C4.5 Classification Algorithm Based On Particle Swarm Optimization To Determine The Delay Order Production Pattern

H Setiawan¹, K Mukiman², Satria³, S Hanadwiputra⁴, A Suwarno⁵
¹,²,³,⁴,⁵ STMIK Bani Saleh, Bekasi, Indonesia

Abstract. Manufacturing companies engaged in the production of cosmetic products and herbal products on a national and international scale, this company is required to be able to carry out production planning by using classification methods in determining the pattern of delay in production orders for the C4.5 classification algorithm based on Particle Swarm Optimization. The research method used in this experiment is the Cross Industry Standard Model for Data Mining (CRISP-DM). From the results of the C4.5 algorithm optimization analysis with particle swarm optimization (PSO), it can be concluded that the accuracy value obtained from the PSO-based C4.5 algorithm model is 82.52%. This is better than the C4.5 algorithm model which produces an accuracy value of 80.17%. The difference of 0.033 with the details of the C4.5 Algorithm producing an AUC value of 0.855 with a diagnosis of Good Classification. The results of the optimization analysis of the C4.5 algorithm with PSO, the pattern of delay that is formed is that the production status is the status that causes the most delay in production orders. then followed by the status of waiting for pre-production and administrative status. the emerging status must be a correction and attention so that management becomes better.

1. Introduction
Manufacturing company engaged in the production of cosmetic products and herbal products on a national and international scale. To achieve success the company is required to be able to plan production well so that it can increase trust in consumers. If consumer trust can be done properly, it can be ascertained that the products made will be used by them and to guarantee the consumers’ needs for products produced by the company, the company needs to control existing production to be ready to respond to customer needs at any time and on time. In 2015, this manufacturing company experienced a delay of around 32 days in production planning. Planned planning is always outside the predetermined or often late production schedule (leadtime) so that product delivery to consumers is always late. The availability of prepared materials is not sufficient for the required production, payment of advances that have not been paid by customers, the formulation of production processes that still use the old method, and the readiness process of Raw Material Packaging Materials (RMPM) for the production process. Some of the above cause delays. These problems can be overcome, they need to analyze the delay. Analysis of delay is used to determine the cause of delays in production orders and through the results of the delay analysis, it can be seen whether that can cause production orders in this company to experience delays and can find out which parts cause delays in production orders. Analysis of delays is important in the scope of risk in production orders, therefore the need for analysis. Large data and number of parameters certainly need an effective and efficient tool to analyze delays and be able to find out the causes of late production. Measuring devices that can be used as
delay analysis data are production history data that has been running in this company. We analyze all order processes in each part that has a role in determining the production process. Every part that we will look for is the cause of the delay in production orders. Some of the problems mentioned above, in this study use the decision tree model C4.5 algorithm to determine the process of delaying production orders. To help process numerical data and choose good attributes, the Particle Swarm Optimization (PSO) algorithm will be used to select the attributes used and give the weight of the root calculation to be simpler so that it can form a reliable model to determine the pattern of delays in production orders.

From the explanation above, it can be formulated that How to determine the accuracy between the C4.5 and C4.5 algorithms based on Particle Swarm Optimization (PSO) in determining the pattern of delay in production orders with the aim of applying the Particle Swarm Optimization (PSO) optimization algorithm with weighting attributes to improve accuracy of the C4.5 classification algorithm in obtaining a pattern of late production orders.

One of the most popular classification techniques used in data mining is classification and decision trees. Decision trees are used to predict membership of objects for different categories (classes), taking into account the values that match their attributes or predictor variables [1]. C4.5 algorithm or decision tree resembles a tree where there are internal nodes (not leaves) that describe attributes, each branch describes the results of the attributes being tested, and each leaf describes the class. Decision trees can easily be converted to classification rules. There are several previous studies that used a linear regression classification model, including:

1. Integrating OpenStreetMap crowdsourced data and Landsat time-series imagery for rapid land use/land cover (LULC) mapping: Case study of the Laguna de Bay area of the Philippines
   This paper aims at of for LULC classification the noise level is high, because Landsat images all contain various levels of cloud cover (cause of noise attributes) and OSM polygons contain location errors and class labeling errors (causes of class noise)).[2]

