ASTES

*Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 2, 1084-1090 (2021)*

www.astesj.com

*Special Issue on Multidisciplinary Sciences and Engineering*

**ASTESJ**

**ISSN: 2415-6698**

# Supporting the Management of Predictive Analytics Projects in a Decision-Making Center using Process Mining

Marlene Ofelia Sanchez-Escobar*, Julieta Noguez, Jose Martin Molina-Espinosa, Rafael Lozano-Espinosa

*School of Engineering and Science, Tecnologico de Monterrey, Mexico City, 14380, Mexico*

ARTICLE INFO

*Article history:*
*Received: 11 January, 2021*
*Accepted: 12 April, 2021*
*Online: 28 April, 2021*

*Keywords:*
*Process design*
*Process modeling*
*Project Management*
*Data Mining*
*Best Practices*
*Decision making*

ABSTRACT

*A Decision-Making Centers (DMCs) Environment facilitates stakeholders' decision-making processes using predictive models and diverse what-if scenarios. An essential element of this environment is the management of Decision Support Components (e.g., models or systems) that need to be created with mature methodologies and good delivery time. However, there has been a gap in the understanding of project management best practices in DMC environments and in the application of methodologies to ease project execution. In the following paper, we address that gap by analyzing six predictive analytics projects executed in a Mexican DMC using Process Mining techniques. We perform process discovery using a detailed activity event log, which has not been possible in previous studies. Additionally, we perform a compliance evaluation versus the de facto methodology to identify the current process alignment gaps, and finally, we analyze the social networks present in the process execution. The research reveals that (1) process mining models are helpful to address management issues of PA/DM projects (2) PA/DM projects require alignment to mature methodologies to improve process performance and avoid execution problems (3) PA/DM project execution should be revised at the activity level to identify issues and to propose specific strategies. This study's findings can help project managers to perform process analyses and to make informed decisions in PA/DM projects. The following paper is an extension of the article "Applying Process Mining to Support Management of Predictive Analytics/Data Mining Projects in a Decision-Making Center¨ presented in the 2019 International Conference on Systems and Informatics (ICSAI 2019).*

## 1. Introduction

Decision-Making Centers (DMCs) are immersive virtual environments used to understand complex problems, simplify decision-making, and visualize the results of predictive and scenario-based models [1]. These environments depend on the creation process of tools like: Predictive Analytics/Data Mining (PA/DM) models to operate [2]. Nevertheless, the authors have demonstrated in previous studies that DMC processes focus on high-level tasks and exclude detailed and standard PA/DM activities [2]. The absence of commonality in PA/DM project execution, generates issues, since (1) models are built using empiric methodologies and (2) managers cannot follow up specific technical activities since they are different in every project.

In this research, we propose three approaches to overcome the mentioned issues and help managers and modelers make informed decisions about PA/DM projects. In the first, we apply process mining techniques to a set of PA/DM processes to discover the timing, flow, frequency, and performance of activities from diverse perspectives (e.g., process, organizational, and case). Second, we compare a real PA/DM project execution with an accepted PA/DM methodology, to identify how aligned are the real processes to the formal methodology (i.e., CRISP-DM) and what gaps need to be closed to achieve compliance. Third, we perform complementary human resources analyses to visualize the relationship between resources and communication channels during process execution.

We expect that managers in DMCs use the models presented in this study to evaluate their processes and to consider the implementation of specific management strategies.

*Corresponding author: Marlene Ofelia Sanchez-Escobar, Tel.: +521 5544995733, Email address: A00704709@itesm.mx

https://dx.doi.org/10.25046/aj0602123

Finally, the organization of the paper is as follows: Section 2 and 3 presents the background and the literature review respectively. Section 4 describes the experimental design. Section 5 provides final results and discussion, and section 6 describes conclusions and future work.

## 2. Background

### *2.1. Process Mining Techniques and Project Management applications*

The Process Mining (PM) technique is a reverse engineering approach where process models are generated using event logs [2]. In [3], the author classifies the following PM techniques that we use for our analysis: discover, conformance, and enhancement.

