Decision tree analysis approach to determine factors that affect the quote order lead time fulfillment

Nabil Gangsarwijaya, Rischa Anindhita, Dyah Lestari Widaningrum*
Industrial Engineering Department, Faculty of Engineering, Bina Nusantara University, Jakarta, Indonesia 11480

Email: dwidaningrum@binus.edu

Abstract. Problem identification is a critical step to resolve the delay problem in many industries. Fishbone Diagram is one method that is widely used to explore various alternative causes of problems. Other approaches are still needed to define the level of significance of the relationship between each alternative cause of the problem with the perceived problem. The Decision Tree with C5.0 and J48 algorithm approaches are used to classify enormous amounts of data. This study used these algorithms to obtain the root cause of delay. The calculation of information gain shows that most influence variables for the classification process of delay are warehouse capacity and quote order lead time.

Keywords: Order Fulfillment, Lead Time, Data Mining, Decision Tree Analysis

1. Background
Countries in Asia have the highest paper consumption in 2014 [1], including Indonesia. The industrial growth that occurred in Indonesia has a relative chain to the demand and growth of the paper industry. The growth of several industry groups, one of which is the paper industry, influenced a high industrial growth that occurred in the third quarter of 2018. Paper and paper products industry gained a growth of 4.80% (YoY), caused by the increase of large and medium industries [2].

Companies generally make various efforts to improve the quality of products and services to meet customer needs and expectations [3]. Lots of studies have examined various factors that can affect the level of customer satisfaction. Among them is the exactness of order fulfillment, both in terms of quantity, quality, and time [4, 5]. This research will focus on the issue of order fulfillment according to the quote order lead time. Many companies still have problems regarding delays in fulfilling orders that often occur due to issues in the production process, inventory management, and shipping processes. It is not uncommon for delays to occur due to interrelated problems between departments. Therefore, it is indispensable to discover the root of the problem and gain an optimal solution so that quality can be improved.

Data mining is now widely used in the practice of Quality Improvement in manufacturing, due to the complexity of the process and the availability and development of the system of data collection [6]. The data mining approach is frequently used to obtain predictions and visualize correlations [7], as well as to gain knowledge from data, determine the primary variables, classify records, and so forth [8]. This study aims to present the data mining approach to obtain specific issues that cause delays in order fulfillment, which can reduce the level of customer satisfaction. Determination of the main variables that influence the delay is significant as a basis to define the root cause of the problem.
decision tree technique is selected to achieve this aim. The structure of this paper includes related literature in section 2, detailed data and methods used in section 3, discussion of the results in section 4, and review conclusions in section 5.

2. Data Mining Methods for Lead Time Analysis
Zaki & Meira Jr. [9] stated the data mining definition that “Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data.” Various problems have applied data mining techniques in various types of industries, including the manufacturing industry. Production management is currently developing various systems to produce and manage data that can be analyzed into knowledge and models to improve manufacturing process performance in various problems such as defect prevention, defect detection, quality improvement, processing time reduction, and so forth [10]. Some cases include the data mining applied to develop a framework for continuous improvement program in energy consumption reduction in manufacturing industry [11]; a neural-based decision tree to predict the performance of the petroleum production process [12]; predict the pattern of manufacturing processes based on production data [13].

Companies must take consideration to define lead-time to quote customers, not only so that customers are interested in the promised timeline, but also to ensure the company's ability to fulfill orders as promised [14]. Fang et al. [15] present recommendations in implementing improvement in supply chain operations on three parameters, namely lead time average, the variance of lead time, and variance of demand [15].

3. Research Methods
There were several stages in this study, which included problem definition, data exploration, data preparation, modeling, evaluation, and deployment [16]. Customer order historical data is one of the necessary information and is very important in finding variables that cause delays in cargo ready. Historical data consists of all information on each order number, and information obtained is Estimated Shipment Date (ESD), paperweight, material number, and lead time. The quote order lead time fulfillment calculated from the time the company receives the order until the item is ready to be sent and gram per square meter (GSM) of paper. Data mining concepts using the C5.0 (R software) and J48 (WEKA) algorithms are used to classify enormous amounts of data.

