International Journal on Advanced Science Engineering Information Technology

Vol.11 (2021) No. 1
ISSN: 2088-5334

Performance Analysis of Heuristic Miner and Genetics Algorithm in Process Cube: a Case Study

Rachmadita Andreswari[a,1], Ismail Syahputra[b], Muharman Lubis[a,2]

[a]*Information System, Telkom University, Ters. Buah Batu No.1, Bandung, 40257, Indonesia*
[b]*Information System, Widyatama University, Jl. Cikutra No.204A, Bandung, 40125, Indonesia*
Corresponding author: [1]andreswari@telkomuniversity.ac.id, [2]muharmanlubis@telkomuniversity.ac.id

***Abstract*— Databases that are processed in the form of Online Analytical Processing (OLAP) can solve large query loads that cannot be resolved by transactional databases. OLAP systems are based on a multidimensional model commonly called a cube. In this study, OLAP techniques are applied in process mining, a method for bridging analysis based on business process models with database analysis. Like data mining, process mining produces process models by implementing the algorithms. This study implements the heuristic miner algorithm compared with genetic algorithms. The selection of these two algorithms is due to the characteristics to be able to model the event log correctly and can handle the control-flow. The capability in handling control-flow including the ability to detect hidden task, looping, duplicate task, detecting implicit/explicit concurrency, non-free-choice, the ability to mine and exploiting time, overcoming noise, and overcome incompleteness. The results of conformance checking on the heuristic miner algorithm for all data, fitness values, position, and structure are 1, 0.495, and 1, while the results of the genetic algorithm are 0.977, 0.706 and 1. Both algorithms have good ability in modeling processes and have high accuracy. The results of the F-score calculation on the heuristic miner algorithm for all data is 0.622, while the result in the genetic algorithm is 0.820. It indicates that genetic algorithms have better performance in modeling event logs based on process cube.**

***Keywords*— Process mining; process cube; OLAP; heuristic miner algorithm; genetics algorithm.**

Manuscript received 2 Apr. 2020; revised 30 Aug. 2020; accepted 25 Nov. 2020. Date of publication 28 Feb. 2021.

I. INTRODUCTION

Process mining is currently a rapidly growing study. Process mining is very promising in understanding processes. This technique has been widely used in various industry segments such as health, manufacturing, information and communication technology, education, finance, and logistics [1]. Process mining can construct a process model by mining event log [2]. It can extract knowledge about business processes and obtain process models that can be used to improve business processes [3], [4]. On-Line Analytical Processing (OLAP) is an approach to performing processing and analysis in the data warehouse. There are two OLAP category models, namely MOLAP (multidimensional OLAP) and ROLAP (relational OLAP) [5]. OLAP is oriented explicitly towards analysis to support the decision-making process [6]. OLAP databases are also capable of resolving large query loads and a variety of data that cannot be resolved by relational databases [7]. OLAP systems are based on a multidimensional model commonly called a cube. OLAP is able to extract knowledge in a data warehouse or data mart to provide navigation through data to users [8]. Nowadays, the implementation of the OLAP concept is not only done in the data warehouse, but also in the mining process. The mining process bridges the analysis of process-based process models (e.g. movement) with database analysis (e.g. data mining, and machine learning) [9]. The process of implementing OLAP in the mining process is called a process cube [10]. Process cube bridges in making process models where event logs are arranged based on different dimensions.

Process cube implements the operating concepts that exist in OLAP, such as slice, dice, roll-up, and drill-down [6]. There are two challenges in implementing the process cube on an event log. First, the issue of performance, namely the level of unloading sparsity cells, while the latter is performance problems. There may be two variants of the same model, which the process cube is closely related to the concept of decomposition in process mining, where both of them do an event log split [11].

Research on process cube has been conducted in the last few years, including the implementation of the process cube

in solving software defect problems. The research uses a case study from the Google Chromium project, which there are nine dimensions in the process cube. The results of this study compare to the process mapping with the better result of process mining regarding performance and security issues [12]. Subsequent research shows that an organization is interested in comparing process mining to see how the process can be improved by understanding the differences between case groups, departments, and others. The use of process cube can be proposed to organize event logs in multidimensional data structures that are suitable for process mining [13]. In his research, Vogelgesang has provided an in-depth and complete description of the cube process [14]–[16]. This study continues the previous research that contributed to implementing the cube process in ERP goods production.

