# Identifying and Explaining Undesirable Traces in Business Processes Using Descriptive and Predictive Analysis

Ali Norouzifar
*Process and Data Science Group (PADS)*
*RWTH Aachen University*
Aachen, Germany
ali.norouzifar@pads.rwth-aachen.de

***Abstract*—This study proposes data analytic approaches to identify possibly undesirable traces in complex business processes. Two types of analysis are employed: descriptive and predictive. In the descriptive analysis, cohorts of traces with different control flows are identified based on a case attribute or process indicator. A hierarchical clustering approach is used for this purpose by employing the Earth Mover's Distance to measure differences between the cohorts. In the predictive analysis, machine learning models are trained to predict the cohort for unseen cases, and SHAP values are used as an explainable AI technique to extract the most critical features. The approach is demonstrated through experiments that show its practical and effective use in identifying possibly undesirable traces and providing insights to improve the process or discover better process models.**

***Index Terms*—process mining, process comparison, undesirable behavior, business process improvement.**

## I. Introduction

Process mining employs data science methods to analyze event data generated by business processes with the aim of deriving descriptive models and improving performance. However, real-life processes are often complex and involve different handling procedures depending on the case's specifications [1]. This raises the question of whether all cases follow a desirable procedure [2]. The process owners are not always aware of the deficiencies occurring in their processes. In this study, we propose some data analytic approaches to assist them in finding the cases that behave undesirably. For example, consider the duration of cases in a business process be between 10 days to 300 days. The output of this research may propose that cases with duration below 20 days and cases with the duration above 100 days have a very different handling procedure and therefore, based on a specialist decision could be categorized as undesirable.

Our study involves two types of analysis: descriptive and predictive. In the descriptive analysis, we focus on identifying cohorts, which are groups of traces with different control flows. These cohorts are generated based on the values of a case attribute or a process indicator. In the predictive analysis, we use these cohorts as class labels to train machine learning models that predict the cohort for unseen cases. We then apply explainable AI techniques to determine the features that play a critical role in the predictions.

This research was supported by the research training group "Dataninja" (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia.

## II. Cohorts Analysis

### A. Descriptive Analysis

The stochastic distance between cohorts is measured using the Earth Mover's Distance (EMD) in this study [3]. EMD is used in process mining literature to find the distance between the stochastic language of two event logs [4]. The proposed strategy is to partition the event log into smaller segments based on a feature of interest and then use hierarchical trace clustering to merge the partitions based on the earth movers' distance. A user-defined distance threshold is used to stop the hierarchical clustering when distance between all the clusters is larger than this threshold. In Fig. 1, an overview of the descriptive analysis is illustrated. The resulting cohorts are different from each other considering the control flow. In the next step a process expert should check the cohorts to see if a cohort should be considered as undesirable or not.

### B. Predictive Analysis

Next to finding interesting cohorts, we may train some machine learning models to predict the cohort based on some designed control flow features. In this step, we can optionally incorporate a normative process model to make more valuable features, e.g., we can capture if an event in the event log could be replayed by the normative process model or which transitions occurred in decision points. After training and evaluating a machine learning model using well-known machine learning evaluation metrics such as accuracy, recall and precision, SHAP (SHapley Additive exPlanations) values are used as a explainable AI technique to extract the most important features contributing in the prediction. With SHAP values, we are also able to find the direction in which these features have effect on the predictions. An overview of the predictive analysis is illustrated in Fig. 1.

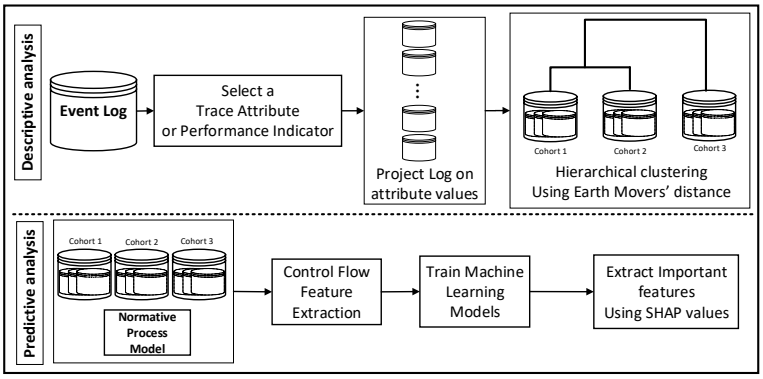


Fig. 1: An overview of the proposed descriptive and predictive frameworks to generate and analyze cohorts.

