Decision Tree to Predict the Color Quality of Refined Bleached Deodorized Palm Oil (RBPO)

Meilita Tryana Sembiring 1, Steven Chailes 2 and Sawaluddin 3
1,2,3 Teknik Industri, Fakultas Teknik, Universitas Sumatera Utara, Medan, Indonesia

Abstract. There are many factors that can affect the quality of Refined Bleached Deodorized Palm Oil produced. Every step in RBDPO processing must be mastered very precisely in terms of process conditions such as temperature, flowrate, dose, pressure, etc. This condition can vary from one facility to another and therefore requires the company's own specifications / rules to implement and integrate all production process facilities until the product is produced. The focus of the study on this particular task is the quality problems faced by companies in the Refinery section where research is carried out using data mining concepts to predict data logsheets that are written every 24 hours by refinery operators. The data can be used as a tool to predict the color quality of Refined Bleached Deodorized Palm Oil (RBPO). The use of Data Mining in this study is to facilitate operators in managing machine settings to achieve the quality required by the company directly. So that the Refinery section does not need to wait for the results of laboratory tests which require approximately 2 hours to find out the RBDPO quality results.

Keywords: Data Mining, Decision Tree, Refined Bleached Deodorized Palm Oil

1. Introduction
Identifying the occurrence of defects as early as possible in the production chain is very valuable, because it saves costs through the physical determination of production process facilities such as temperature, flowrate, dosage, pressure, etc. to get maximum results. According to Arnu's research, RapidMiner software can be used to get information from quality defects faster. The application of RapidMiner is very important in conducting time series analysis of the sensory data of each PT AAJ machine and predicting defects early in the RBDPO production process.

There are many parameters that affect the quality, such as the quality of CPO given, the condition of the engine process such as temperature, pressure on the Deodorizer machine, etc. A slight deviation from the parameters that affect the quality can cause a change in the color of the RBDPO. Identifying the occurrence of quality defects as early as possible in the production chain is very valuable, because it saves costs through remaining production steps for defective end products.

The focus of the study on this particular task is the quality problems faced by companies in the Refinery section where research is conducted using data mining concepts to predict data logsheets that are written every 24 hours by refinery operators. The data can be used as a tool to predict the color quality of Refined Bleached Deodorized Palm Oil (RBPO). The use of Data Mining in this study is to facilitate operators in managing machine settings to achieve the quality required by the company
directly. So that the Refinery section does not need to wait for the results of laboratory tests which require approximately 2 hours to determine the RBDPO quality results.

There are many parameters that affect the quality, such as the quality of CPO given, the condition of the engine process such as temperature, pressure on the Deodorizer machine, etc. A slight deviation from the parameters that affect the quality can cause a change in the color of the RBDPO. Identifying the occurrence of quality defects as early as possible in the production chain is very valuable, because it saves costs through remaining production steps for defective end products.

2. Method

2.1. Criteria of production target
In carrying out the production process, there has 4 criteria as determining the success of achieving the production target. The criteria used are as follows:

1. Quality
   Quality is the overall characteristics and characteristics of a product or service whose ability can satisfy needs, whether expressly or disguised (ISO 8402).

2. Capacity
   Capacity is the optimum level of production capability of a facility usually expressed as the amount of output in a given time period.

3. Production Costs
   Production costs are all costs associated with the product (goods) obtained, where there are elements of product costs in the form of raw material costs, direct labor costs, and factory overhead costs.

4. Losses
   Losses is the reduction in assets or economic resources of the company that is not due to taking the owner and for the reduction of assets or sources there is no benefit obtained by the company.

2.2. Data mining
Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

2.2.1. Data Cleaning.
Data cleaning is usually called data cleansing or scrubbing. Data cleaning process is carried out to eliminate information errors in the data [1]. So the data cleaning process can be used to determine inaccurate, incomplete or incorrect data and to improve data quality through detecting errors in the data. There are 4 criteria in calculating the quality of data used [2].

