Post-harvest Soybean Meal Loss in Transportation: A Data Mining Case Study

Emmanuel Jason Wijayanto, Siana Halim(✉), and I. Gede Agus Widyadana

Industrial Engineering Department, Petra Christian Universtiy, Jl. Siwalankerto 121-131, Surabaya, Indonesia
halim@petra.ac.id

Abstract. A poultry company in Indonesia has a problem, i.e., losing raw material, the so-called Soybean Meal (SBM), during transportation from the port to the factory. To reduce material loss, the company created a raw material transport (RMT) system, which recorded the time and activities during loading-unloading and transporting the material from the port to the factory warehouses. Therefore, this study aims to mine the data on the loss of raw materials through RMT. The application used is Orange data mining to find the relationship between lost material and other attributes, create clusters, and classify the standardized lost. The clustering exhibits two classes, namely, the standard and non-standard conditions. The classification process uses five different algorithms. The random forest algorithm was chosen because it produces the second-best AUC value and can produce a classification visualization through a decision tree. This classification process also produces rules based on the decision tree.

Keywords: Data mining · Clustering · Classification · Random forest

1 Introduction

The problem faced by a poultry company in Surabaya is losing raw material, so-called Soybean Meal (SBM), which was imported from e.g., Brazil and Argentina to Surabaya when it is transported from the port to the factory. There are two ports in Surabaya, Tanjung Perak, which is 34 km from the factory, and Teluk Lamong, is 45 km from the factory. Most of the SBM is shipped to Teluk Lamong. Three types of SBM imported by the company right now, e.g., SBM Argentine HiPro (50%), SBM Brazil Lopro (26.4%), and SBM Brazil HiPro (23.6%). The company occupies third-party logistics (TPL) to transport the SBM from the port to the factory. The TPL uses dump trucks as the SBM transporter; each dump truck has a capacity of 25 tons. The material is lost during transportation from the port to the factory. So, the company created a monitoring system called raw material transport (RMT) to reduce the lost material. The RMT recorded the vehicle's plate number and the weight of the dump truck in an empty condition before and after transporting the SBM, the weight of the loaded dumped truck scaled in the port, and when it arrived in the factory. Additionally, it also recorded each activity's times. The activities are departure time from the port, arrival time in the factory, queuing time for scaling and unloading in the factory, and unloading time in the warehouse. The scaling and unloaded process in the factory is hectic.

P. Vasant et al. (Eds.): ICO 2023, LNNS 853, pp. 316–324, 2024.
https://doi.org/10.1007/978-3-031-50327-6_33

Post-harvest losses during transportation can occur due to many factors, such as physical damage to the crops during loading and unloading the crops, temperature and humidity, contamination, pest infestation, and poor packaging [1]. Many researchers study this problem, e.g., Medeiros et al. [2], studying the post-harvest soybean loss during transportation in Brazil. In this study, Medeiros et al. recorded the loss statistics for the distance from the farm to the destination and the road condition. Wang and Shi [3] were researching the function optimization of bulk grain transportation. Iordăchescu et al. [4] investigated transportation and storage losses for Romania's fresh fruits and vegetables. In their finding, the losses due to transportation could happen because of long shipping and delivery times, transportation of unsuitable products together, product damage due to rough and rugged mechanical processing, and not transporting products in a suitable atmosphere. Jia et al. [5]. Provided a systematic literature review on supply chain management of soybeans. Machine learning, and data mining have also been used widely in agricultural problems. Borse and Agnihotri [6] used fuzzy logic rule to predict the crop yields, Vasilyev et al. [7], processing plants for post-harvest disinfection of grain. This study aims to investigate the loss of soybean meal in transportation through data mining.

