of events. It is the first algorithm used in process mining applications and aims at bridging the gap between collected event log data and process model discovery, creating the model in terms of workflow nets (a subclass of the Petri net) without any supplementary knowledge. Some limitations concerning the original alpha miner have been reported, including: (i) dealing with noisy data, (ii) not being able to discover duplicate and hidden tasks, and (iii) dealing with loops with lengths of one or two. These limitations are addressed in subsequent versions of the algorithm: the α^+ (alpha plus) algorithm is capable of handling short loops, the α^{++} (alpha plus plus) algorithm can handle more complex patterns in the process, and the $\alpha^\#$ algorithm can discover hidden and unobserved tasks.

Inductive Visual Miner:

Inductive visual miners (IvMs) have not yet been explored by many researchers, especially in education datasets, in order to understand learning analytics [37]. Compared to alpha miners and heuristic miners, inductive miners can deal with large event logs and can ensure a soundness of build in process learning models [38]. Furthermore, they can handle unnoticed transitions by and large on the skipping and/or looping portion of a process model. Hence, the idea behind using a splitting operation (sequential, parallel, circumstantial, and prewired) in an event log is to obtain a protruding split (or sub-logs) so that the inductive miner algorithm recurs on the protruding split until the model's

expected case is recognized. The inductive miner, unlike the alpha miner, does not present a Petri net rather than a process tree; however, it is always possible to convert process trees to Petri nets. The strength of an inductive miner lies in its ability to discover robust process models that can efficiently deal with noisy and incomplete data. A general architecture using an IvM is shown in Figure 5.

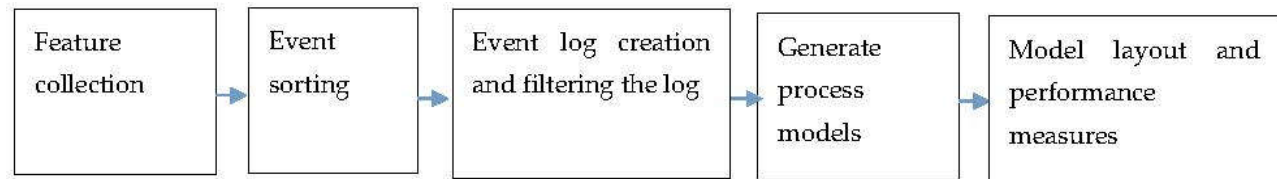


Figure 5. The architecture of an inductive visual miner.

In Figure 5, the arrows show the constructs where a task is followed after the completion of a previous task in a sequential manner. Initially, attributes or features of the recorded process data are collected, followed by the sorting of the event to create an event log. If any anomalies are found in the event log, they are filtered out until final clean event logs are obtained. Next, a process mining algorithm (IvM) is applied to the clean event log created in the previous step and models are built. Finally, one can visualize the model and obtain several performance measures, such as the precision, recall, and F1 score in order to gain insights into the process.

Heuristics Miner:

The heuristic miner (HM) makes several improvements over the alpha miner by taking frequency and significance into account. It is also capable of dealing with short loops and non-local dependencies along with unnoticed event logs; however, the problem lies in the fact that it cannot guarantee the soundness of the process model. The HM algorithm acts on the directly follows graph (DFG) and generates a heuristic net as an output, which later can be converted into a Petri net graph [39]. DFGs are visual representations in which events are denoted as nodes, with directed edges appearing between the nodes with frequency and performance calculations. The frequency in a DFG represents the number of times the source event follows the destination event, and through performance measures one can understand the time elapsed between the source and destination events during the process model discovery.

Evaluation metrics

The evaluation of educational processes and their intricate components is crucial in order to improve process discovery implementations. In our study, the performance of the discovered process model was evaluated in terms of its throughput time, fitness, simplicity, precision, and generalization [40].

- Throughput time denotes the time taken for a process to be executed completely from the start to the end
- Fitness (or recall) measures the aptness of the model to apprehend the recorded behavior in the collected event log. This quantifies how many of the observed behaviors in the event log fit well into the process model.
- Precision is used to enumerate the extent to which a process model overapproximates the behavior seen in an event log without allowing too many non-existent behaviors in the event logs.
- Simplicity in a process model provides a wool-gathering process and is induced by independent process instances. The model discovered should be as simple as possible, but too much simplicity will reduce its precision.
- Generalization indicates that the discovered model should generalize the example behavior, as seen in event logs that do not have any variations.

- Soundness deals with whether the process model is free from anomalies, such as deadlocks or livelocks. This provides an understanding as to whether a process terminates properly, provided that every activity is a participant in a process instance.
- F1 score is a performance measure of a process model's accuracy that is often defined as a harmonic mean of the precision and recall. Furthermore, it presents a score for the accuracy of the fitting or positive event traces and the accuracy of the non-fitting (or negative) ones. Therefore, if a process model classifies all traces as positive, the F1 score will be one (100%). On the other hand, if it classifies all traces as negative, the F1 score will be 0 (0%). In general, the value of the F1 score lies between 0% and 100%.
- Execution time or processing time is the time required to process an event log. A longer or shorter duration in process model building can have a negative or positive impact on the business activities, respectively.

The main challenge of process mining is that all these criteria are conflicting, which makes it very difficult to consider all of them concomitantly. Hence, we chose to use some of them in our proposed research.

6. Experimental Result and Discussion

All the experiments were performed on a laptop in a Windows environment using ProM 6.10 [41] with IvM and DFvM process mining techniques on an xAPI student learning analytics dataset to build a process mining model to understand the behavior of students in the learning process.