The variety of algorithms with their respective criteria used for mining makes it difficult for us to choose.[17]. Thus, this study compares the performance of algorithms in their application to multidimensional process mining and analysis of the results obtained following conformance checker measurements. The algorithm used in this study is the heuristic miner algorithm and genetic algorithm. Also, genetic algorithms have several advantages, which it can detect short-loops, hidden task, robust against logs that contain noise, incompleteness, and can see non-free choices. Meanwhile, the disadvantages of genetic algorithms are less stable with AND split, OR join, especially if there is an extended parallel branching model, and the execution time needed is long enough, especially when compared to the execution time of the alpha ++ algorithm [18].

The advantage of the heuristic miner algorithm can calculate the frequency of relations between activities that occur in the event log to determine the possibility of causal dependence or causal dependency between these activities. This advantage makes the heuristic miner able to handle noise in the event log [19]. It is also able to detect non-free-choice, noise, loops, incompleteness, and having a short execution time in the process. The selection of these two algorithms is due to the characteristics to be able to model the event log correctly and able to handle the control-flow [20]. Some of the capabilities of the algorithm according to the flow-control perspective are detecting hidden tasks, repetition, duplicate tasks, detecting implicit/explicit concurrency, non-free choices, mining and exploiting time, overcoming noise, and overcoming incompleteness[21].

II. Materials and Method

The methods of conducting this research are as follows (Figure 1). The first is to identify activities in the production-planning module in ERP applications, in this case, SAP. When the activity has been identified, the next step is to make product categorization according to the data obtained from the ERP application event log. This categorization is required to find out the types of products that companies make according to software specifications and business requirements. After that, the researcher downloaded the user log, which should follow the format file with the extension of .xls, .csv, .txt, or other compatible ones. Also, data cleaning is required so that the application logs can be made into a process model (Table 1)

TABLE I
Sample Event Log of Production Planning

ID	Case	Timestamp
31539965	Start	12/27/12 16:00
31539966	Start	12/27/12 16:00
31539966	Run MRP	12/27/12 16:30
31539966	Create Plan Order	12/28/12 6:00
31539966	Change Plan Order Date	1/2/13 11:00
31539966	Run MRP	1/3/13 16:30
31539966	Create Plan Order	1/4/13 6:00
31539966	Change Plan Order Date	1/9/13 11:00
31539966	Run MRP	1/10/13 16:30
31539966	Create Plan Order	1/11/13 6:00
31539966	Change Plan Order Date	1/16/13 11:00
31539966	Run MRP	1/17/13 16:30
31539966	Create Plan Order	1/18/13 6:00

Afterward, a cube operation is performed, which is slicing data according to the product category, namely male, female or kids. Furthermore, the process involved converting logs into files in the (.mxml) extension to adjust input in process mining. Process modeling using ProM, the application, which was developed by Technische Universitat Eindhoven. It is executed with the heuristic miner algorithm and genetic algorithm. Both algorithms have excellent and proven performance in the modeling processes, and this study will compare them in this process modeling. On the other hand, the modeling results are converted into Petri net diagrams. It is necessary to do conformance through checking measurement to determine the performance of each algorithm in modeling the process cube. This process will calculate and evaluate how well the event log matches the process model, called business alignment [22]. The most important measurement of conformance checking is fitness, but in this study, it also used the dimensions of advanced behavioral appropriateness and structural appropriateness.

A. Fitness

The fitness dimension calculates how many events from a record are recorded in the business process model. Fitness values are in the range 0 - 1 and can be calculated using the following formula (1) :

$$f = \frac{1}{2}\left(1 - \frac{\sum_{i=1}^{k} nimi}{\sum_{i=1}^{k} nici}\right) + \frac{1}{2}\left(1 - \frac{\sum_{i=1}^{k} niri}{\sum_{i=1}^{k} nipi}\right) \quad (1)$$

With information:

k = the number of different trace records. For each log trace i ($1 \leq i \leq k$)
ni = number of process instances from trace i
mi = number of tokens missing from trace i
ci = number of tokens consumed by trace i
ri = the number of tokens left from trace i
pi = number of tokens produced from trace i.