The process discovery technique aims to mine process models using discovery algorithms, so the process helps managers answer specific questions [4]. Examples of discovery algorithms include alpha algorithm, heuristic miner, fuzzy miner, genetic miner, region miner, and integer linear programming (ILP). Differently, the process conformance technique aims to measure the process quality through metrics like: fitness, precision, generalization, and simplicity [4]. In this category, conformance checking is used to compare the expected model and the reality obtained from event logs. Likewise, it is possible to identify processes, commonalities, similarities, and deviations [3]. Finally, the process enhancement technique aims to extend the model with relevant information [4]. For instance, statistical metrics based on timestamps (e.g., throughput time, working time, and waiting time) or the use of replay analysis to visualize process execution. In this research, we use Disco and ProM 6 tools to implement the process mining techniques previously explained.

Finally, in [5], the authors explain that the project management field requires the process mining discipline to identify optimal workflows within project life cycles. In this regard, we consider that the following managerial issues identified in PA/DM projects can be analyzed using process mining techniques: establishing realistic goals, the creation of good teams, gaining knowledge of data, lack of infrastructure, poor project communication methodology, lack of risks management and change management [6], [7].

### *2.2. CRISP-DM Framework*

The Cross-Industry Standard Process for Data Mining (i.e., CRISP-DM) is the most accepted methodology in the field for executing data mining projects [8]-[10]. In the framework, the project life cycle includes the following key phases [11]: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. The methodology provides a universal process with generic tasks that can be executed for all data mining projects. The CRISP-DM reference model can be regarded in Figure 1, and the generic tasks to be performed in each phase are listed in Table 1. As can be noticed, the generic tasks are defined at the activity level, which facilitates its integration to high-level DMC process. In this study, we examine the data understanding, data preparation, modeling, and evaluation phases of CRISP-DM methodology, given information constraints.

### *2.3. PA/DM processes with CRISP-DM*

We discussed in [2] the importance of integrating PA/DM processes with DMC processes, and we make an integration effort. However, PA/DM processes at the level of activity have not been studied separately. In this study, we focus just on PA/DM processes, as a part of DMC processes, since there is no work in the literature that performs such analysis using process mining techniques. Finally, we assume that the CRISP-DM methodology matches DMC's PA/DM processes.

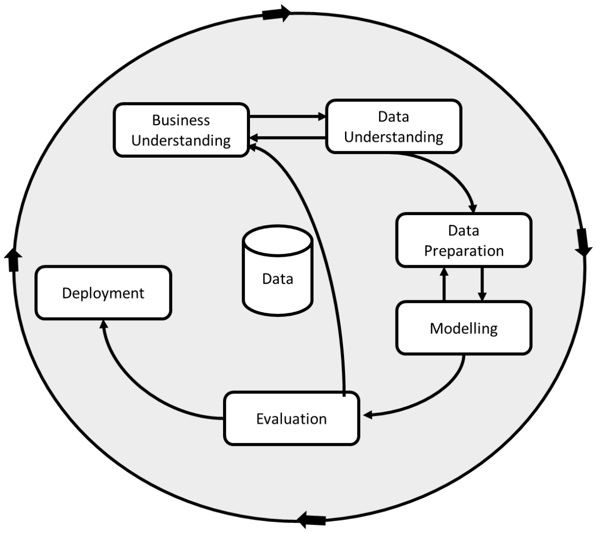

Figure 1: CRISP-DM phases and activity flow. Adapted from (Chapman, 2000)

Table 1: CRISP-DM phases and generic tasks

| Phase | Generic Task |
|---|---|
| Business understanding | Determine business objectives<br>Assess situation<br>Determine Data Mining goals<br>Produce Project Plan |
| Data Understanding | Collect Initial Data<br>Describe Data<br>Explore Data<br>Verify Data Quality |
| Data Preparation | Select Data<br>Clean Data<br>Construct Data<br>Integrate Data<br>Format Data |
| Modeling | Select Modeling Technique<br>Generate Test Design<br>Build Model<br>Assess Model |
| Evaluation | Evaluate Results<br>Review Process<br>Determine Next Steps |
| Deployment | Plan Deployment<br>Plan Monitoring and Maintenance<br>Produce Final Report<br>Review Project |

## 3. Literature review

Limited research has focused on Project Management using Process Mining techniques. In the literature, studies concentrate on the analysis of project management processes using data mining [12]-[14] and predictive analytics [15]-[17] techniques, but not process mining techniques. Likewise, we identify that most analyses are focused on software life cycles [12]-[18], but not on DA/PM processes. For instance, in [18], the authors use conformance checking techniques to reveal aspects of processes and identify deviations in software project execution. Additionally, the author presented an application to optimize the software development life cycle of projects using process mining [5]. However, in both cases, research is not focused on PA/DM projects or DMC environments.