Initially, the event log collected from the xAPI learning analytics dataset was converted to an XES format, which is suitable for log analysis. Next, to remove any noise in the dataset, log filtering was performed using simple heuristics. The filtered event log is shown in Figure 5, where the events are reduced from 480 to 210 instances. The descriptive statistics, as the distribution of process instances for students' classes in the event logs and event classes with their transitions, after using a simple heuristic filter, are presented in Tables 2 and 3, respectively.

Table 2. Distribution of event classes in a log for students' classes.

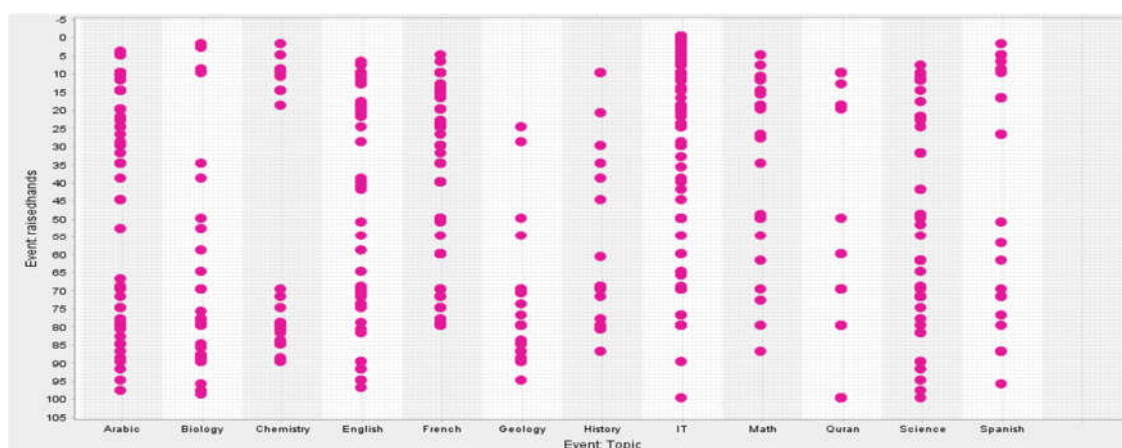
Total Number of Process Instances: 6; Total Number of Events: 210		
All Events		
Total number of classes: 3		
Class	Absolute occurrences	Relative occurrences
M	95	45.238%
H	67	31.905%
L	48	22.857%
Start events		
Total number of classes: 3		
Class	Absolute occurrences	Relative occurrences
H	2	33.333%
L	2	33.333%
M	2	33.333%
End Events		
Total number of classes: 2		
Class	Occurrences (absolute)	Occurrences (relative)
H	4	66.667%
M	2	33.333%

Table 3. Distribution of process instances for students' classes in the event logs.

Event Classes Defined by (Event Name AND Lifecycle Transition)		
All Events		
Total number of classes: 3 Class	Absolute occurrences	Relative occurrences
M + complete	95	45.238%
H + complete	67	31.905%
L + complete	48	22.857%
Start events		
Total number of classes: 3 Class	Absolute occurrences	Relative occurrences
H + complete	2	33.333%
L + complete	2	33.333%
M + complete	2	33.333%
End Events		
Total number of classes: 2 Class	Occurrences (absolute)	Occurrences (relative)
H + complete	4	66.667%
M + complete	2	33.333%

The dotted chart view of the event correlations between the topics studied by the students and their responses in the class (measured by them raising their hands) is shown in Figure 6. A dotted chart (with two orthogonal dimensions, such as time and component types) is a pictorial representation of how the process events are spread out over time, involving plotting a dot for each event in an event log in order to gain some knowledge about the process being created, interesting patterns in the process, and its performance. In Figure 6, topics studied by the students are presented on the horizontal axis of the chart, while one of the students' interactional behaviors, such as them raising their hands during learning a topic of interest, is shown on the vertical axis. Each row indicates a different task or learning event in the course, while the column represents the size of the dots, indicating how many students have raised their hands when learning about a particular topic.

It can be observed from Figure 6 that student participation in all topics was satisfactory; however, more interest is detected in Arabic, IT, English and Science subjects in particular. To provide a more detailed understanding of the students' behavior in all topics, a connected event graph is shown in Figure 7.

**Figure 6.** Dotted chart analysis showing the topic studied and hands raised.

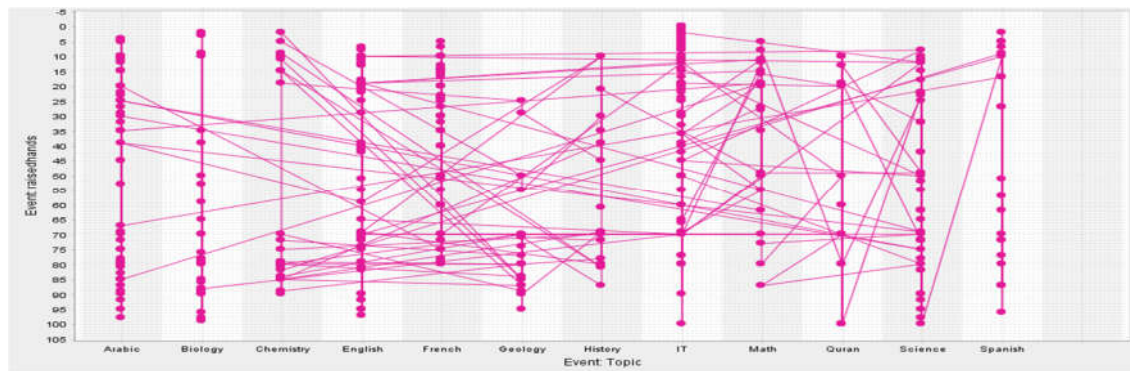


Figure 7. Dotted chart showing connected events between the topic studied and hands raised.