B. Precision

The precision dimension describes how many events might be formed, but not based on event data. Precision values are in the range 0 - 1, and can be calculated using the following formula “Advanced Behavioral Appropriateness” (2):

$$a'B = \left(\frac{|S_F^1 \cap S_F^m|}{2\,.\,|S_F^m|}\right) + \left(\frac{|S_P^1 \cap S_P^m|}{2.\,|S_P^m|}\right) \quad (2)$$

With information: relationship

S_F^m = "Sometimes follows" relations for the process model
S_P^m = "Sometimes precedes" relation for the process model
S_F^1 = "Sometimes follows" relations for event logs
S_P^1 = relationship "Sometimes precedes" for the event log

C. Structure

This dimension shows the ability of the model to handle the XOR and AND processes. This measurement is related to the perspective of control flow, and often there are several syntactic ways to express behavior in the model process. Structural values can be calculated using the following formula (3):

$$a's = \frac{|T| - (|T_{DA}| + |T_{IR}|)}{|T|} \quad (3)$$

With information:

T = transition in the Petri net model
T_{DA} = a collection of alternative duplicate tasks
T_{IR} = a collection of redundant invisible tasks

Based on the measurement results, the performance and the ability of the algorithm will be analyzed in modeling the event log based on the process cube.

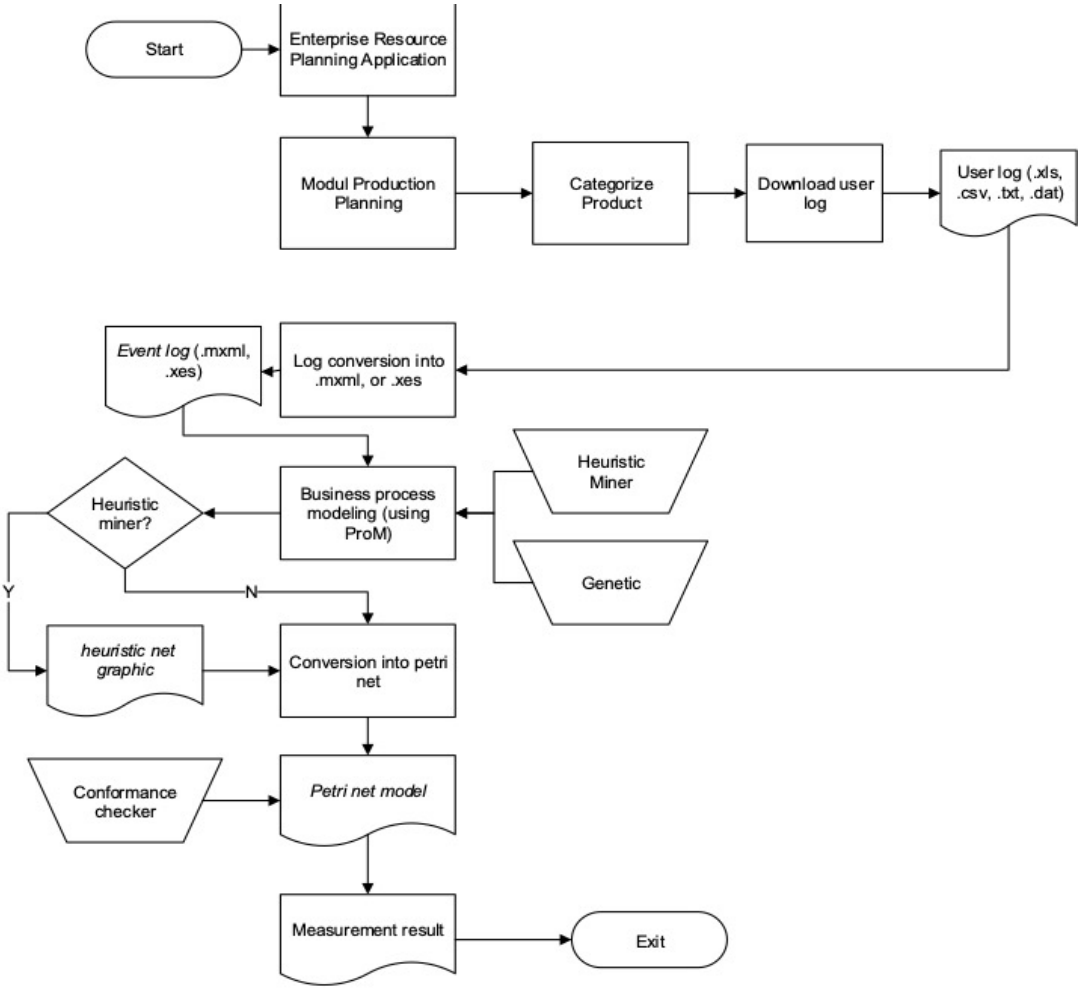

Fig. 1 Research methodology

III. Results and Discussion

In this study, the result of process cube implementation in process modeling production planning can be seen in Figure 2. Each process model is an implementation of the heuristic miner algorithm and genetic algorithm in each slicing category, namely male, female, and kids. The slicing categories are based on the type of goods produced following the target consumers.