2. Fault diagnosis of chemical processes with incomplete observations: A comparative study
   This paper aims at of for estimating errors with incomplete observations. This work discusses various approaches to dealing with lost data, and the performance of data-based error diagnosis schemes. Exploiting classifications and combined methods are assessed in the Tennessee-Eastman process, which results in a variety of incomplete observations.[3]

3. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data
   This paper aims at of To improve the accuracy of housing price predictions, this paper analyzes housing data of 5359 townhouses in Fairfax County, Virginia, collected by Multiple Listing Service (MLS) from Metropolitan Information Information Systems (MRIS).[4]

4. Implementation of decision tree using C4.5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region)
   This paper aims at of a tool to support loan analysts in decision making of loan application.[5]

5. Credit analysis using data mining: application in the case of a credit union
   This paper aims at of this study was to develop a model to analyze the capacity of credit union members to complete their commitments, using decision trees C4.5 algorithm and artificial neural networks multilayer perceptron algorithms.[6]
2. Methods
Based on figure Figure 1, the research method used in this experiment is the Cross-Standard Industry for Data Mining (CRISP-DM) model which consists of 6 phases [2], namely:

1. Business Understanding
At this stage it can be said as a stage of understanding research, determining the purpose of research in the formulation of defining data mining problems. Based on production order data for the 2015 and 2016 periods which can indicate the travel status of production orders.

Status of production orders that always experience delays will cause a decrease in the level of satisfaction with the customer. In this study the pattern of the causes of order delay was done by using the C4.5 algorithm and determining the pattern used with the C4.5 classification algorithm based on Particle Swarm Optimization (PSO) by then determining the comparison between the two with the aim of increasing the accuracy of the calculation of the C4.5 classification algorithm.

2. Data Understanding
At this stage, data collection is carried out, analyzing data investigations (order status data) to further identify the data and search for initial knowledge and then evaluate the quality of the data. So that it can be seen the percentage of the group of data to be analyzed.

3. Data Preparation
The amount of data obtained for this study of transaction status, both problematic and non-problematic, was taken in the last 2 years, namely 2015 and 2016. To obtain quality data, several preprocessing techniques were used [9], namely:

Data validation, to identify and delete odd data (outliers / noise), inconsistent data, and incomplete data (missing value)

Data integration and transformation, to improve the accuracy and efficiency of the algorithm. The data used in this writing is categorical and continuous.

Data size reduction and discretization, to obtain data sets with fewer attributes and records but are informative. In the training data used in this study, attribute selection and data duplication were deleted using RapidMiner software.

4. Modeling
At this stage the training data is processed so that it will produce several rules and will form a decision tree. The model that will be used is two, namely the C4.5 classification algorithm and the C4.5 classification algorithm based on Particle Swarm Optimization (PSO).

5. Evaluation
In the evaluation phase, it is called the classification stage because at this stage testing will be determined for accuracy. The testing phase is to see the results of accuracy in the classification process of C4.5 Algorithm and classification of C4.5 Algorithm based on Particle Swarm Optimization and evaluation with ROC Curve.

6. Deployment
At the deployment stage, the application of the Particle Swarm Optimization-based C4.5 classification algorithm model is applied to determine the pattern of delay in product orders.
Figure 1. Method of Cross-Standard Industry for Data Mining.

3. Results and Discussion

a. Data collection

Data Understanding. Data obtained from an order at PT Cedefindo in the period of 2015 and 2016. The amount of data used is 1321 records, with 11 Predictor attributes, namely product type, DP repayment status, service model, customer contract, down payment, status administration, pre-production waiting status, production status, waiting post production status, and overdue. This data is data that shows customer orders for a year, of which there are data orders that experience delays. Then the preprocessing technique is performed using the Rapidminer application to produce a candidate split like Table 1:

| Candidate Split | Child Nodes | Remarks                          |
|-----------------|-------------|----------------------------------|
| 1               | type of product | Liquid
Dry
CDF          |
| 2               | Model service | Semi
Full Fill Pack               |
| 3               | DP Repayment Status | Discharge: if there is no DP
Get off
Not yet DP
Late DP
DP OK          |
| 4               | Customer Contract | Large cust: > 1 billion
> 1 billion
100 million to 1 billion
<100 million |
| 5               | Down Payment   | Non DP
DP                                    |
### Candidate Split | Child Nodes | Remarks
--- | --- | ---
6 | **Status Administration** (Date of order up to the date of calculation)
- **Administration**
  - > 16 days
  - <16 days
|  | > 16 Days: Late
<= 16 Days: Exactly |
7 | **Status Waiting** Pra produktion Calculation date until production date
- **Waiting**
  - > 1 day
  - <1 day
|  | > 1 Day: Late
<= 1 Day: Exactly |
8 | **Status Production** Production date until the date of PTN
- **Production**
  - > 11 days
  - <11 days
|  | > 11 Days: Late
<= 11 Days: Exactly |
9 | **Status Waiting post production** PTN date until the date the letter is sent
- **Waiting post production**
  - > 2 days
  - <= 2 days
|  | > 2 Days : Late
<= 2 Days : Exactly |
10 | **Overdue** SJ Date - send date
- **Overdue**
  - <= 15
  - > 15
|  | <= 15: Right on time
> 15: Not Timely |

b. Results of Testing the C4.5 Algorithm Model

At this stage the training data is processed so that it will produce several rules and will form a decision tree.

The following are the steps in the C4.5 classification algorithm model:

Calculate the number of cases with Current and Loss and Entropy conditions from all cases

\[
\text{Entropy}(\text{total}) = - \left( \frac{905}{3121} \log_2 \left( \frac{905}{3121} \right) + \frac{416}{3121} \log_2 \left( \frac{416}{3121} \right) \right) = 0.899
\]

Then calculate the Entropy value and gain on each attribute, as the example below calculates the entropy and gain values for the administration status attribute:

The number of records of the administration status exactly consists of 502 on time, and 176 not on time, while the administration status is late consisting of 403 on time, and 240 not on time. Then the entropy can be calculated as follows:

\[
\text{Entropy}(i) = - \sum_{j=1}^{m} f(i,j) \cdot \log_2 f[(i,j)]
\]

\[
\text{Tepat} = \left( \frac{502}{672} \cdot \log_2 \left( \frac{502}{672} \right) + \frac{176}{672} \cdot \log_2 \left( \frac{176}{672} \right) \right)
\]

\[
\text{Telat} = \left( \frac{403}{645} \cdot \log_2 \left( \frac{403}{645} \right) + \frac{240}{645} \cdot \log_2 \left( \frac{240}{645} \right) \right)
\]

\[
E_{\text{split tepat}}(T) = \frac{672}{3121} \cdot 0.321 + \frac{645}{3121} \cdot 0.422 = 0.422
\]

\[
\text{Gain} \ OTR = 0.899 - 0.370 = 0.528
\]
The results of the calculation of entropy and gain can be seen in Table 2

| CONCLUSION                      | ENTROPY | GAIN       |
|---------------------------------|---------|------------|
| Number of Cases                 | 0.899   |            |
| Candidate Split                 |         |            |
| Types Of Products               |         |            |
| LIQUID                          | 0.880   | 0.0049784931 |
| DRY                             | 0.949   |            |
| CDF                             | 0.000   |            |
| Model Service                   |         |            |
| Full                            | 0.204   | 0.5390723296 |
| semi                            | 0.442   |            |
| Fill.Pack                       | 0.000   |            |
| Fill & Pack                     | 0.000   |            |
| Customer Contract               |         |            |
| BIG                             | 0.298   | 0.5316947723 |
| ONLY                            | 0.441   |            |
| SMALL                           | 0.301   |            |
| DownPayment Repayment status    |         |            |
| Yes                             | 0.399   | 0.5257195792 |
| No                              | 0.351   |            |
| Status Administration           |         |            |
| Right                           | 0.321   | 0.5283593364 |
| Late                            | 0.422   |            |
| Status Waiting Pre Production   |         |            |
| Right                           | 0.157   | 0.5456342529 |
| Late                            | 0.454   |            |
| Status Production               |         |            |
| Right                           | 0.228   | 0.5900936117 |
| Late                            | 0.524   |            |
| Status Waiting Post Production  |         |            |
| Right                           | 0.306   | 0.5399661639 |
| Late                            | 0.502   |            |

The data above shows the production status attribute by having the highest gain value of 0.5900936117, so that the production status attribute will be the main root of the model. Calculate
entropy and gain until the last root formation. Test results with K-Fold Cross Validation C4.5 Algorithm (See in Figure 2).