4. Findings and Discussion
In a data-driven approach study, one of the most critical processes is identifying variables for the analysis process. This research conducted a process study to see the various possible conditions that would affect order fulfillment [4]. The results of the study show a variety of conditions that can affect the fulfillment of the quote order lead time, as can be seen in Figure 1.

![Figure 1. Factors considered affecting the order fulfillment delays](image_url)
Fishbone Diagram is one method that is widely used to explore various alternative causes of problems. Other approaches are still needed to define the level of significance of the relationship between each alternative cause of the problem with the perceived problem. Warehouse capacity is one of the factors that affect the fulfillment of the quote order lead time. This condition conduces a lack of storage layout so that operators cannot work optimally and lead to non-value-added jobs. One of the causes of the ready cargo delay is the length of picking time. Lack of the number of work teams has an impact on the time needed by workers to complete the picking process. Lack of workers will cause workers to get tired faster. Customer requests can also affect the delay in cargo ready. Variation of items requested by the customer and order quantity are classified into external factors. The variation of items requested also has an impact on the production lead time. The lead time difference in each item is an obstacle in estimating cargo ready. The delay also occurs due to the addition of a production batch, to meet the demand quantity of items that are always reduced in each batch due to the defect of each production.

Fishbone diagram shows the primary factors from the brainstorming processed, namely items ordered by the customer, order quantity (tonnage), number of the worker, the quote order lead time, and total warehouse capacity in a month. Data regarding those variables are structured for the decision tree analyzing process using C5.0 and J48 algorithm to define the root cause of the problem. This study will compare the result of the decision tree using two different algorithms, which are the C5.0 algorithm and J48 algorithm to obtain the best result. There will need several data to run the algorithm, such as historical data and capacity of the warehouse based on the results of determining variables using fishbone diagrams.

Figure 2 shows 221 instances and six attributes that will be processed. Instances indicate the amount of data contained in the file, while the attribute indicates the variable that causes tardiness in cargo ready. One of these attributes becomes the classifier or target attribute. Each attribute can be visualized into a histogram, as can be seen in Figure 2. The histogram in the blue shading area is classified as "late" class, while the red shading area is classified as "on-time" class.

This study used two decision tree classification models, namely the J48 and C5.0 algorithms. J48 is another name for the C4.5 algorithm, which is the development of the ID3 algorithm [17], while C5.0 is the development of the C4.5 algorithm by using different measurements as a classifier determinant. This study used splitting techniques for the classification process. Two scenarios for this splitting process are determined to obtain a comparison, specifically 90:10 and 55:45 splitting (training:testing).
According to the confusion matrix in Table 1, the accuracy value calculated for each algorithm as well as the splitting scenarios that have been conducted. The accuracy value for the 90:10 splitting technique for both algorithms has a good accuracy value, 91.3% for C5.0 Algorithm, and 86.4% for J48 Algorithm. Meanwhile, the accuracy value for the 55:45 splitting technique is lower for both algorithms, 74% for the C5.0 Algorithm and 71.7% for the J48 Algorithm. Figure 3 shows the accuracy of the model for both algorithms for each splitting technique. C5.0 and J48 algorithms have different decision tree results. The accuracy value, the precision value (called PPV – Positive Prediction Value), and Cohen’s Kappa coefficient of each algorithm can be the parameters to determine the best decision tree to choose. These parameters computed from confusion matrix data. Table 2 shows that the C5.0 algorithm with 90% splitting is the best-classified result according to these three parameters. This classifier shows that the concern for the main problems of delay is the variable warehouse capacity and lead time. The subsequent analysis is based on computational results using the decision tree C5.0 algorithm which has the best accuracy value, as can be seen in figure 4.