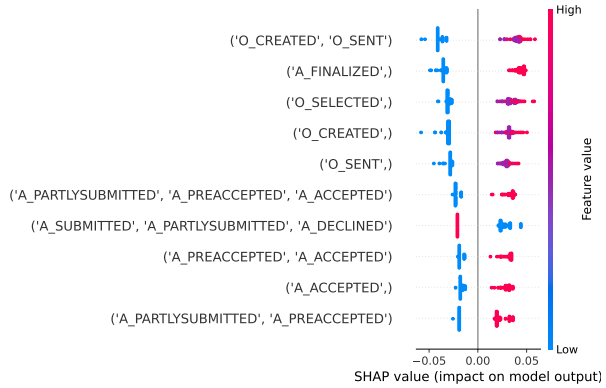


Fig. 2: The 10 most important features extracted from random forest model.

## III. Experiment

The proposed framework has been evaluated using several real-life event logs. In this paper, we present an example experiment conducted on the BPIC 2012 event log, which contains the handling process of loan applications and includes the application and offer sub-processes. The BPIC 2012 event log comprises 13087 applications and provides a realistic publicly available dataset for evaluating process mining methods [1].

The event log is divided into smaller partitions based on the throughput time of traces, represented by the process indicator $dur$. To achieve this, the equal frequency binning technique is used, resulting in 20 equally frequency bins. Then, a hierarchical clustering approach is applied to group similar traces together and form cohorts. The clustering algorithm stops when the distance between all clusters is greater than the user-defined threshold of 0.25. In this experiment, the algorithm produced three distinct cohorts: $coh_1$ with $0 \leq dur \leq 36$ $hours$, $coh_2$ with 36 $hours < dur \leq 31$ $days$, and $coh_3$ with 31 $days < dur \leq 3$ $months$. The Earth Mover's Distance (EMD) is used to measure the stochastic distance between cohorts, with $EMD(coh_1, coh_2)=0.77$ and $EMD(coh_2, coh_3)=0.30$. The resulting cohorts suggest that the process behavior may vary significantly when case duration is very short or very long.

Several machine learning models are trained using control-flow features. We use $n$-gram encoding with $n \in \{1, 2, 3\}$ to capture the behavior patterns. In this experiment, we focus on binary classifiers that distinguish between cases labeled as $coh_1$ or $coh_2$. To ensure generalization, we split the event log into training and testing sets with 30% of the cases in the test set. We compared different machine learning techniques, including random forest, gradient boosting, logistic regression and support vector machines. We used accuracy as the evaluation metric to select the best model, which is the random forest model with an accuracy of 92% in this experiment. The 10 most important features using SHAP values are represented in Fig 2. This plot suggest which control-flow related features contributed more to the predictions.

## IV. Conclusion

The experimental results show that this approach offers valuable insights into identifying potentially undesirable traces. It provides a practical and effective way to analyze and understand complex processes, which can be used for process improvement and better process discovery. Although the initial results are promising, further justification is required for the design choices, and the framework could be enhanced to make it applicable to real-life use cases.

[1] https://data.4tu.nl/articles/BPI_Challenge_2012/12689204

## References

[1] M. De Leoni, W. M. P. van der Aalst, and M. Dees, "A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs," *Information Systems*, vol. 56, pp. 235–257, 2016.

[2] A. Norouzifar and W. M. P. van der Aalst, "Discovering process models that support desired behavior and avoid undesired behavior," in *SAC '23: The 38th ACM/SIGAPP Symposium on Applied Computing, March 27-March 31, 2023, Tallinn, Estonia*, 2023.

[3] S. J. Leemans, S. Shabaninejad, K. Goel, H. Khosravi, S. Sadiq, and M. T. Wynn, "Identifying cohorts: Recommending drill-downs based on differences in behaviour for process mining," in *Conceptual Modeling: 39th International Conference, ER 2020, Vienna, Austria, November 3–6, 2020, Proceedings 39*. Springer, 2020, pp. 92–102.

[4] S. J. Leemans, A. F. Syring, and W. M. P. van der Aalst, "Earth movers' stochastic conformance checking," in *Business Process Management Forum: BPM Forum 2019, Vienna, Austria, September 1–6, 2019, Proceedings 17*. Springer, 2019, pp. 127–143.