2 Methods

2.1 RMT Process

The RMT flow starts with recording the loaded dump truck weight before leaving the port. The truck is identified by its vehicle plate number and travel document. The travel document number is barcoded and attached to the truck body. In every security post, the security will scan the barcode and record the time. Then the truck will travel from the port to the factory. Once it arrives in the factory, security will scan the barcode; the truck will enter the queuing line to weigh the loaded truck. The weighting queuing time is recorded in the RMT. After the weighing process, the truck will enter the queue to unload the cargo. The unloading queuing time is also recorded in the RMT. Then the truck will unload the cargo in the specific warehouse. Again, the RMT records the unloading process time. In the final state, the empty truck is weighed, the security scans the factory departure ticket, and the truck returns to the port.

2.2 Data Preprocessing

All previously collected data will be processed through merging, selecting, and transforming. First, the merging process is carried out to unify the two data types obtained to calculate the amount and percentage of difference. Next, the selection process is a process to remove unnecessary data attributes from the combined data so that the amount of data can be minimized but still represents the actual data. In the data cleaning, we cleaned the data that deviated significantly and will ultimately damage the data distribution (outliers). Finally, after the transformation, where the data will be changed to the selection, the data will undergo processing in another form to add the information needed in the mining process.

2.3 Data Mining

The next stage is the core of this research, which is to perform data mining from data that has been processed and cleaned. Data mining is a study to collect, clean, process, analyse, and obtain useful information from data [8]. The mining process will be carried out to look for attributes that may be related to the loss of this raw material through correlation tests and analysis of variance (ANOVA).

The clustering process is carried out using the k-Means algorithm to determine the grouping of losses that occur and can be used as a standard benchmark for a loss. In this method, the algorithm examines the data to find groups of similar items [9]. The clustering process carried out in this study uses the k-Means algorithm and produces two clusters, namely standard and non-standard conditions.

Then the classification process is carried out, dividing objects into each data into one of several categories commonly known as classes [9]. This classification process will be carried out using five algorithms: naive Bayes, KNN, tree, random forest and neural network. This classification process aims to make predictions from a condition that will enter a standard or non-standard group. This classification process can obtain a rule for companies to classify.

3 Results and Discussions

3.1 Data

This study uses two datasets recorded from January 2020 to August 2022. The first dataset was obtained from the RMT report data. It contains 28 attributes. It includes information about the dump trucks used for shipping, details of the time and name of the user involved in the occurrence of each flow stop, and the weight of raw materials transported at the port. There are seven files of this type of data, separated by year, period, and type of raw material transported. The second dataset is Factory's SAP Data, which internal factories use to update after raw materials are weighed at the factory location. Therefore, in this type of data, there are only 9 data attributes, including delivery truck information, time details, and weight of raw materials that are weighed at the factory. Those two datasets are preprocessed to obtain a clean dataset consisting of 27 attributes (14 are numerical data, 8 are categorical data, and five are in the meta category in the text form) and 12,027 rows of data.

3.2 Data Descriptive

This study aims to mine the SBM lost due to transportation from the port to the factory. Here, the loss is defined as the dump truck's weight difference when weighted in the port and the factory. Among previous studies, post-harvest loss happened due to distance; Time traveled, temperature and humidity [1, 2]. Therefore, in this study, we do the correlation analysis and hypothesis test to infer the loss. Furthermore, since the distance from the port to the factory is constant, we start to infer with Time traveled. The company did not record the temperature and humidity. The company did not record the temperature and humidity. However, the temperature is related to the time. The day and night temperature are different. At the same time, the humidity is related to the month.

3.2.1 Time Traveled

Time traveled records the length of travel time since the truck driver scans the barcode at the port until the driver scans the barcode again at the factory gate. This attribute was chosen for research because there was an initial assumption from the company that the Time travelled would be related to the SBM loss. Therefore, the company inferred that the longer the Time traveled, the more significant the SBM loss. The Time traveled has an average of 01:33:01. If the Time traveled is grouped every 10 min, the highest frequency occurred from 01:10:00 to 01:20:00, with a total of 2779 trips (23.11% of the total data). The fastest Time traveled is 00:58:40 (58 min 40 s), while the slowest is 03:31:07 (3 h 31 min 7 s). Correlations are performed to see the relationship between SBM loss and Time traveled. The resulting correlation coefficient between the Time traveled, and SBM loss is +0.034. This result is not significant and negligible [10]. Therefore, the Time traveled attribute is not related to the SBM loss.