It is evident from Figure 8 that most of these students visited educational resources in order to learn better in topics such as English, IT, science, French, and Arabic. This shows the students' preferences regarding learning subjects that will improve their perspectives for obtaining jobs inside and outside their respective countries in more detail.

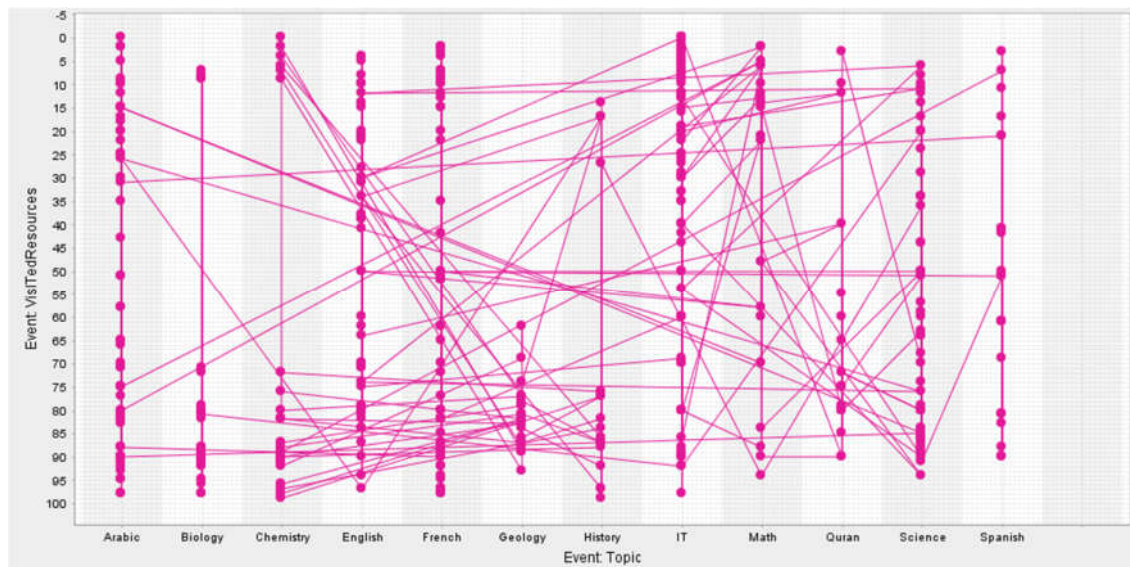


Figure 8. Dotted chart showing the connections between topics and the resources visited by students.

Figure 9 shows the relationship between class events and the topics taught to the students, where all the class events are shown in different shapes to provide a better visualization. The students attending the class for a particular topic are noted, and the connection between each event is shown in order to illustrate whether a student remains absent for one or several topics; this will enable teachers and administrators to take effective actions to motivate students to attend classes, as it will lead to a greater understanding of the reasons why a student might remain absent from a class or a particular topic. This is shown in Figure 10.

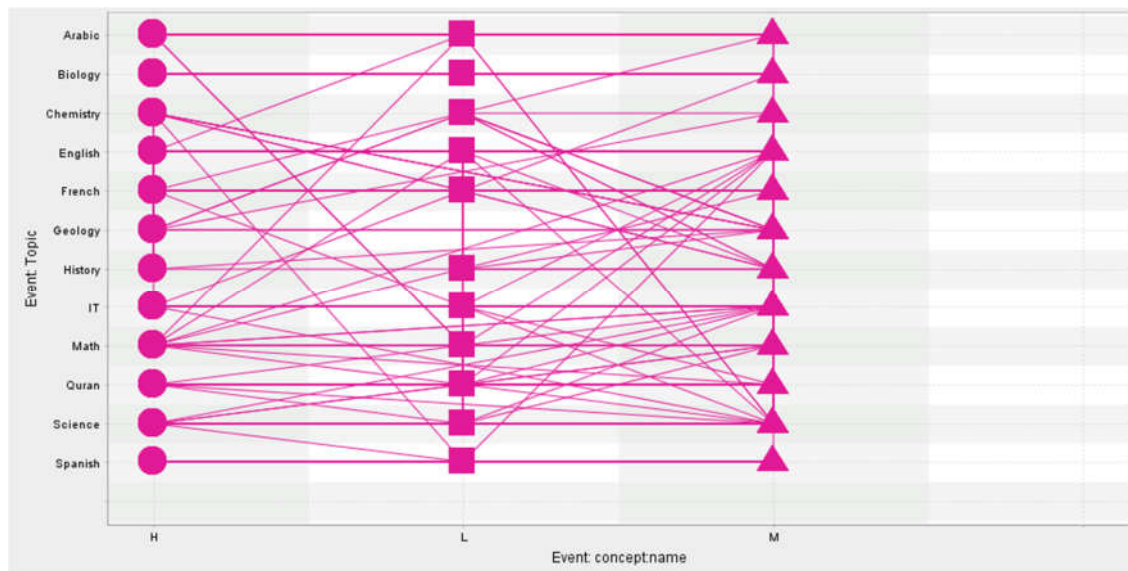


Figure 9. Frequency of distribution between class and topic in different shapes for events.