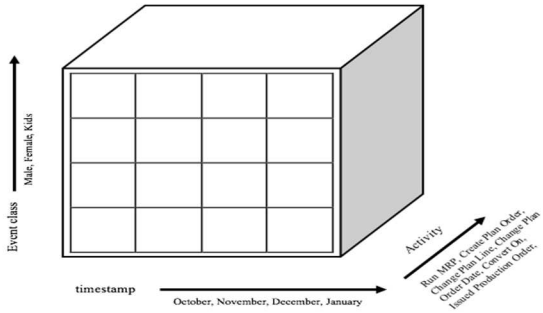

Fig. 2 Production planning cube design [16]

As shown in Figure 2, the dimensions of the process cube used are the event class, timestamp, and activity. The following are the statistics of the data used in this study; there are 3 product categories: male, female, and kids. There are 1293, 3186, and 1638 event logs, respectively (Table II).

TABLE II
EVENT LOG STATISTIC

	Male	Female	Kids	All
Event	1293	3186	1638	6117
Case	115	317	150	582
Activity	8	8	8	8
Start Date	December 11th, 2012	October 15th, 2012	December 6th, 2012	October 15th, 2012
End Date	January 21st, 2013	January 26th, 2013	January 22nd, 2013	January 26th, 2013

Every event log has eight activities, i.e., Start, Run MRP, Create Plan Order, Change Plan Line, Change Plan Line Order Date, Convert On, Issued Production Order, and Start Production. The event log is then modeled using the ProM application, in which the results are grouped according to product categories for each algorithm, as shown in Figure 3-8. Figures 3 and 4 are modeling for the Male category, where Figure 3 uses heuristic miners, and Figure 4 uses genetic algorithms. Based on the picture, in Figure 4, there are more hidden tasks (grey squares) and places (small white circles). The same thing was also found in the female category (figure 5 uses heuristic miners, and figure 6 uses genetic algorithms. Heuristic miners handle better-hidden tasks in both categories. In the kids' category shown by Figure 7 (heuristic miner) and Figure 8 (genetic algorithm), the number of hidden tasks and places is almost identical. This number indicates that both algorithms have a good ability to handle hidden tasks.