3.2.2 Temperature

The temperature is related to the port departure time. Port departure time is a record when a truck starts leaving the port to deliver SBM to the factory. The departure time can be started from 00:00 to 23:59 and it is well known that temperature and win speed are different from time to time [11]. Most trucks (766 trips or 6.37%) depart from the port between 16.00 and 17.00, and the lowest (108 trips or 0.90%) occur between 07:00 and 08:00. The correlation between port departure time to the SBM loss in percentage is +0.012. Therefore, the correlation is not significant. In Surabaya, the day and night temperature are not extremely different and the uncorrelated condition between SBM loss and temperature is reasonable. Surabaya, the capital of East Java Province, is located on the northern coast of East Java Province and is a tropical city. It is between 7° 9′–7° 21′ South Latitude and 112° 36′–112° 54′ East Longitude. Topographically, Surabaya is 80% flatland, with a height of 3–6 m above sea level [12]. Additionally, the temperature in Surabaya does not differ daily. It is between 24 and 32 °C. Figure 1 shows the monthly average temperature in Surabaya [13].

3.2.3 Humidity

Indonesia has two seasons, dry and wet seasons. Officially, the dry season starts in April and ends in September; the wet season starts in October to March. However, the starting month of the dry or wet seasons is shifting. Therefore, in this study, we related the humidity to the month when the imported SBM was anchored in the port of Surabaya. Figure 2 shows the monthly average humidity in Surabaya. During wet season the humidity is higher than in the dry season. Additionally, the data set recorded that July is the most hectic month with 2,235 trips (18.58%) from port to the factory, while September is a bit of slack with only 78 trips (0.65%). However, the SBM loss in September is higher than in July. The loss tends to be higher in the wet seasons than in the dry seasons. The monthly average SBM loss in percentage is depicted in Table 1. The ANOVA test shows that the monthly average percentage SBM loss is significantly different (F-value 73.25, p-value 0.000).

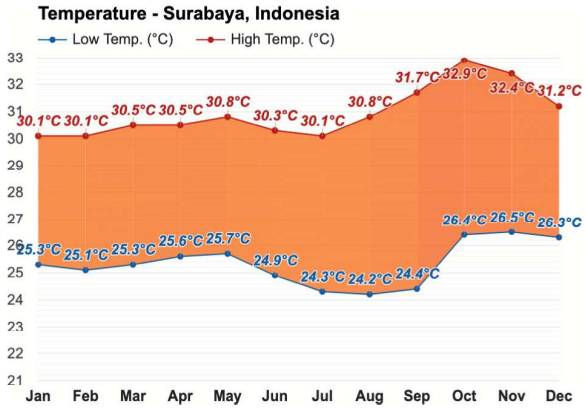

Fig. 1. Monthly average temperature in Surabaya [13]

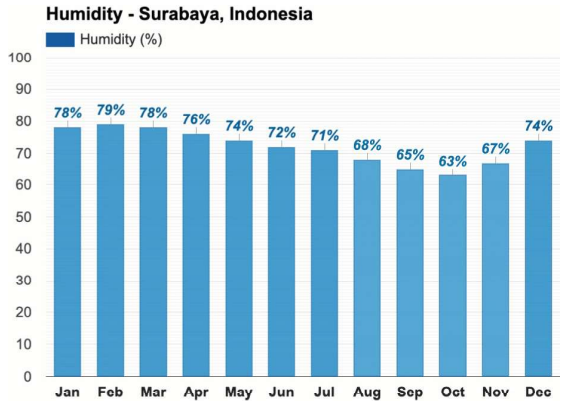

Fig. 2. Monthly average humidity in Surabaya [13].