Finally, the authors developed a previous study to analyze PA/DM processes in a DMC [2]. Nevertheless, the paper is oriented to analyze project phases and not project activities. Thus, that limits the possibility to discover low-level issues and applying target strategies. Besides, the study uses only in its majority enhancement techniques and a limited number of discovery algorithms. In the present study, we address those limitations.

## 4. Experimental Design

The DMC located at Tecnologico de Monterrey, Mexico City campus also referred as "Decision Laboratory", is a room with seven big format and high-definition screens that offers a space to make consensual decisions and to present solution proposals to a group of decision-makers [19]. This last with the goal of selecting the best possible solution. A picture of Tecnologico de Monterrey's DMC can be regarded in Figure 2.

Figure 2: Tecnologico de Monterrey's DMC

In the Mexican DMC, managers execute PA/DM projects to create models that support decision-making. The team organization is defined according to the knowledge of resources and their affinity to develop specific models. The project manager role is performed by one resource, and one or more product owners define the business requirements. After project execution, the modelers report that no formal methodology is applied to create the models.

Even though project managers try to deliver models with quality and on time, modelers and the supervisor report the following issues during project execution: (1) Wrong selection of modeling technique and (2) lack of standardization of data glossary among models.

With this in mind, we examine the possible causes behind the reported issues and perform a complete analysis of the PA/DM process execution in the Mexican center.

### *4.1. Question to be answered*

For this experiment, we aim to answer the following questions about the execution of PA/DM projects in a DMC.

1) RQ1:What do the dependency, frequency, and performance statistics of the process model reveal?

2) RQ2: How compliant is the discovered model vs. the CRISP-DM reference model?

3) RQ3: How is the interaction among resources during process execution?

4) RQ4: What are the possible causes for the reported issues?

### *4.2. Information Gathering*

We obtained qualitative and quantitative information from six real PA projects executed at Tecnologico de Monterrey DMC by interviewing modelers and managers. The format utilized for quantitative data gathering is available in Appendix A.

During this phase, five modelers and one manager were interviewed. The requested data include information from four CRISP-DM process stages (i.e., data understanding, data preparation, modeling, and evaluation), since we do not have access to data from the business understanding and deployment phases

Finally, the following data was obtained from stakeholders: start and finish dates of activities, the average number of hours per activity per day, and the number of resources involved in each activity. At the end of the interview, we request impressions about execution processes to identify specific issues.

### *4.3. Event Log Generation*

We create 4945 records with timestamps based on the provided information. The corresponding records per project can be regarded in Table 2. In this phase, no assumptions were considered since the modelers provide specific times and dates for each activity.

Table 2: Event Log record generation by the project.

| Project | Records generated |
|---|---|
| P1 | 504 |
| P2 | 1262 |
| P3 | 1496 |
| P4 | 463 |
| P5 | 599 |
| P6 | 621 |

### *4.4. Event Log Analysis*

We use Disco and ProM 6 applications to perform process mining. The first is a commercial tool that provides accurate process models [20]. While the second supports other types of functionalities like Petri nets and Social Networks [21]. Table 3 shows the modules used in each application.

Table 3: Models used in PM applications

| Application | Models used |
|---|---|
| Disco | (1) Map (2) Statistic 3) Filter by case |
| ProM | (1) Social Network Miner |

We use the Map view from the disco application to visualize the flow of activities, dependencies, frequencies, and performance. Likewise, the statistics view is used to identify the process event distribution, the activities, and the frequency of resources. Finally, we use the Filtering functionality to analyze the process model by specific cases. From ProM application, we use Social Network Miner to identify relationships among resources.

## 5. Results and Discussion

During process analysis, we document the global statistics shown in Table 4. As can be noted, the number of events represents the total records in the event log, and the cases correspond to the number of processes. The activities represent 16 generic tasks of the following phases of the CRISP-DM methodology: data understanding, data preparation, modeling, and evaluation. Finally, we examine projects that were executed in the next time range (April 25th, 2016 to June 30th, 2019). The statistics reveal that, on average, the project duration is 30 weeks.