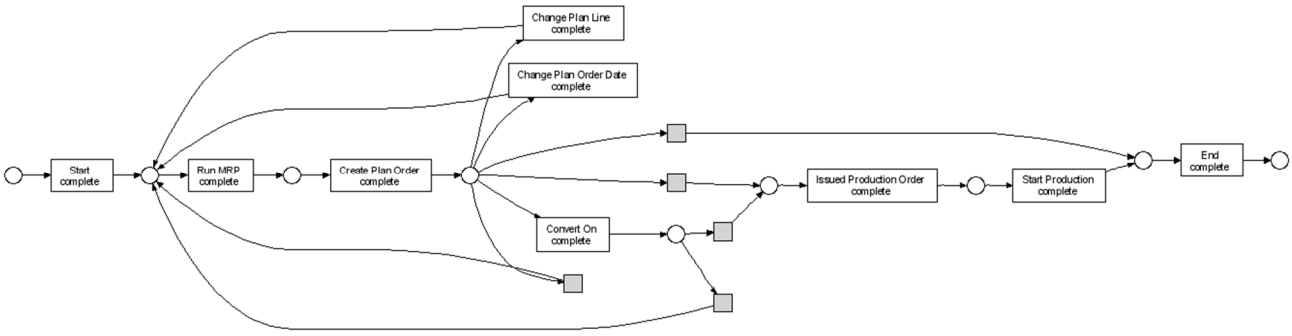

Fig. 3 Male process model using heuristic miner

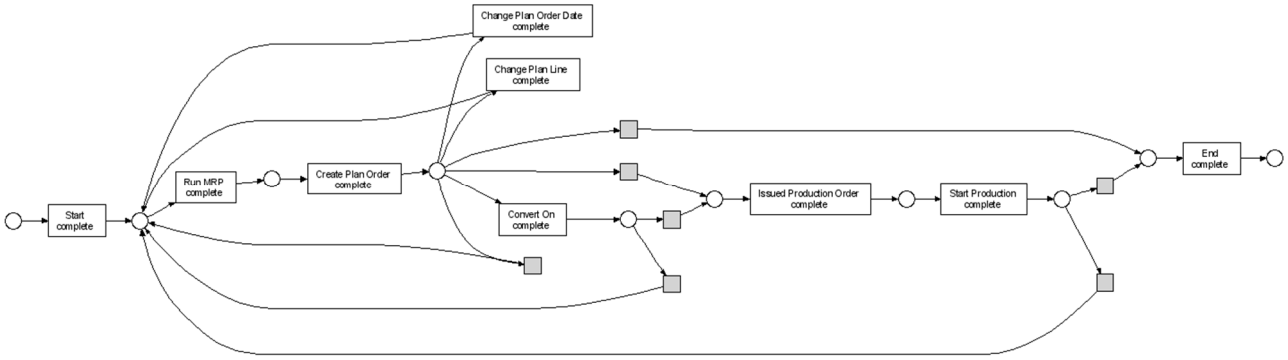

Fig. 4 Male process model using the genetic algorithm

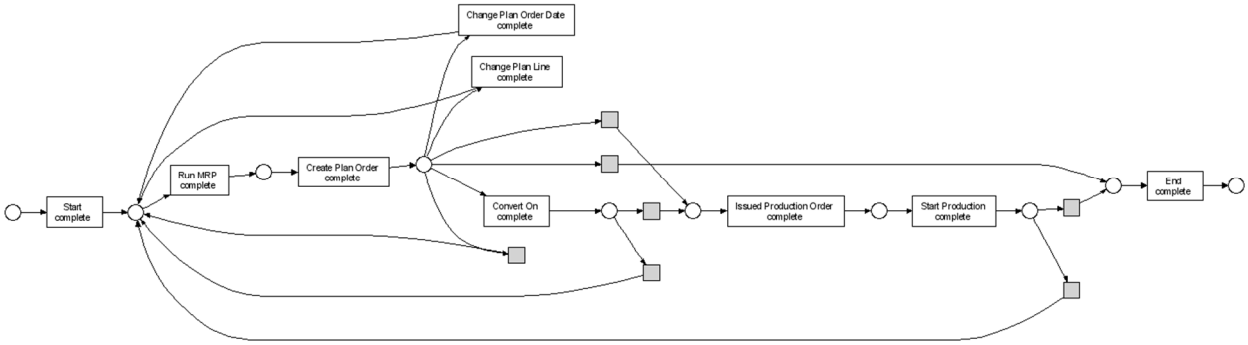

Fig. 5 Female process model using heuristic miner

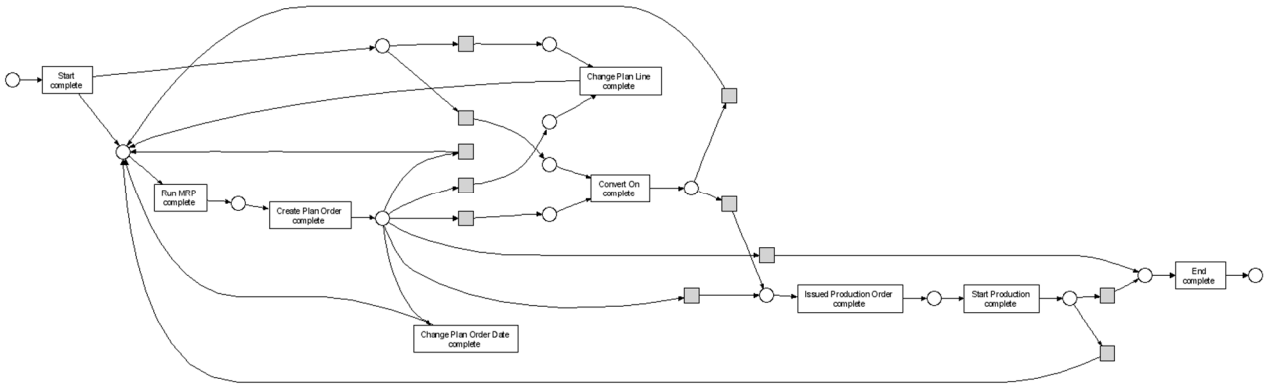

Fig. 6 Female process model using the genetic algorithm

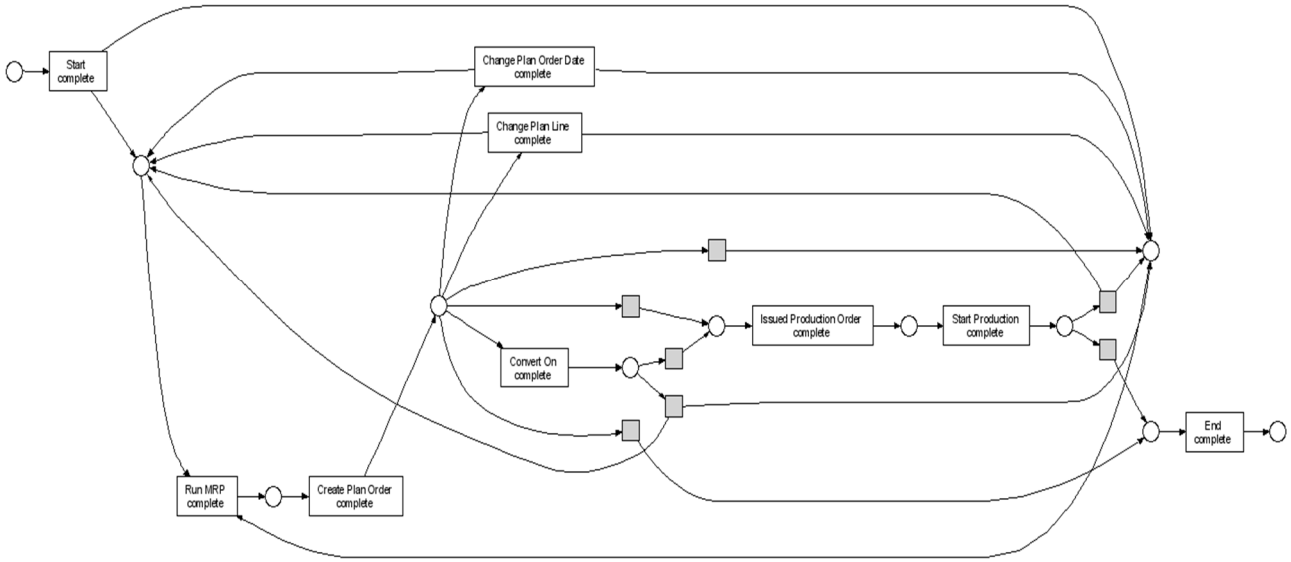

Fig. 7 Kids process model using heuristic miner

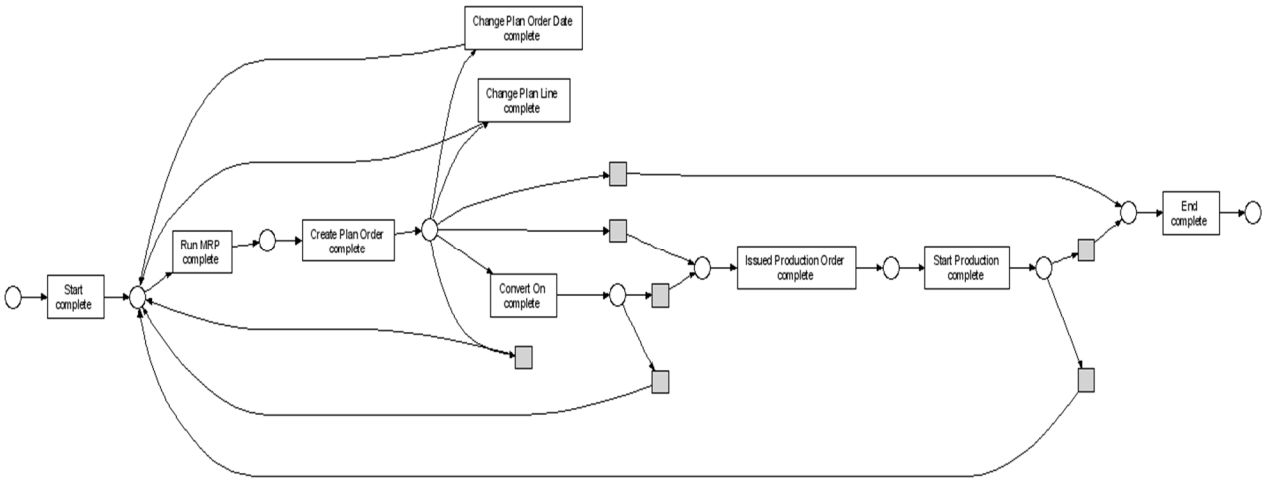

Fig. 8 Kids process model using the genetic algorithm

It is necessary to measure the performance based on conformance checking to determine the performance of each algorithm in the process cube. Based on Table 3 the highest fitness value in the male and kids' categories is obtained from modeling using genetic algorithms. In contrast, in the female category, the best fitness value is obtained from modeling using the heuristic miner algorithm. In precision measurements using advanced behavioral appropriateness, it was found that the best precision value in the male category was obtained from modeling using the heuristic miner algorithm. In contrast, the female category was obtained using genetic algorithms. In the kids' class, both algorithms show the same precision value. Moreover, the model processing time uses in the heuristic miner algorithm is faster than the

genetic algorithm; this is following previous studies [18], [19].