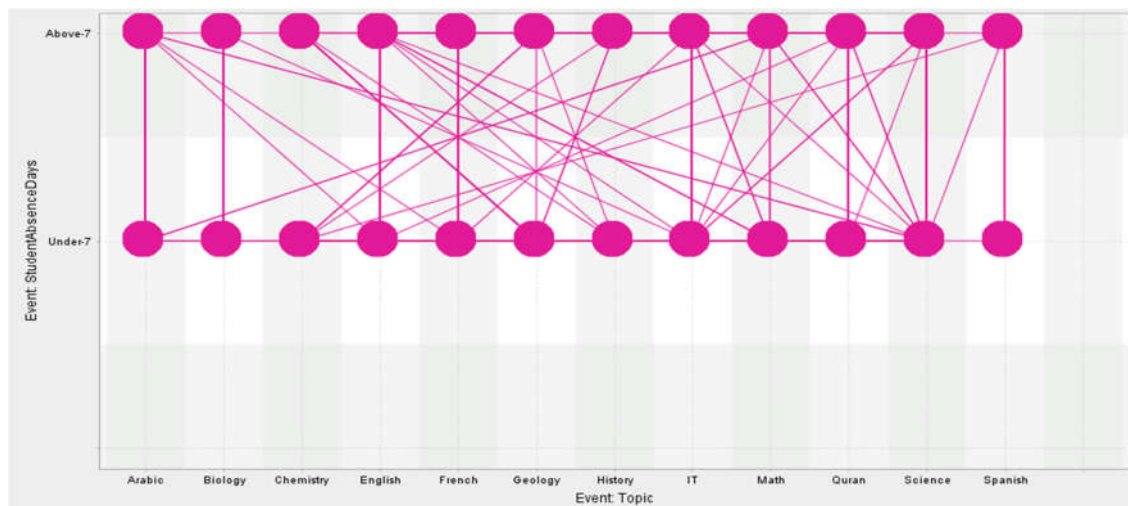


Figure 10. Visualization of the interconnection between a topic and the number of days the students were absent.

Experiment 1: Inductive Visual Miner (IvM)

Next, we develop the process models using inductive visual miner (IvM) to discover the learning process models and visualize the different paths the process instances follow in terms of class (L, M, and H) and student grades, as shown in Figure 11. In this study, we used the default IMf (Inductive Miner framework) to build the process model and determined the paths or points at which deviations or bottlenecks occurred during the process execution; these are marked in red in Figure 12. This red color with a mark of 1 in the model indicates that activity H has been skipped once in the event log, while the model said it should have been executed. To understand more about the IvM process model created, with its different classes, a relative path graph is shown in Figure 13.

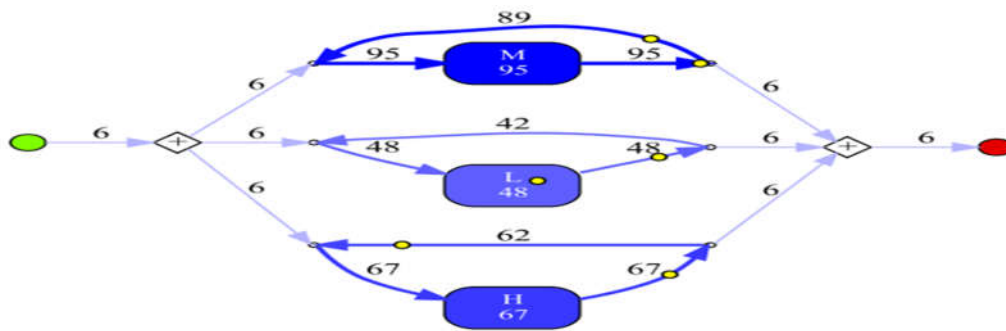


Figure 11. IvM using a default IMf for three classes (lower (L), middle (M), and higher (H)).

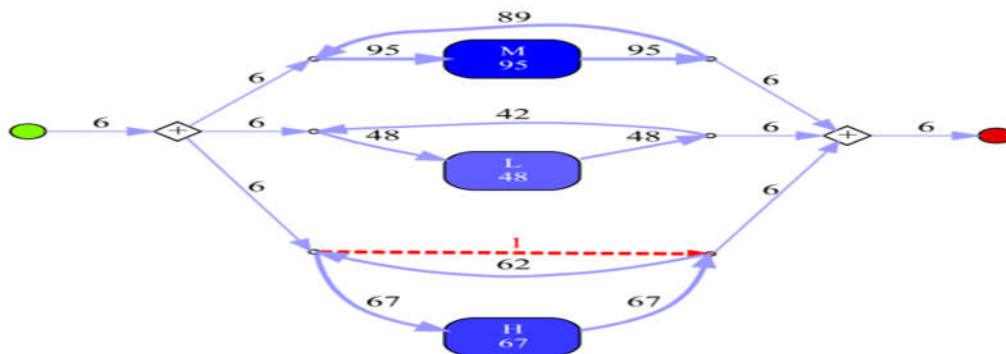


Figure 12. IvM process model with deviations or bottlenecks in Class H.

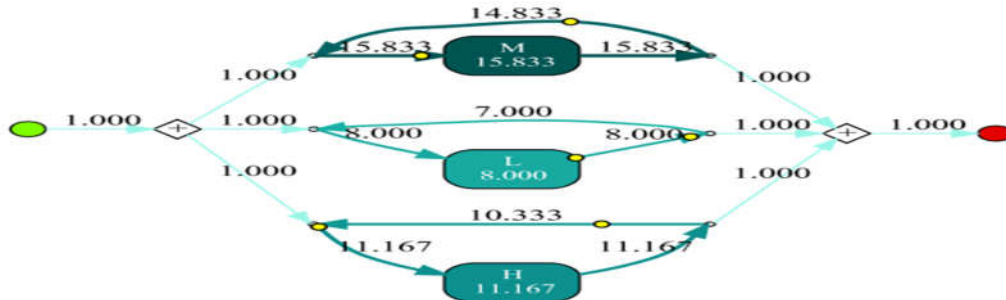


Figure 13. IvM process model with the relative paths shown for all student classes.