Table 4: Process global statistics

| Metric | Value |
|---|---|
| Events | 4945 |
| Cases | 6 |
| Activities | 16 |
| Median Case Duration | 27 weeks |
| Mean Case Duration | 30 weeks |
| Start | 4/25/2016 9:00:00 |
| End | 06/30/2019 18:00:00 |

On the other hand, the statistics per activity showed in Table 5 reveal the most, the average, and the least executed activities in the project.

Table 5: Statistics per activity

| Activity | Relative Frequency |
|---|---|
| Build model | 19.38% |
| Evaluate results | 10.11% |
| Select modeling technique | 8.15% |
| Verify data quality | 7.28% |
| Construct data | 6.23% |
| Assess model | 6.15% |
| Explore data | 6.09% |
| Integrate data | 5.46% |
| Collect initial data | 5.42% |
| Review process | 5.04% |
| Determine next steps | 4.99% |
| Generate test design | 4.77% |
| Select data | 3.94% |
| Describe data | 3.28% |
| Clean data | 2.93% |
| Format data | .79% |

In the following subsections, we respond to the defined research questions.

*RQ1: What do dependency, frequency, and performance statistics of the process model reveal?*

Figure 3 shows the process map of the event log. As can be noticed, there are four thick arrows in the diagram that represents significant dependence among activities. For instance, the most substantial and unique bidirectional dependency is present between the review process and the determination of Next steps activities. Likewise, a significant reliance is visible between the process review and next steps activities, which means that those tasks execution order is the same in cases majority. Besides, managers should pay attention to the iteration that involves all data manipulation activities (i.e., collect, explore, verify, select, clean, construct, and integrate data) with the modeling selection technique. The evidence support that the team is having trouble with gaining knowledge of the data, which is a common problem in these kinds of models. We can assume that the lack of consistent execution of the description and selection of data could be the cause of the described problem.

Lastly, the diagram shows a dependency between the modeling technique selection and the initial data collection, which should be revised. Strangely, a change in data impacts the modeling technique and also the model construction. We recommend the inclusion of roles with expertise in modeling techniques to break that dependency.