![Figure 2. Testing K-Fold Cross Validation C4.5 Algorithm](image)

Calculation of the value of accuracy is done by using the rapidminer application. The test results using the C4.5 algorithm are shown in Table 3.

The result of the model testing that has been done is to measure the level of accuracy and AUC (Area Under Curve).

1) Confusion Matrix

True Positive (TP) amount is 770 records classified as appropriate and False Negative (FN) as many as 127 records are classified as precise but not correct (intolerant). The next 289 records for True Negative (TN) are classified as incorrect, and 135 False Positive (FP) records classified as incorrect are correct.

|                | True No accurate | True accurate | class precision |
|----------------|------------------|---------------|-----------------|
| pred. unjustified | 289              | 135           | 68.16%          |
| pred. right     | 127              | 770           | 85.84%          |
| class recall    | 69.47%           | 85.08%        |                 |

Based on Table 3 it shows that, the level of accuracy using the C4.5 algorithm is 90.99%, and can be calculated to find the value of accuracy, sensitivity, specificity, ppv, and npv in the equation below:

\[
acc = \frac{tp + tn}{tp + tn + fp + fn} \\
acc = \frac{770 + 289}{770 + 289 + 135 + 127}
\]
Conclusion the results of the calculation of the above equation are shown in Table 4 below:

**Table 4. Results of calculation of the C4.5 algorithm**

| Value (%)          |
|--------------------|
| Accuracy           | 80.17            |
| Sensitivity        | 85.84            |
| Specitivity        | 68.16            |
| PPV                | 85.08            |
| NPV                | 69.47            |

2) ROC Curve
In figure 3 shows the ROC graph with an AUC (Area Under Curve) value of 0.822 with a Good Classification diagnostic level

![Figure 3. The AUC value in the ROC Graph C4.5 algorithm](image)

C. Test Results of C4.5 Algorithm based on PSO
In determining the results of the delay pattern using the PSO-based C4.5 Algorithm on RapidMiner as follows in Figure 4.
Figure 4. Testing of K-Fold Cross Validation C4.5 Algorithm Based on PSO

The results of the tests by using the PSO-based C4.5 algorithm are shown in Table 5. The result of the model testing that has been done is to measure the level of accuracy and AUC (Area Under Curve).

1) Confusion Matrix
True Positive (TP) amount is 804 records classified as appropriate and False Negative (FN) as many as 130 records are classified as exact but Not Right. The next 286 records for True Negative (TN) are classified as Improper, and 101 False Positive (FP) records classified as incorrect are correct.

|            | true No accurate | true accurate | class precision |
|------------|------------------|---------------|-----------------|
| pred. tidak tepat | 286              | 101           | 100.00%         |
| pred. tepat    | 130              | 804           | 87.08%          |
| class recall   | 68.75%           | 88.84%        |                 |

Based on Table 5 shows that, the level of accuracy using the C4.5 algorithm is 90.99%, and can be calculated to find the value of accuracy, sensitivity, specificity, ppv, and npv in the equation below:

\[
\text{accuracy} = \frac{tp + tn}{tp + tn + fp + fn} = \frac{804 + 286}{804 + 286 + 101 + 130}
\]

\[
\text{sensitivity} = \frac{tp}{tp + fn} = \frac{804}{804 + 130}
\]

\[
\text{specificity} = \frac{tn}{tn + fp} = \frac{286}{286 + 101}
\]
Conclusion The results of the calculation of the above equation are shown in Table 6 below:

Table 6. Results of calculation of the C4.5 PSO-based algorithm

| Value ( %) | Value          |
|-----------|----------------|
| Accuracy  | 82.52          |
| Sensitivity | 86.08        |
| Specificity | 73.90        |
| PPV       | 88.84          |
| NPV       | 68.75          |

2) ROC Curve
In figure 20 shows a ROC graph with an AUC (Area Under Curve) value of 0.855 with a Good Classification diagnostic level (See in Figure 5).