The prime objective of using IvM is to ensure the model's soundness so that the discovered model can present all possible event paths throughout the process life cycle. It can be seen from the graphs obtained from the IvM experiments that multiple case paths are reflected in the best possible way. From left to right in Figure 11, we start with six cases as learning lessons with the model branching in parallel. This is indicated by an icon mimicking a diamond shape with a plus sign inside, as well as by the measures for the outgoing branches (which are a diamond shape with a plus sign inside), with the number of cases unchanged at 6. This stipulates that in the overall process flow, any of these paths can be taken at any given point. Furthermore, it can be seen from the small green node at the left side of the figure that, in the beginning, the path is split into several directions. The one for Class M tells us that, in six cases, Class M is engaged, while, in 89 cases, we observe that this process flows back on itself, leaving a total engagement frequency of 95. The flow loopback of 89 indicates that there were transitions in the same micro-level process within a single learning period. This micro-level process frequency depicts students' learning behavior.

For a more detailed analysis, we extended the case ID and event IDs to develop a more complex model, as shown in Figure 14, using concept names. Here, log deviations are shown as loops that should have been executed by the model during the process discovery.

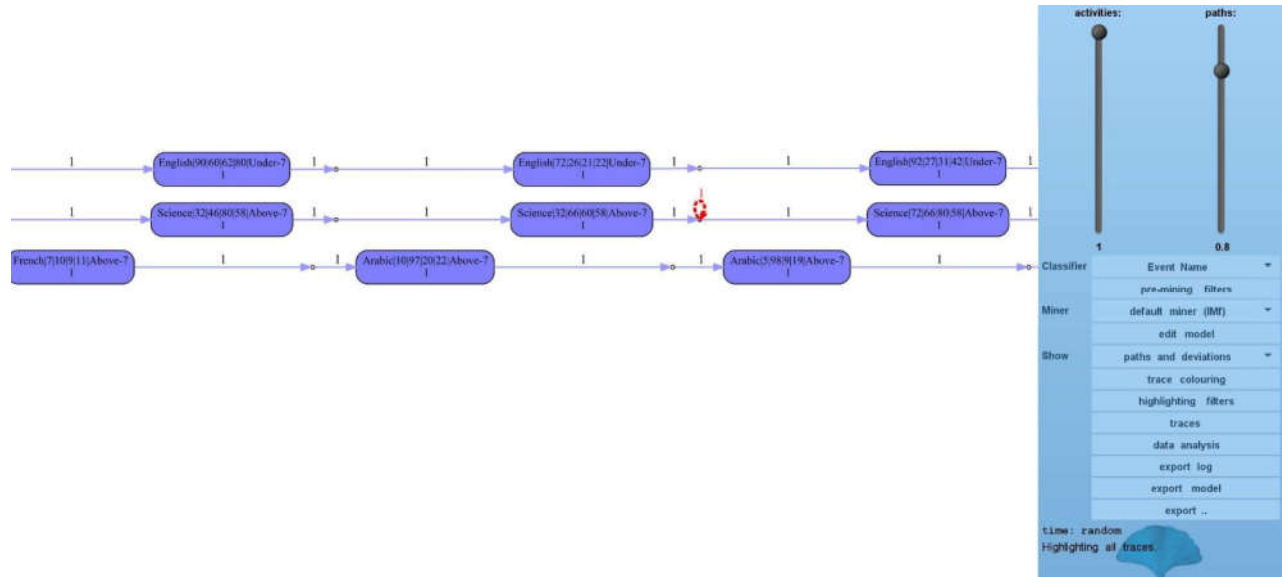


Figure 14. IvM model with log deviations.

Experiment 2: Directly Follows visual Miner (DFvM)

Secondly, we used a relatively new miner, a directly follows visual miner (DFvM) [42], which is an extension to IvM that has several new characteristics compared to IvM. DFvM performs process discovery automatically and iteratively. It can choose to use filter logs based on the event and trace attributes before and after the model discovery, the application of classifiers, conformance checking, and model assessment. Here, the green and red circles indicate the start and end of the process, respectively. Figure 15 shows the DFvM model discovery for three class activities, H, L, and M, while Figure 16 shows the possible bottlenecks. As can be seen, there are no red marks in Figure 16; hence, DFvM was not able to trace any path deviations in the discovered process models, contrary to IvM, which detected one path deviation. This shows the soundness of the DFvM model in comparison to the IvM model. Finally, the relative path scenario for the DFvM model is shown in Figure 17.

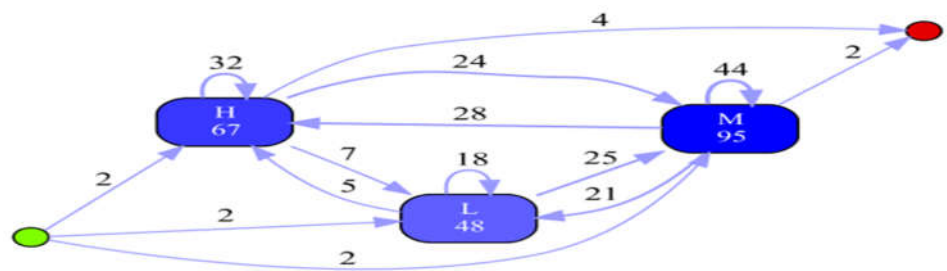


Figure 15. DFvM model discovery showing paths.