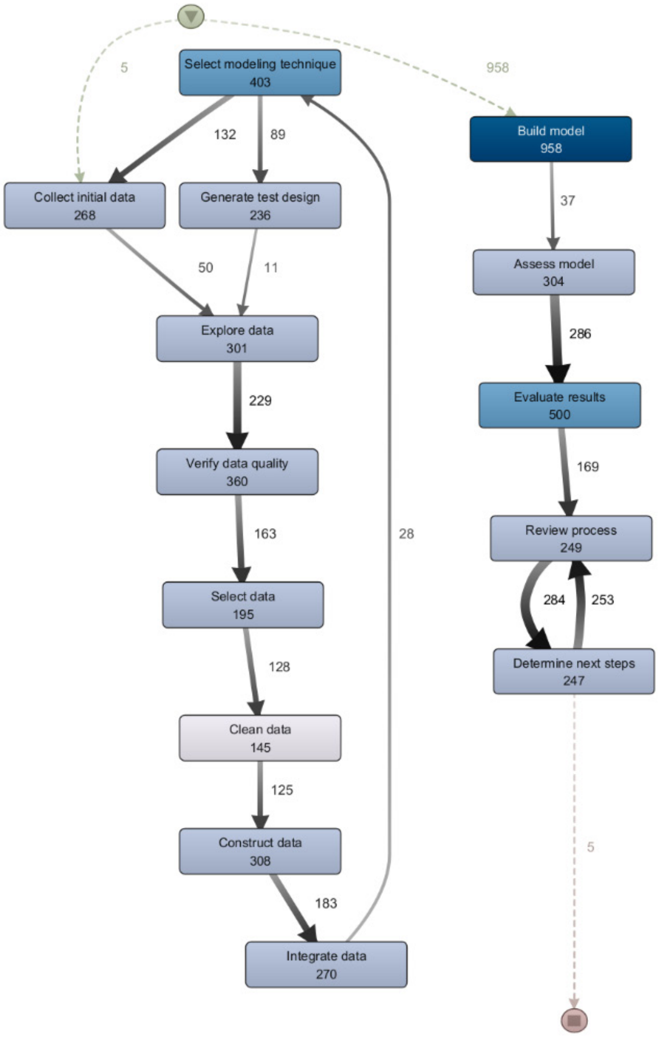

Figure 3: Process map by absolute frequency

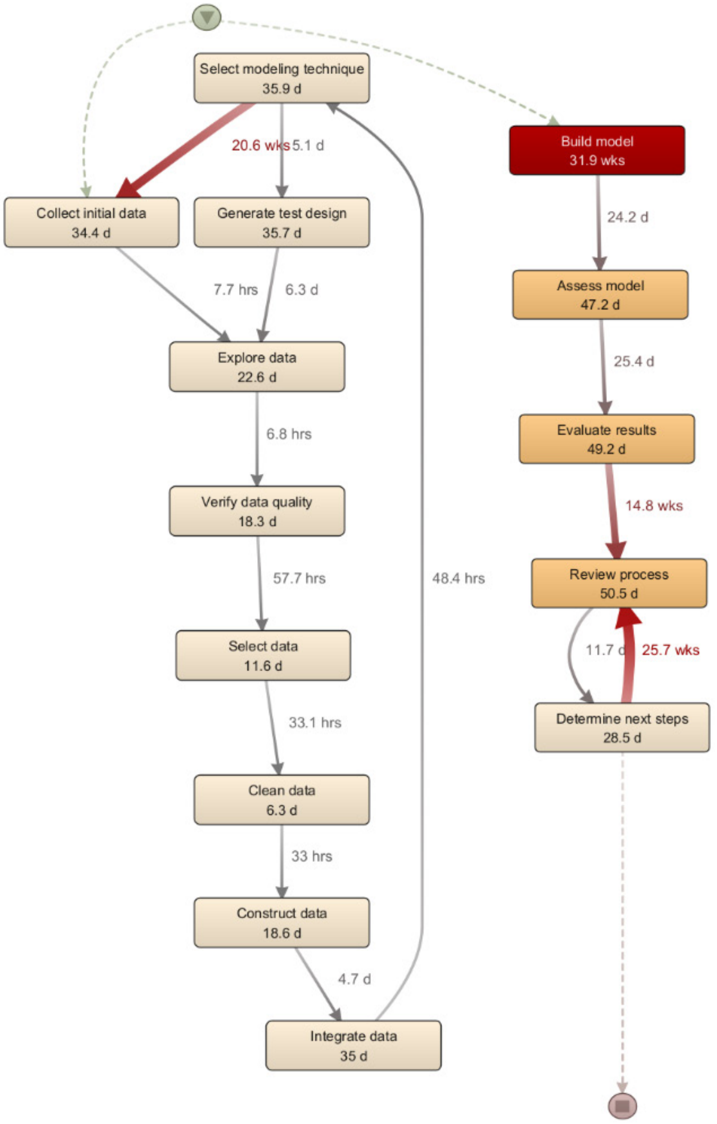

Figure 4: Process map by total duration

On the other hand, the absolute frequency of activities is represented with color. A high-frequency task is painted with intense blue, while one with low frequency is depicted with light blue. For this process model, the activity with the most significant frequency is the model's construction, followed by evaluating results and selecting the modeling technique. As we have mentioned, the model construction is affected by previous executed or non-executed activities. We can assume that previous activities improvement has a positive impact on the construction activity. In this case, we recommend using lean prototypes to facilitate the technique selection and diminish the time devoted to the model construction.

Finally, the performance of the model can be regarded in Figure 4. The model shows the total task duration and delays between activities. The model´s construction is the most significant task with 31.9 wks. Likewise, a delay of 20.6 weeks is present between the select modeling technique and collect initial data activities. In this case, managers should focus on diminishing the time between those two activities, by involving more resources or/and experts to the project.

A second delay is exposed between the process review and the definition of the next steps; however, this case should be analyzed separately since all resources execute these activities simultaneously, and that variable could affect the metric and not represent the real delay.

*RQ2: How compliant is the discovered model vs the CRISP-DM reference model?*

Figure 5 shows the process map by case frequency that is useful to analyze compliance. For instance, we are examining six cases, and theoretically, all activities should be present in all cases; however, in reality, this is not the case. Specifically, for this DMC, we discovered that the activities with a lower presence in the execution are the data formatting and data description. This last represents an issue in subsequent tasks since modelers report that the lack of data description has delayed the model's development and integration processes. Likewise, data cleaning and construction activities have problems with compliance in one of the cases, so DMC managers need to review these deviations.