Table 1. Confusion Matrix

| Algorithm | Splitting | 90 : 10 | 55 : 45 |
|-----------|-----------|---------|---------|
|           | Late      | On-time | Late    | On-time |
| J48       | 9         | 2       | 33      | 13      |
|           | 1         | 10      | 15      | 38      |
| C5.0      | 9         | 0       | 32      | 16      |
|           | 2         | 12      | 10      | 42      |

Figure 3. Accuracy of the model

Table 2. Comparison of Accuracy, Precision (Positive Prediction Value), and Cohen’s Kappa

| Algorithm/Splitting | Accuracy | PPV   | Cohen’s Kappa |
|---------------------|----------|-------|---------------|
| C5.0 (90:10)        | 0.913    | 1     | 0.8244        |
| C5.0 (55:45)        | 0.74     | 0.8077| 0.4767        |
| J48 (90:10)         | 0.8636   | 0.9   | 0.7273        |
| J48 (55:45)         | 0.7172   | 0.688 | 0.4331        |
The classifier classifies variables that are judged to have an influence on the occurrence of delay, which makes the quote order lead time unfulfilled. The pattern of delays shown in Figure 4 becomes a reference for determining improvement systems. Entropy calculation for each attribute and overall is done to obtain information on gain values. The information gain values obtained form the basis for determining variables that significantly influence the classification process. Attributes with the most significant gain values mean that these attributes significantly influence the classification process, in this case the occurrence of delays. Table 3 shows the result of the capacity attribute entropy calculation.

\[
\text{Entropy}(S) = -\frac{103}{221} \log_2 103 - \frac{118}{221} \log_2 118
\]

\[
\text{Entropy}(S) = -\frac{103}{221} (-1.1014) - \frac{118}{221} (-0.9053)
\]

\[
\text{Entropy}(S) = 0.9967
\]

| Capacity | + (LATE) | - (ON TIME) | SUM | Entropy       |
|----------|----------|-------------|-----|--------------|
| 72       | 3        | 12          | 15  | 0.721928095  |
| 77       | 6        | 21          | 27  | 0.764204507  |
| 84       | 8        | 9           | 17  | 0.997502546  |
| 92       | 12       | 20          | 32  | 0.954434003  |
| 94       | 5        | 16          | 21  | 0.791858353  |
| 100      | 11       | 8           | 19  | 0.981940787  |
| 105      | 14       | 16          | 30  | 0.996791632  |
| 106      | 23       | 9           | 32  | 0.857148437  |
| 107      | 21       | 7           | 28  | 0.811278124  |
Based on entropy calculations, information gain calculations can be performed to determine the position and role of each variable. A large gain value describes the significant role of the attribute. Examples of calculating the gain from the warehouse capacity attribute are shown below. Table 4 shows the gain values for all attributes, which are the result of entropy data calculation.

\[
\text{Gain}(S, \text{Capacity}) = 0.9967 \times \left( \frac{15}{221} \times 0.722 \right) - \left( \frac{27}{221} \times 0.764 \right) - \left( \frac{17}{221} \times 0.998 \right) - \left( \frac{32}{221} \times 0.954 \right) - \left( \frac{21}{221} \times 0.792 \right) - \left( \frac{19}{221} \times 0.982 \right) - \left( \frac{32}{221} \times 0.857 \right) - \left( \frac{28}{221} \times 0.811 \right)
\]

Gain(S, Capacity) = 0.1175

| Variable                    | Gain          |
|-----------------------------|---------------|
| Item                        | 0.059479798   |
| Quantity                    | 0.029959889   |
| Quote Order Lead Time       | 0.094113453   |
| Capacity                    | 0.117506653   |
| Worker                      | 0.013558954   |

According to the calculation results of information gain for all variables, this study determined that the most significant variables which influence the classification process of the unfulfillment of quote order lead time or the delays are the capacity and quote order lead time.