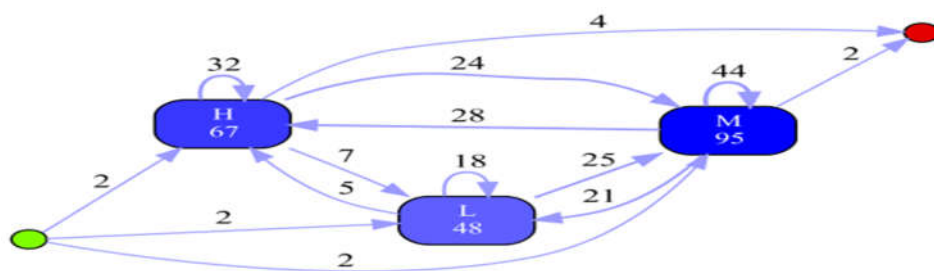


Figure 16. DFvM model showing paths and deviations (no deviations found).

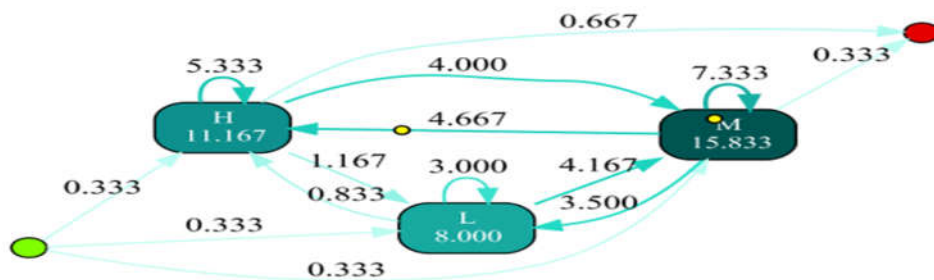


Figure 17. Relative paths for the DFvM model discovery.

Evaluation Metrics:

Table 4 shows the results of the four algorithms in the evaluation metric in terms of fitness, precision, F1 score, and soundness.

Table 4. Comparison of the process mining algorithms.

Event Logs	Algorithms	Fitness or Recall	Precision	F1 Score	Execution Time in Sec	Soundness
X-API learning analytics dataset	IvM	0.986	1	0.993	random	yes
	DFvM	1	1	1	random	yes

Table 4 shows that even though both IvM and DFvM are sound models, as they do not have any deadlocks, DFvM outperforms IvM with a 100% accuracy in terms of precision, recall, and F1 score. The low recall score of 0.986 for the IvM-based process model indicates that some of the observed behaviors in the event log do not fit well (98.6% fit) with the process model in comparison to DFvM, in which all the observed behaviors fit well (100% fit) in the model. The algorithm uses random times for different attributes while building the process model.

Threats to Validity:

The first threat to validity lies in the potential selection prejudice and imprecision in the data extraction, selection, and analysis, which is quintessential of existing literature. To deal with such issues, we chose to use appropriate literature pertaining to automatic process discovery. Next, the performance evaluation of process mining algorithms was carried out in this study through several experiments and was limited to Petri nets and causal nets (C-nets). Finally, the generalization of the experiments was limited to student learning pertaining to the event logs in the xAPI e-learning dataset.

7. Conclusions

This paper presents a methodical literature analysis of automated process discovery methods and analyzes their advantages and disadvantages in process mining applications. The x-API education learning dataset used for the experiments shows the students'

behavior in the learning process. This paper highlighted the log interpretation in terms of absolute and relative metrics in order to provide a better understanding of self-regulated learning. It was observed from the experiments that learning model discovery using process mining techniques could be useful for preventing dropouts or burdens in learning management systems. From the experimental results, it is evident that DFvM outperformed IvM, presenting a more accurate automatic process discovery model in terms of soundness, precision, recall, and F1 score. Finally, as the online academic learning context has become more important as a consequence of the COVID-19 pandemic, more research in this area is warranted in order to better understand the social networks involved and enable educational institutions to invest in the right resources to address the problems causing students to drop out of classes. Hence, to generalize the performance of process mining algorithms in improving quality education, other benchmarking datasets obtained from alternate LMS or MOOCs should be explored, while more universal measures of performance in process mining should be identified in the future.

Choosing the right process discovery tool, such as an inductive miner (IvM) and its variant (DFvM), is an important matter that should be considered, while further improvements should be made in developing interactive heuristic miners with conformance checking in future research. The implications of the use of these methods on an unsampled and larger dataset should also be explored in future studies.