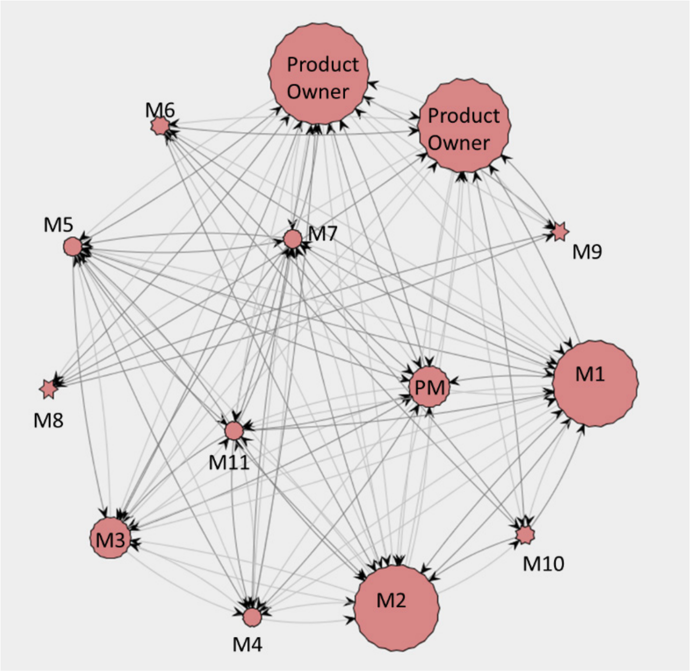

Figure 6: Social Networks of the process execution

Finally, the flow of activities has some compliance issues. In contrast to CRISP-DM, the actual execution is iterative in the data understanding and preparation phases. With this information, managers can create initiatives to align the model to CRISP-DM by stages and increase its performance.

*RQ3: How is the interaction among resources during process execution?*

To answer this question, we use the Social Network mining capability of ProM. As can be regarded in Figure 6, the two product owners are critical intermediaries in the project execution. In this case, the PM role is key in the process; however, the manager seems distant from individual modelers. On the other hand, M1 and M2 modelers seem to be more connected to the group. This last can have two explanations: (1) The existence of a functional dependency among parties (e.g., infrastructure, software, etc.) or (2) the modeler has participated in several projects which allow him to collaborate with more people. Finally, managers must pay attention to isolated modelers M8 and M9 and understand why they are separated from the group.

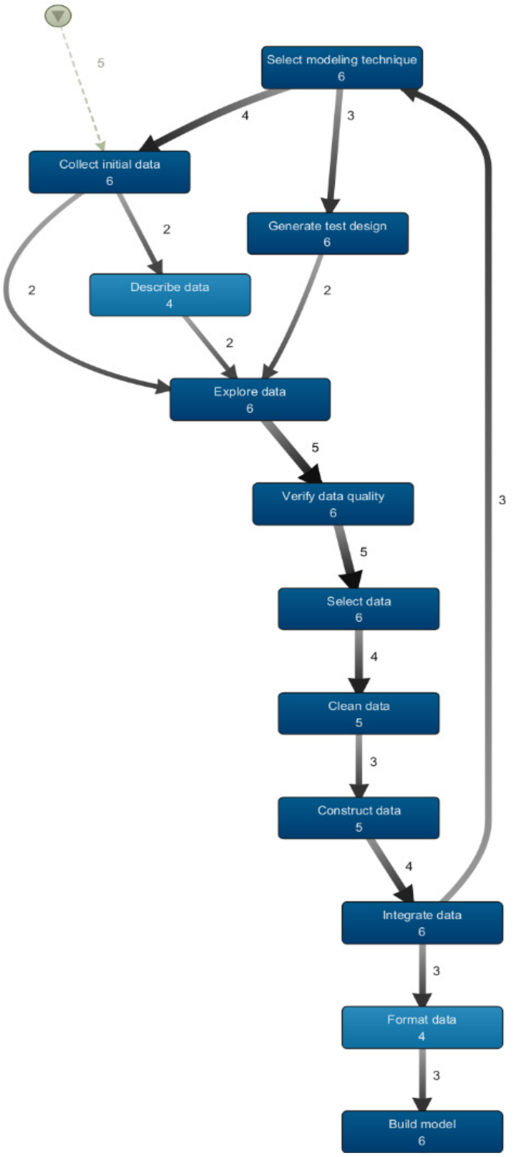

Figure 5: Process map by case frequency

*RQ4: What are the possible causes for the reported issues?*

We analyze the possible causes of the reported issues using previous process maps.