TABLE III
CONFORMANCE CHECKING PRODUCTION PLANNING MODEL PROCESS

Conformance Measure	Male	Female	Kids	All
Heuristic Miner Algorithm				
Fitness	0.996	1	0.956	1
Adv. Behavioral Appropriateness	0.56	0.495	0.495	0.495
Structure	1	1	1	1
Genetic Algorithm				
Fitness	1	0.995	1	0.977
Adv. Behavioral Appropriateness	0.495	0.542	0.495	0.706
Structure	1	1	1	1

In general, to measure the overall event log, the heuristic miner algorithm will generate the highest fitness value while the genetic algorithm generates the highest precision value. Meanwhile, the measurement of the overall structure, each category gets the maximum value. The F-score calculation considers the value of precision and recall. It was stated that one of the matrices in precision value is a'B (advanced behavioral appropriateness), and matrix recall is fitness [23]. This analysis is for calculating combinations of fitness and precision values with the F-score formula as follows:

$$F_\beta = 2.\frac{precision\ x\ recall}{precision + recall} \quad (4)$$

The β value on the F-score is weight, while at the real event log, fitness weights are twice as high as precision because fitness tends to be more important than precision. The results of the F-score calculation can be seen in Table IV. In the calculation, the heuristic miner algorithm has a better F-score in the male category. The female, kids, and overall event log categories found that genetic algorithms produce better values than the heuristic miner algorithm.

TABLE IV
F-SCORE MEASUREMENT

	Male	Female	Kids	All
Heuristic Miner	0.717	0.662	0.652	0.662
Genetics	0.662	0.702	0.974	0.820

IV. CONCLUSION

This study model and evaluate the business processes using mining process techniques with the process cube approach. Modeling business processes comes from event logs by comparing two process-mining algorithms: genetic algorithms and heuristic miner algorithms. In the process modeling results, which can be seen from the control flow, the model produced by the heuristic algorithm is different from the model produced by the genetic algorithm. The conformance checking on the heuristic miner algorithm shows that the fitness values for male, female, kids, and all are 0.996, 1, 0.956, and 1. Meanwhile, in the genetic algorithm, the fitness values are 1, 0.995, 1, and 0.977. For precision measurements, the results for heuristic miner are 0.56, 0.495, 0.495, and 0.495.

Furthermore, for genetic algorithms, the results are 0.495, 0.542, 0.495, and 0.706, respectively. The last conformance checking measurement, namely structure, obtained a value of 1 overall for the heuristic miner algorithm and genetics, the maximum value. The F-score value is calculated to find algorithms that can model the process better. In this calculation, it was found that genetic algorithms could model with accuracy better than the heuristic miner. In other results, Miner heuristic algorithms can model processes faster than genetic algorithms in terms of speed in modeling processes.

REFERENCES

[1] C. dos S. Garcia *et al.*, "Process Mining Techniques and Applications – A SystematicMapping Study," *Expert Syst. Appl.*, vol. 133, pp. 260–295, 2019, doi: 10.1016/j.eswa.2019.05.003.

[2] X. Zhang, Y. Du, L. Qi, and H. Sun, "An Approach for Repairing Process Models Based on Logic Petri Nets," *IEEE Accsess*, vol. 6, pp. 29926–29939, 2018, doi: 10.1109/ACCESS.2018.2843137.

[3] W. Li, Y. Fan, W. Liu, M. Xin, H. Wang, and Q. Jin, "A Self-Adaptive Process Mining Algorithm Based on Information Entropy to Deal with Uncertain Data," *IEEE Access*, vol. 7, pp. 131681–131691, 2019, doi: 10.1109/ACCESS.2019.2939565.

[4] X. Zhang, Y. Du, L. Qi, and H. Sun, "Repairing Process Models Containing Choice Structures via Logic Petri Nets," *IEEE Access*, vol. 6, pp. 53796–53810, 2018, doi: 10.1109/ACCESS.2018.2870727.

[5] Y. zhang, Y. Zhang, S. Wang, and J. Lu, "Fusion OLAP: Fusing the Pros of MOLAP and ROLAP Together for In-Memory OLAP," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 9, pp. 1722–1735, 2019, doi: 10.1109/TKDE.2018.2867522.