Author Contributions: Conceptualization, H.A. and M.P.; Data curation, H.A. and M.P.; Formal analysis, H.A. and M.P.; Investigation, H.A. and M.P.; Methodology, M.P.; Project administration, H.A.; Resources, H.A. and M.P.; Software, M.P.; Supervision, M.P.; Validation, H.A. and M.P.; Visualization, H.A. and M.P.; Writing—original draft, H.A. and M.P.; Writing—review & editing, H.A. and M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The used data can be accessed at <https://www.kaggle.com/aljarah/xAPI-Edu-Data> (accessed on 20 November 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dutt, A.; Ismail, M.A.; Herawan, T. A Systematic Review on Educational Data Mining. *IEEE Access* **2017**, *5*, 15991–16005. <https://doi.org/10.1109/access.2017.2654247>.
2. Romero, C.; Ventura, S. Educational data mining: A survey from 1995 to 2005. *Expert Syst. Appl.* **2007**, *33*, 135–146.
3. Reimann, P.; Markauskaite, L.; Bannert, M. e-Research and learning theory. *Br. J. Educ. Technol.* **2014**, *45*, 528–540. <https://doi.org/10.1111/bjet.12146>.
4. van der Aalst, W. Data Science in Action. In *Process Mining*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 3–23. https://doi.org/10.1007/978-3-662-49851-4_1.
5. Pemberton, A.; Moallem, M. The impact of personalized learning on motivation in online learning. In *Society for Information Technology & Teacher Education International Conference*; Association for the Advancement of Computing in Education (AACE): Waynesville, NC, USA, 2013; pp. 907–914.
6. Felder, R.M.; Silverman, L.K. Learning and teaching styles in engineering education. *Eng. Educ.* **1988**, *78*, 674–681.
7. Bogarn, A.; Romero, C.; Cerezo, R.; Sanchez-Santillan, M. Clustering for improving educational process mining. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, Indianapolis, IN, USA, 24–28 March 2014; pp. 11–15.
8. Reisig, W.; Rozenberg, G. (Eds.) *Lectures on Petri Nets I: Basic Models*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 1998; Volume 1491. ISBN 978-3-540-65306-6.
9. Bogarín, A.; Cerezo, R.; Romero, C. A survey on educational process mining. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2017**, *8*, e1230. <https://doi.org/10.1002/widm.1230>.
10. Van der Aalst, W.M.P. *Process Mining: Discovery, Conformance and Enhancement of Business Process*; Springer: Berlin/Heidelberg, Germany, 2011. ISBN 978-3-642-19345-3.

11. Porouhan, P.; Jongsawat, N.; Premchaiswadi, W. Process and deviation exploration through Alpha-algorithm and Heuristic miner techniques. In Proceedings of the 2014 Twelfth International Conference on ICT and Knowledge Engineering, Bangkok, Thailand, 18–21 November 2014; pp. 83–89.
12. Zhou, X.; Zacharewicz, G.; Chen, D.; Chu, D. A Method for Building Service Process Value Model Based on Process Mining. *Appl. Sci.* **2020**, *10*, 7311.
13. Martin, N.; De Weerd, J.; Fernández-Llatas, C.; Gal, A.; Gatta, R.; Ibáñez, G.; Johnson, O.; Mannhardt, F.; Marco-Ruiz, L.; Mertens, S.; et al. Recommendations for enhancing the usability and understandability of process mining in healthcare. *Artif. Intell. Med.* **2020**, *109*, 101962. <https://doi.org/10.1016/j.artmed.2020.101962>.
14. Maldonado-Mahauad, J.; Pérez-Sanagustín, M.; Kizilcec, R.F.; Morales, N.; Muñoz-Gama, J. Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses. *Comput. Hum. Behav.* **2018**, *80*, 179–196.
15. Salazar-Fernandez, J.P.; Sepulveda, M.; Muñoz-Gama, J. Influence of student diversity on educational trajectories in engineering high-failure rate courses that lead to late dropout. In Proceedings of the 2019 IEEE Global Engineering Education Conference (EDUCON), Dubai, United Arab Emirates, 8–11 April 2019; pp. 607–616.
16. Kabra, R.R.; Bichkar, R. Performance Prediction of Engineering Students using Decision Trees. *Int. J. Comput. Appl.* **2011**, *36*, 8–12.
17. Iam-On, N.; Boongoen, T. Generating descriptive model for student dropout: A review of clustering approach. *Human-Cent. Comput. Inf. Sci.* **2017**, *7*, 1. <https://doi.org/10.1186/s13673-016-0083-0>.
18. Jing, L. *Data Mining Applications in Higher Education*; SPSS Executive Report; 2004; Volume 7. pp. 1–20. Available online: <http://www.insol.it/software/modeling/modeler/pdf/Data%20mining%20applications%20in%20higher%20education.pdf> (accessed on 6 January 2022).
19. Werner, M.; Wiese, M.; Maas, A. Embedding process mining into financial statement audits. *Int. J. Account. Inf. Syst.* **2021**, *41*, 100514. <https://doi.org/10.1016/j.accinf.2021.100514>.
20. Hamdan, K.M.; Al-Bashaireh, A.M.; Zahran, Z.; Al-Daghestani, A.; Al-Habashneh, S.; Shaheen, A.M. University students' interaction, Internet self-efficacy, self-regulation and satisfaction with online education during pandemic crises of COVID-19 (SARS-CoV-2). *Int. J. Educ. Manag.* **2021**, *35*, 713–725. <https://doi.org/10.1108/ijem-11-2020-0513>.
21. Viberg, O.; Khalil, M.; Baars, M. Self-regulated learning and learning analytics in online learning environments. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, Frankfurt, Germany, 23–27 March 2020; ACM: New York, NY, USA, 2020; pp. 524–533.
22. Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *WIREs Data Min. Knowl. Discov.* **2020**, *10*, e1355.
23. Buijs, J.; Dongen, B.; Aalst, W. On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery. In *On the Move to Meaningful Internet Systems: OTM 2012. OTM 2012. Lecture Notes in Computer Science*; Meersman, R., Ed.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7565. https://doi.org/10.1007/978-3-642-33606-5_19.
24. Kas, S.; Post, R.; Wiewel, S. Automated Machine Learning in a Process Mining Context. 2020. Available online: https://icpm-conference.org/2020/wp-content/uploads/sites/4/2020/10/ICPM_2020_paper_44.pdf (accessed on 6 January 2022).
25. Azeta, A.; Agono, F.; Falade, A.; Azeta, E.; Nwaocha, V. A Digital Twin Framework for Analysing Students' Behaviours Using Educational Process Mining. 2020. pp. 1–19. Available online: <https://www.researchsquare.com/article/rs-51184/v1> (accessed on 6 January 2022).
26. Omori, N.J.; Tavares, G.M.; Ceravolo, P., Jr.; Barbon, S. Comparing Concept Drift Detection with Process Mining Software. *iSys Revista Brasileira de Sistemas de Informac, ˜ao (Braz. J. Inf. Syst.)* **2020**, *13*, 101–125. <https://doi.org/10.5753/isys.2020.832>.
27. Okoye, K.; Nganjji, J.T.; Hosseini, S. Learning analytics: The role of information technology for educational process innovation. In Proceedings of the International Conference on Bioinspired Computing and Applications (IBICA), Gunupur, India, 16–18 December 2019; AISC 1180; Springer: Cham, Switzerland, 2020; pp. 272–284.
28. Kurniati, A.P.; Rojas, E.; Hogg, D.; Hall, G.; Johnson, O.A. The assessment of data quality issues for process mining in healthcare using Medical Information Mart for Intensive Care III, a freely available e-health record database. *Health Inform. J.* **2019**, *25*, 1878–1893.
29. Mukala, P.; Buijs, J.; Leemans, M.; van der Aalst, W. Exploring Students' Learning Behaviour in MOOCs Using Process Mining Techniques. Computing Conference. 2015; pp. 1–12. Available online: <http://bpmcenter.org/wp-content/uploads/reports/2015/BPM-15-10.pdf> (accessed on 6 January 2022).
30. Hachicha, W.; Ghorbel, L.; Champagnat, R.; Zayani, C.A.; Amous, I. Using Process Mining for Learning Resource Recommendation: A Moodle Case Study. *Procedia Comput. Sci.* **2021**, *192*, 853–862. <https://doi.org/10.1016/j.procs.2021.08.088>.
31. van der Aalst, W. Process mining: Overview and opportunities. *ACM Trans. Manag. Inf. Syst.* **2012**, *3*, 1–17. <https://doi.org/10.1145/2229156.2229157>.
32. Abu Amrieh, E.; Hamtini, T.; Aljarah, I. Preprocessing and analyzing educational data set using X-API for improving student's performance. In Proceedings of the 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, Jordan, 3–5 November 2015; pp. 1–5. <https://doi.org/10.1109/AEECT.2015.7360581>.
33. Madnaik, S.S. Predicting Students' Performance by Learning Analytics. Master's Projects. 2020. Available online: https://scholarworks.sjsu.edu/etd_projects/941 (accessed on 6 January 2022).
34. Abu Amrieh, E.; Hamtini, T.; Aljarah, I. Mining educational data to predict student's academic performance using ensemble methods. *Int. J. Database Theory Appl.* **2016**, *9*, 119–136.