*Wrong selection of modeling technique*. It seems that the modeling technique selection is an exploratory process that takes too much time to define. As we have recommended before, inclusion of an expert and the implementation of prototypes can address this problem. Since the modeling technique can be evaluated with a prototype and with the expert support.

*Lack of standardization of data glossary among models*: This problem is caused by the lack of execution of the data documentation activity. We believe that an alignment to CRISP-DM methodology can solve this problem.

Finally, it is relevant to mention the limitations of the present study, which can be addressed in future research.

First, the presented model represents PA/DM execution processes of DMCs exclusively, so other PA/DM processes outside this environment are not represented in the research. Second, we don't include the business understanding and deployment phases in the modeling given data restrictions. So we represent part of the PA/DM process in this study. Lastly, the absence of previous research limits the possibility to compare and contrast our model with others and evaluate its completeness.

## 6. Conclusions and Future Work

In this research, we reveal the value of process mining as a tool to support project management of PA/DM projects in DMCs. Likewise, we expose the need to implement mature PA/DM processes in DMCs that facilitate (1) project management and (2) process improvement.

In this study, we create a process model to identify project execution issues, gaps in compliance, and the interaction of resources. We perform interviews to obtain detailed data from the PA process execution from modelers and managers. An event log at the level of activities was created considering CRISP-DM generic tasks, timestamps, and resources. Disco application was used to apply process discovery and process enhancement techniques. ProM application was used to perform Social Network mining. The results of the study reveal that: (1) Process mining models are helpful to analyze and address common management issues of PA/DM projects (2) PA/DM projects require alignment to mature methodologies to improve process performance and avoid execution problems (3) PA/DM project execution should be revised at the activity level to identify issues and to propose specific strategies (4) PA/DM projects should be analyzed from different perspectives to obtain valuable information for the management team.

Although it has been proved that Process Mining techniques are useful tools to support the management of PA/DM projects, there is work that needs to be addressed in the future. For instance, we need to use ProM tool to obtain compliance metrics of process models. Likewise, we need to use additional social network algorithms to analyze other organizational relationships.

## Conflict of Interest

The authors of this paper certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

## Acknowledgment

This work was supported by CONACYT Project 266632, identified as "Binational Laboratory on Smart Sustainable Energy Management and Technology Training." Additionally, the authors would like to thank the following experts for the provided information about PA project execution: Jose Hugo Carmona, Luis Alberto Lira, Ricardo Granados, Aban Moreno, Eduardo Meza, and Paolo Paez.

## References

[1] R. Edsall, K. Larson, "Decision making in a virtual environment: effectiveness of a semi-immersive 'Decision Theater' in understanding and assessing human-environment interactions," 2006.

[2] M.O.S. Escobar, R.L. Espinosa, J.M.M. Espinosa, J.J.N. Monroy, G. V Solar, "Applying Process Mining to Support Management of Predictive Analytics/Data Mining Projects in a Decision Making Center," in 2019 6th International Conference on Systems and Informatics (ICSAI), 1527–1533, 2019, doi:10.1109/ICSAI48974.2019.9010135.

[3] W. van der Aalst, "Process Mining: Overview and Opportunities," ACM Trans. Manage. Inf. Syst., **3**(2), 2012, doi:10.1145/2229156.2229157.

[4] C. dos S. Garcia, A. Meincheim, E.R. Faria Junior, M.R. Dallagassa, D.M.V. Sato, D.R. Carvalho, E.A.P. Santos, E.E. Scalabrin, "Process mining techniques and applications – A systematic mapping study," Expert Systems with Applications, **133**, 260–295, 2019, doi: 10.1016/j.eswa.2019.05.003.

[5] J. Joe, T. Emmatty, Y. Ballal, S. Kulkarni, "Process mining for project management," in 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), 41–46, 2016, doi:10.1109/SAPIENCE.2016.7684142.