1. Missing: A lot of missing data is divided by lots of variable data
2. ID-Ness: Many different values divided by many variable data
3. Stability: Many values are the same (non-missing) divided by many variable data
4. Validity: Divided by values not included by missing, ID, stable.

2.2.2. Data Pre-processing.
The preprocessing phase is carried out as an initial and important stage in research [3]. The method used in this research is data transformation using normalization. Normalization is the process of scaling attribute values from data so that they can lie within a certain range.

\[ Z = \frac{X - \mu}{\sigma} \]
2.3. Correlation
Correlation is one of the statistical analyzes used to find the relationship between two quantitative variables [4]. Correlation analysis is a study of the degree of relationship or degree of association between two variables, for example the variable X and the variable Y. The more specific understanding of correlation, which implies a relationship that is substantive numeric (numbers / numbers). From this definition, it also shows that the purpose of correlation analysis is to see / determine how closely the relationship between two variables is.

\[
\text{Correlation}= r_{xy}= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}
\]

(1)

2.4. Decision Tree
Decision Tree is a prediction model using tree structure or hierarchical structure. The concept of a decision tree is to change data into a Decision Tree and decision rules. The main benefit of using the Decision Tree is its ability to formulate complex decision-making processes to be simpler, so that decision makers will interpret solutions to problems. Decision Tree is also useful for exploring data, finding hidden relationships between a number of potential input variables and a target variable [5]. Decision Tree combines data exploration and modeling, so it is very good as a first step in the modeling process even when used as the final model of several other techniques. In some applications, the accuracy of a classification or prediction is the only thing that is highlighted in this method.

\[
\text{Entropi} (s)= \sum_{j=1}^{k} P_j \log_2 P_j
\]

(2)

\[
\text{Gain} (S,A) = \text{Entropi} (s)- \sum_{i=1}^{n} \frac{|s_i|}{|s|} \times \text{Entropi} (S_i)
\]

(3)

2.5. Root mean square error
Root mean square error is an alternative method for evaluating forecasting techniques used to measure the accuracy of the results of a model forecast [6]. RMSE is the average value of the sum of the squares of error, it can also state the size of the error generated by a forecast model. A low RMSE value indicates that the variation in values produced by a forecast model approaches the variation in the value of obesity.

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}{n}}
\]

(4)

Absolute Error is the number of errors in your measurement. This is the difference between the predicted value and the actual value. Mean Absolute Error (MAE) represents the average absolute error between forecasting results and actual values [6]. Other variant of error percentage can be obtained using Mean Absolute Percentage Error (MAPE)[8], which can be considered in the future.

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|
\]

(5)

3. Results and discussion
3.1. Variable in data used
The focus of the study on this particular task is the quality problems faced by companies in the Refinery section where research is carried out using data mining concepts to predict data logsheets that are written every 24 hours by refinery operators. The data can be used as a tool to predict the color quality of Refined Bleached Deodorized Palm Oil (RBPO).

| No. | Variable                        |
|-----|---------------------------------|
| 1   | Storage Tank CPO No.            |
| 2   | Flow MT/H CPO                   |
| 3   | Flow MT/H BPO                   |
| 4   | Storage Tank RBDPO No.          |
| 5   | Storage Tank PFAD No.           |
The variables used in this study are as follows. Problem solving can be done using logsheet data that has been written every hour with 168 data (45 variables) recorded from July 1, 2019 to July 7, 2019. The written sheet will be moved into Excel software which is the source of input for certain algorithms.