35. Bharara, S.; Sabitha, A.S.; Bansal, A. Application of learning analytics using clustering data Mining for Students' disposition analysis. *Educ. Inf. Technol.* **2018**, *23*, 957–984. <https://doi.org/10.1007/s10639-017-9645-7>.
36. van der Aalst, W.; Weijters, A.T.; Maruster, L.L. Workflow mining: Discovering process models from event logs. *IEEE Trans. Knowl. Data Eng.* **2004**, *16*, 1128–1142. <https://doi.org/10.1109/tkde.2004.47>.
37. Bogarín, A.; Cerezo, R.; Romero, C. Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs). *Psicothema* **2018**, *30*, 322–329. <https://doi.org/10.7334/PSICOTHEMA2018.116>.
38. Leemans, S.J.J.; Fahland, D.; van der Aalst, W.M.P. Discovering Block-Structured Process Models from Event Logs Containing Infrequent Behaviour. In *Business Process Management Workshops. BPM 2013. Lecture Notes in Business Information Processing*; Lohmann, N., Song, M., Wohed, P., Eds.; Springer: Cham, Switzerland, 2014; Volume 171. https://doi.org/10.1007/978-3-319-06257-0_6.
39. Nuritha, I.; Mahendrawathi, E. Structural Similarity Measurement of Business Process Model to Compare Heuristic and Inductive Miner Algorithms Performance in Dealing with Noise. *Procedia Comput. Sci.* **2017**, *124*, 255–263.
40. Naderifar, V.; Sahran, S.; Shukur, Z. A Review on Conformance Checking Technique for the Evaluation of Process Mining Algorithm. *TEM J.* **2019**, *8*, 1232–1241, ISSN 2217–8309. <https://doi.org/10.18421/TEM84-18>.
41. Dixit, P.M.; Verbeek, H.M.W.; Buijs, J.C.A.M.; van der Aalst, W.M.P. Interactive data-driven process model construction. In *Conceptual Modeling—37th International Conference, ER 2018, Proceedings, Xi'an, China, 22–25 October 2018*; Du, X., Li, G., Li, Z., Trujillo, J.C., Ling, T.W., Davis, K.C., Lee, M.L., Eds.; Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2018; pp. 251–265.
42. Leemans, S.J.; Poppe, E.; Wynn, M.T. Directly Follows-Based Process Mining: Exploration & a Case Study. In *Proceedings of the 2019 International Conference on Process Mining (ICPM), Aachen, Germany, 24–26 June 2019*; pp. 25–32. <https://doi.org/10.1109/ICPM.2019.00015>.