![Figure 5. AUC value in the ROC Graph C4.5 based PSO algorithm](image)

The results of the decision tree C4.5 based PSO classification algorithm with K-fold cross Validation testing are as follows in Figure 6.

![Figure 6. Decision Tree of the Late Classification C4.5 algorithm based on PSO](image)

D. Research Implications
From the results of the evaluation carried out above, both confusion matrix and ROC curve showed that the C4.5 classification algorithm based on Particle Swarm Optimization has a higher accuracy
value than just using the C4.5 classification algorithm. The accuracy value for the C4.5 classification algorithm is 80.17% and the accuracy value of the PSO-based C4.5 classification algorithm (Particle Swarm Optimization) is 82.52% with an accuracy difference of 2.35%, can be seen in Table 7 below.

### Table 7. Testing of the C4.5 classification algorithm and C4.5 based on PSO

|          | Accuracy | AUC  |
|----------|----------|------|
| C4.5     | 80.17%   | 0.822|
| C4.5 PSO based | 82.52% | 0.855|

Based on table 7 above, it can be analyzed that the PSO-based C4.5 algorithm has an accurate value better in determining the pattern of delay in production orders with the highest accuracy value of 82.52% and has a high AUC test value of 0.855 (Good Classification).

From the results of the C4.5 algorithm optimization analysis with particle swarm optimization (PSO), it can be concluded that the accuracy value obtained from the PSO-based C4.5 algorithm model is 82.52%. This is better than the C4.5 algorithm model which produces an accuracy value of 80.17%. From this result, the difference between the two models is 2.35%. While evaluating using the ROC curve, the two models have a difference of 0.033 with the details of the C4.5 algorithm model which produces an AUC value of 0.822 with the level of Good Classification and C4.5 based PSO algorithm that produces AUC value of 0.855 with a Good Classification diagnosis. So it can be concluded that the application of swarm optimization particle optimization techniques can increase the value of accuracy in the C4.5 algorithm.

## 4. Conclusion

The results of C4.5 can be concluded that the pattern of delays that are formed is the status and status of production followed by the status of waiting for pre-production and administrative status. From the results of the optimization analysis of C4 algorithm with particle swarm optimization (PSO), it can be concluded that the pattern is the status that causes the most delay in production orders. Also followed by pre-production pending status and administrative status. the three statuses that cause delays in production orders, it should be considered in teaching

### References

[1] Gorunescu, Florin. (2011). *Data Mining: Concepts, Models and Techniques*. Chennai: Springer.

[2] Johnson, Brian A.Iizuka, Kotaro. (2016). Integrating OpenStreetMap crowdsourced data and Landsat time-series imagery for rapid land use/land cover (LULC) mapping: Case study of the Laguna de Bay area of the Philippines. *Applied Geography*

[3] M. Escudero, Askarian, G. Graells, M. Zarghami, R. Jalali-Farahani, F. Mostoufi, N. (2016). *Fault diagnosis of chemical processes with incomplete observations: A comparative study*. Computers and Chemical Engineering

[4] Park, ByeonghwaKwon Bae, Jae.(2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications.*

[5] Rafik, Amin, Khairul, Indwiarti Sibaroni, Yuliant. (2015). *Implementation of decision tree using C4.5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region).* 2015 3rd International Conference on Information and Communication Technology, ICoICT 2015
[6] Sousa, Marcos de Moraes Figueiredo, Reginaldo Santana. (2014). *Credit analysis using data mining: application in the case of a credit union*. Journal of Information Systems and Technology Management.

[7] Sousa, Tiago, Arlindo Silva & Ana Neves. (2004). *Particle Swarm based Data Mining Algorithms for classification tasks*. Parallel Computing 30 (2004) 767–78.

[8] Turban, E., dkk, (2005), *Decision Support Systems and Intelligent Systems*. Yogyakarta : Andi Offset.

[9] Vercellis, C. (2009). *Business Intelligence Data Mining And Optimization For Decision Making*. United Kingdom: A John Wiley And Sons, Ltd., Publication.