[6] A. Vaisman and E. Zimanyi, *Data Warehouse Systems Design and Implementation*, 1st ed. Berlin, Heidelberg: Springer, 2016.

[7] F. Davardoost, A. Babazadeh Sangar, and K. Majidzadeh, "Extracting OLAP Cubes from Document-Oriented NoSQL Database Based on Parallel Similarity Algorithms," *Can. J. Electr. Comput. Eng.*, vol. 43, no. 2, pp. 111–118, 2020, doi: 10.1109/CJECE.2019.2953049.

[8] M. R. Llave, "Business Intelligence and Analytics in Small and Medium-sized Enterprises: A Systematic Literature Review," *Procedia Comput. Sci.*, vol. 121, pp. 194–205, 2017, doi: https://doi.org/10.1016/j.procs.2017.11.027.

[9] W. Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, vol. 136. 2011.

[10] W. M. P. Van Der Aalst, "Process Cubes: Slicing, Dicing, Rolling Up and Dilling Down Event Data for Process Mining," in *Asia Pacific Business Process Management*, 2017, pp. 1–22, doi: 10.1007/978-3-319-02922-1_1.

[11] W. M. P. Van Der Aalst, "Process Cube: Turning Data into Value," 2014.

[12] M. Gupta and A. Sureka, "Process Cube for Software Defect Resolution," in *2014 21st Asia-Pacific Software Engineering Conference*, 2014, vol. 1, pp. 239–246, doi: 10.1109/APSEC.2014.45.

[13] A. Bolt and W. M. P. Aalst, van der, "Multidimensional Process Mining Using Process Cubes," in *Enterprise, Business-Process and Information Systems Modeling (16th International Conference, BPMDS 2015, 20th International Conference, EMMSAD 2015, Held at CAiSE 2015, Stockholm, Sweden, June 8-9, 2015, Proceedings)*, 2015, pp. 102–116, doi: 10.1007/978-3-319-19237-6_7.

[14] T. Vogelgesang and H. J. Appelrath, "Multidimensional Process Mining with PMCube Explorer," *CEUR Workshop Proc.*, vol. 1418, pp. 90–94, 2015.

[15] T. Vogelgesang, S. Rinderle-Ma, and H.-J. Appelrath, "A Framework for Interactive Multidimensional Process Mining," in *Business Process Management Workshops*, 2017, pp. 23–35.

[16] T. Vogelgesang, G. Kaes, S. Rinderle-Ma, and H.-J. Appelrath, "Multidimensional Process Mining: Questions, Requirements, and Limitations," in *CAISE 2016 Forum*, 2016, pp. 169–176.

[17] P. Weber, B. Bordbar, and P. Tino, "A Framework for the Analysis of Process Mining Algorithms," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 43, no. 2, pp. 303–317, Mar. 2013, doi: 10.1109/TSMCA.2012.2195169.

[18] A. J. M. M. Weijters, W. M. P. Aalst, van der, and A. K. Alves De Medeiros, *Process Mining with The Heuristics Miner Algorithm*. Technische Universiteit Eindhoven, 2006.

[19] R. Andreswari and M. Er, "Analisis Kinerja Algoritma Penggalian Proses untuk Pemodelan Proses Bisnis Perencanaan Produksi dan Pengadaan Material pada PT.XYZ dengan Kriteria Control-Flow," *J. SISFO*, vol. 5 No.1, pp. 1–8, 2014, doi: 10.24089/j.sisfo.2014.03.008.

[20] W. van der Aalst, T. Weijters, and L. Maruster, "Workflow Mining: Discovering Process Models from Event Logs," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1128–1142, 2004, doi: 10.1109/TKDE.2004.47.

[21] A. Rozinat and W. Aalst, "Conformance Checking of Processes Based on Monitoring Real Behavior," *Inf. Syst.*, vol. 33, pp. 64–95, 2008, doi: 10.1016/j.is.2007.07.001.

[22] R. Andreswari and M. Rasyidi, "OLAP Cube Processing of Production Planning Real-life Event Log: A Case Study," Telkom University, 2017.

[23] J. De Weerdt, M. De Backer, J. Vanthienen, and B. Baesens, "A Multidimensional Quality Assessment of State-of-the-Art Process Discovery Algorithms Using Real-Life Event Logs," *Inf. Syst.*, vol. 37, no. 7, pp. 654–676, Nov. 2012, doi: 10.1016/j.is.2012.02.004.