[6] G. Cordeiro, F. Deschamps, E. Pinheiro de Lima, Managing a Big Data/Analytics project: a systematic literature review, 2017.

[7] G. Kabanda, "An Evaluation of Big Data Analytics Projects and the Project Predictive Analytics Approach," Oriental Journal of Computer Science and Technology, **12**, 132–146, 2020.

[8] R. Nisbet, G. Miner, K. Yale, The Data Mining and Predictive Analytic Process, Elsevier: 39–54, 2018, doi:10.1016/b978-0-12-416632-5.00003-7.

[9] V. Kotu, B. Deshpande, Data Science Process, 19–37, 2019, doi:10.1016/B978-0-12-814761-0.00002-2.

[10] A. Azevedo, M. Santos, KDD, semma and CRISP-DM: A parallel overview, 2008.

[11] R. Wirth, J. Hipp, "CRISP-DM: Towards a standard process model for data mining," Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 2000.

[12] R. Nayak, T. Qui, "Data Mining Application in a Software Project Management Process," in Proceedings 3rd Australasian Data Mining Conference (AusDM04), University of Technology Sydney: 99–109, 2004.

[13] P. Pospieszny, "Application of Data Mining Techniques in Project Management – an Overview," Collegium of Economic Analysis Annals, 199–220, 2017.

[14] J. Balsera, V. Montequín, F. Ortega-Fernández, C. Alba, Data Mining Applied to the Improvement of Project Management, 2012, doi:10.5772/48734.

[15] G. Schuh, M. Riesener, C. Dölle, "Implementation and assessment of a predictive analytics model for development project management," 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 696–700, 2017.

[16] G. Guillaume-Joseph, J. Wasek, "Improving software project outcomes through predictive analytics: Part 2," IEEE Engineering Management Review, **43**, 39–49, 2016, doi:10.1109/EMR.2015.2469471.

[17] E.S. Awolumate, Using Predictive Analytics to Deliver an Improved IT Project Cost Performance Model, 2020.

[18] E. Kouzari, L. Sotiriadis, I. Stamelos, "Process mining for process conformance checking in an oss project: An empirical research," in IFIP International Conference on Open Source Systems, Springer: 79–89, 2018.

[19] C.N. Corella, J. Noguez, E.J.M. Molina, E. Sotkoeva, R. Salla, "Use of a decision-making laboratory to support student's visual analysis for the solution of a transportation problem in Mexico City," in 2019 IEEE Frontiers in Education Conference (FIE), 1–5, 2019, doi:10.1109/FIE43999.2019.9028550.

[20] C.W. Günther, A. Rozinat, "Disco: discover your processes," in: Lohmann, N. and Moser, S., eds., in Proceedings of the Demonstration Track of the 10th International Conference on Business Process Management (BPM 2012), CEUR-WS.org: 40–44, 2012.

[21] B. Dongen, H. Verbeek, A. Weijters, W. Aalst, The ProM Framework: A New Era in Process Mining Tool Support, 2005, doi:10.1007/11494744_25.

## Appendix A. Data gathering format

| | | | | | |
|---|---|---|---|---|---|
| **Project Name** | | | | | |
| **Start date** | | **End date** | | ▨ | |
| | | | | | |
| **Phase: Data Understanding** | | | | | |
| **Activity** | **Start date** | **End date** | **Hours/day/rol** | **Resource** | **Order** |
| Collect initial data | | | | | |
| Describe data | | | | | |
| Explore data | | | | | |
| Verify data quality | | | | | |
| **Phase: Data Preparation** | | | | | |
| **Activity** | **Start date** | **End date** | **Hours/day/rol** | **Resource** | **Order** |
| Select data | | | | | |
| Clean data | | | | | |
| Construct data | | | | | |
| Integrate data | | | | | |
| Format data | | | | | |
| **Phase: Modeling** | | | | | |
| **Activity** | **Start date** | **End date** | **Hours/day/rol** | **Resource** | **Order** |
| Select modeling technique | | | | | |
| Generate test design | | | | | |
| Build model | | | | | |
| Assess model | | | | | |
| | | | | | |
| **Phase: Evaluation** | | | | | |
| **Activity** | **Start date** | **End date** | **Hours/day/rol** | **Resource** | **Order** |
| Evaluate results | | | | | |
| Review Process | | | | | |
| Determine Next Steps | | | | | |