5. Conclusion
The decision tree can help decision-makers to determine the focused problem to design the system improvement program. This study using the decision tree model, as one of the data mining approaches to classifying the pattern between the problem (delay or unfulfillment of quote order lead time) and the causes. C5.0 algorithm with 90:10 splitting proves as the best classifier, with the highest values of the parameters (accuracy value, precision value, and Cohen's kappa value). This classifier is execute using R software, while the J48 algorithm is executed using WEKA software. The development of various open-access programs, such as R software and WEKA makes it easier to use this method for practitioners in many industries.

The results obtained from the classification process is a decision tree to find out the root cause of delay, namely warehouse capacity, and quote order lead time. One solution that can be provided is by calculating the min-max inventory value along with the quantity of production to overcome the problem of warehouse capacity. Further research needs to conduct to design a new formulation to determine an accurate lead time quote orders to increase customer satisfaction by getting the product purchased according to the promised lead time.

References
[1] Environmental Paper Network 2018 The State of The Global Paper Industry
[2] Kementerian Perindustrian Republik Indonesia 2018 Analisis Perkembangan Industri (Jakarta)
[3] Kotler P and Keller K L 2012 Marketing Management (New York: Prentice-Hall)
[4] Croxton K L 2003 The Order Fulfillment Process Int. J. Logist. Manag. 14 19–32
[5] Slotnick S A 2014 Lead-time quotation when customers are sensitive to reputation Int. J. Prod. Res. 52 713–26
[6] Köksal G, Batmaz I and Testik M C 2011 A review of data mining applications for quality
improvement in manufacturing industry Expert Syst. Appl. 38 13448–67

[7] Peral J, Maté A and Marco M 2017 Application of Data Mining techniques to identify relevant Key Performance Indicators Comput. Stand. Interfaces 54 76–85

[8] Linoff G S and Berry M J A 2011 Data Mining Techniques - For Marketing, Sales, and Customer Relationship Management (Indianapolis, Indiana: Wiley Publishing, Inc)

[9] Zaki M J and Meira Jr. W 2014 Data Mining and Analysis - Fundamental Concepts and Algorithms (New York: Cambridge University Press)

[10] Cheng Y, Chen K, Sun H, Zhang Y and Tao F 2018 Data and knowledge mining with big data towards smart production J. Ind. Inf. Integr. 9 1–13

[11] Zhang Y, Ma S, Yang H, Lv J and Liu Y 2018 A big data driven analytical framework for energy-intensive manufacturing industries J. Clean. Prod. 197 57–72

[12] Li X and Chan C W 2010 Application of an enhanced decision tree learning approach for prediction of petroleum production Eng. Appl. Artif. Intell. 23 102–9

[13] Vazan P, Janikova D, Tanuska P, Kebisek M and Cervenanska Z 2017 Using data mining methods for manufacturing process control IFAC-PapersOnLine 50 6178–83

[14] Nguyen T H and Wright M 2015 Capacity and lead-time management when demand for service is seasonal and lead-time sensitive Eur. J. Oper. Res. 247 588–95

[15] Fang X, Zhang C, Robb D J and Blackburn J D 2013 Decision support for lead time and demand variability reduction Omega (United Kingdom) 41 390–6

[16] Suagra B, Widyatmoko D, Prakoso B S and Saputro D M 2018 Penerapan Algoritma C5.0 Untuk Deteksi Dini Gangguan Autisme Pada Anak Seminar Nasional Teknologi Informasi dan Komunikasi 2018 (SENTEKA 2018) pp 87–96

[17] Sulaiman A, Akinbowale B, Ronke B and Moshood H 2015 Comparative Analysis of Decision Tree Algorithms for Predicting Undergraduate Students’ Performance in Computer Programming Publ. Fac. Sci. Adeleke Univ. Ede, Niger. Comp. 2 79–92