3.2. Data cleaning
Data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with the most commonly occurring values for that attribute.

| Variable                                | Missing (%) | ID-NeSt (%) | Stability (%) | Valid (%) |
|-----------------------------------------|-------------|-------------|---------------|-----------|
| Storage Tank CPO No.                   | 1,800       | 1,800       | 41,460        | 59,940    |
| QTY MT Storage Tank CPO                | 100,000     |             |               |           |
| QTY MT Storage Tank RBDPO              | 100,000     |             |               |           |
| Storage Tank PFAD No.                  | 1,800       | 1,200       | 93,900        | 3,100     |
| QTY MT Storage Tank PFAD               | 100,000     |             |               |           |
| Level Tank Phosphoric Acid Min 100 L   | 1,800       | 38,920      | 5,490         | 53,790    |
| etc                                     | ...         | ...         | ...           | ...       |

At this stage there were 14 variables that were discarded because the data were completely absent and the data had high stability.

| No. | Variable                                | Parameter |
|-----|-----------------------------------------|-----------|
| 1   | QTY MT Storage Tank CPO                 | Missing 100% |
| 2   | QTY MT Storage Tank RBDPO               | Missing 100% |
| 3   | QTY MT Storage Tank PFAD                | Missing 100% |
| 4   | Setting Stroke Phosphoric Acid mm       | Stability 100% |
| 5   | B/E Dosage %                            | Stability 100% |
| 6   | Level HE 721                            | Stability 100% |
| 7   | Pressure PU 621                         | Stability 100% |
| 8   | Pressure PU 701                         | Stability 100% |
| etc | ...                                     | ...       |

So that data with high missing cannot be processed and data with high stability does not have any impact on the data.

3.3. Data preprocessing
The method used in this research is data transformation using normalization. Normalization is the process of scaling attribute values from data so that they can lie within a certain range.

| Storage Tank CPO No. | Flow MT/H CPO | Flow MT/H BPO | Storage Tank RBDPO No. |
|----------------------|---------------|---------------|-------------------------|
| 0.8                  | -0.9          | -0.9          | -0.5                    |
| 0.8                  | -0.9          | -0.9          | -0.5                    |
| -1.5                 | -0.9          | -0.9          | -0.5                    |
| -1.5                 | -0.9          | -0.9          | -0.5                    |
| -1.5                 | -0.9          | -0.9          | -0.5                    |
| -1.5                 | -0.9          | -0.9          | -0.5                    |
| etc                  | ...           | ...           | ...                     |

Data pre-processing may affect the way in which the outcomes of the final data processing can be interpreted. This aspect should be carefully considered when interpretation of the results is a key point.

3.4. Correlation
Correlation is a number between -1 and +1 that measures the degree of relationship between two Attributes (say X and Y) [7]. A positive value for correlation implies a positive relationship. In this case a large X value tends to be associated with a large Y value and a small X value tends to be associated with a small Y value. A negative value for correlation implies a negative or inverse relationship. In this case a large X value tends to be associated with a small Y value and vice versa.

**Table 5. Correlation Table**

| Variables      | RBDPO Red | CPO FFA | Flow MT/H BPO | Flow MT/H CPO |
|----------------|-----------|---------|---------------|---------------|
| RBDPO Red      | 1         |         |               |               |
| CPO FFA        | -0.20311  | 1       |               |               |
| Flow MT/H BPO  | 0.15057   | -0.150  | 1             |               |
| Flow MT/H CPO  | 0.15057   | -0.150  | 1             | 1             |
| etc.           | ...       | ...     | ...           | ...           |

Weight by Correlation calculates attribute weights with respect to label attributes using correlations [9]. The higher the weight of the attribute, the more relevant the variable is. Weight by Correlation can only be applied to ExampleSets with numeric or binominal labels.

**Table 6. Weight of correlation**

| Attribute                  | Weight |
|----------------------------|--------|
| Phosphoric Acid Dosage %   | 0.7593 |
| Temp. HE 311 IN            | 0.7080 |
| Frequency Hz               | 0.6553 |
| RBDPO Yellow               | 0.5479 |
| Temp. HE 721 OUT RPO       | 0.3896 |
| etc.                       | ....   |

From the calculation of the attribute weights it can be concluded that the Phosphoric Acid Dosage% has the most relevant influence on RBDPO Red. And the lowest attribute relevant to RBDPO Red is Temp. HE 722 OUT.

3.5. Decision tree

The concept of a decision tree is to change data into a Decision Tree and decision rules. The main benefit of using the Decision Tree is its ability to formulate complex decision-making processes to be simpler, so that decision makers will interpret solutions to problems. Here are the results of a comparison between the results of RBDPO Red Decision Tree predictions with the original data.

**Table 7. Prediction Results**

| No. | RBDPO Red | Prediction (RBDPO Red) |
|-----|-----------|------------------------|
| 1   | 2.5       | 2.5                    |
| 2   | 2.5       | 2.4                    |
| 3   | 2.5       | 2.4                    |
| 4   | 2.5       | 2.5                    |
| 5   | 2.4       | 2.2                    |
| Etc.| ...       | ...                    |

Performance Decision Tree:
1. root_mean_squared_error: 0.081 +/- 0.025 (micro average: 0.084 +/- 0.000)
2. absolute_error: 0.052 +/- 0.017 (micro average: 0.052 +/- 0.066)
3. correlation: 0.936 +/- 0.038 (micro average: 0.934)
4. Conclusions
Data mining is a collection of algorithms that is used by office, governments, and corporations to predict and establish trends with specific purposes in mind. The company has equipped a machine that has sensors measuring pressure, temperature, flow, tank level, stroke settings and so on so that available variations of data recorded periodically every hour in detail by the operator. Continuous processes like these tend to create sensory data which often requires a lot of cleaning and normalization before being able to effectively implement data mining techniques. This is what happens in most heterogeneous industries in their dimensions and physical properties.

In this research, we have presented a summary of data mining predictive analysis and results and focus on the decision tree. In addition, tracking the quality of the RBDPO becomes difficult because the product change is affected by many variables and the resulting color is difficult to examine visually. In this study the researchers also held discussions with the head of the company's laboratory to find out more clearly what factors influence the color quality of the Red RBDPO. Analysis of the problem at PT AAJ was carried out using data mining methods to process the sensory results of each machine that aims to predict the variables that affect the quality of the RBDPO, namely RBDPO Red. So it is likely possible to continuing predicting RBDPO quality result using decision tree.

References
[1] Van Den Broeck J, Cunningham S A, Eeckels R, and Herbst K, 2005 Data cleaning: Detecting, diagnosing, and editing data abnormalities PLoS Med., 2(10), pp. 0966–0970
[2] O’Donoghue J and Herbert J, 2012 Data management within mHealth environments: Patient sensors, mobile devices, and databases J. Data Inf. Qual., 4(1)
[3] Chamidah N, and Salamah U, 2016 Pengaruh Normalisasi Data pada Jaringan Syaraf Tiruan Backpropagasi Gradient Descent Adaptive Gain (BPGDAG) untuk Klasifikasi J. Teknol. Inf. ITSmart 1(1), p. 28
[4] Schober P and Schwarte L A, 2018 Correlation coefficients: Appropriate use and interpretation Anesth. Analg. 126(5), pp. 1763–1768
[5] Kakavand S, Mokfi T, and Tarokh M J, 2014 Prediction the Loyal Student Using Decision Tree Algorithms Int. J. Inf. Commun. Technol. Res 4(1), pp. 32–37
[6] Makridakis S et al., 1982 “The Accuracy of Extrapolation (1 ime Series ) Methods : Results
of a Forecasting Competition,” J. Forecast., 1, pp. 111–153

[7] Arnu D et al. 2017 A Reference Architecture for Quality Improvement in Steel Production,” in Data Science – Analytics and Applications, Springer Fachmedien Wiesbaden pp. 85–90.

[8] Rahmat R F, Nurmaryanto, Sembiring S, Syahputra M F and Fadli 2018 Adaptive neuro-fuzzy inference system for forecasting rubber milk production IOP Conference Series : Materials Science and Engineering 308(1), 012014

[9] Waller N G and Jones J A, 2010 Correlation weights in multiple regression Psychometrika, 75(1), pp. 58–69