Table 1. Monthly average SBM loss in percentage

Month	Average loss (%)	Month	Average loss (%)
January	−0.158	July	−0.155
February	−0.123	August	−0.184
March	−0.153	September	−0.175
April	−0.169	October	−0.141
May	−0.206	November	−0.264
June	−0.135	December	−0.219

3.3 Clustering and Classification

3.3.1 Clustering

The data descriptive only shows the relationship between one attribute to the SBM loss. Further, we analyze the data by clustering the SBM loss. The data analysis is carried out using Orange data mining [14]. We used K-means since it is simple and efficient. Moreover, it is commonly used to cluster features with many data types [15]. The data is standardized to reduce the noise in the dataset [16]. To find the number of clusters we used the Silhouette scores [17], and it exhibits the number of clusters is two (see Table 2). Figure 3 exhibits the distribution of the two clusters and the summary statistics can be seen in Table 3.

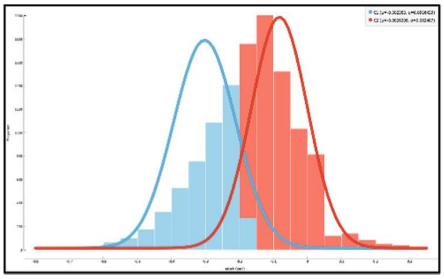

Fig. 3. Distribution of the two clusters

Table 2. Silhouette scores

Number of clusters	Silhouette scores
2	**0.557**
3	0.515
4	0.532
5	0.526
6	0.540
7	0.542
8	0.542

Table 3. Characteristics of Cluster 1 (C1) and Cluster 2 (C2)

Cluster	Mean (in %)	St. Dev (in %)	Mean (in kg)	Characteristic
C1	−0.302	0.091	−76.3	Non-standard
C2	−0.083	0.082	−20.7	Standard

3.3.2 Classification

The classification rules allow class predictions if several variables are known in the study [18]. In classifying, many methods can be used, but in this study, there are only four algorithms, namely KNN [19], Naive Bayes [20], Tree [21], Random Forest [22] and Neural Network [23]. In this study, overfitting the decision tree will be prevented by pre-pruning by limiting the depth of the decision tree to 5 levels. Max depth five was chosen because it increases the AUC value significantly compared to the depth below it (up to 4.7%), even though the AUC value is higher than depth 6. Testing the results of this classification uses the 80:20 system, where the algorithm will use 80% of the data to study data patterns and models (training data). The remaining 20% will be used for testing the algorithm (testing data).

Based on the assessment results in Table 4, there are five metrics parameters, but this study uses the AUC value as a reference for assessment. AUC, which stands for the area under the ROC curve, is a global index used to calculate the accuracy of the estimated area under the Receiver Operating Characteristic (ROC) Curve [24].

Table 4. Classification algorithm metrics

Model	AUC	CA	F1	Precision	Recall
kNN	0.672	0.654	0.645	0.645	0.654
Tree	0.652	0.668	0.681	0.681	0.668
Random forest (2)	0.673	0.662	0.685	0.685	0.662
Neural network	0.524	0.605	0.596	0.596	0.605
Naive Bayes	0.642	0.639	0.635	0.635	0.639

The classification algorithm with the highest AUC value is the kNN algorithm, followed by random forest, tree, and naive Bayes. Although not the best algorithm because it produces the second-highest AUC value, the random forest algorithm has its advantages: it can visualize the classification process into a decision tree. Figure 4 shows the decision tree generated by the random forest algorithm, where there are one root node, 22 internal nodes, and 24 leaf nodes consisting of 16 standard leaves (C2) and eight non-standard leaves (C1). Three features are significant in classifying the SBM loss as standard or non-standard, i.e., month, departure time, and duration.

Three rules that signify the SBM loss is classified as non-standard are:

Month = April and Port departure time < 09:26 and Time-traveled > 01:30.

Month = February and Port departure time > 22:54 and Time-traveled > 01:45.

Month = December and Time-traveled > 03: 21.

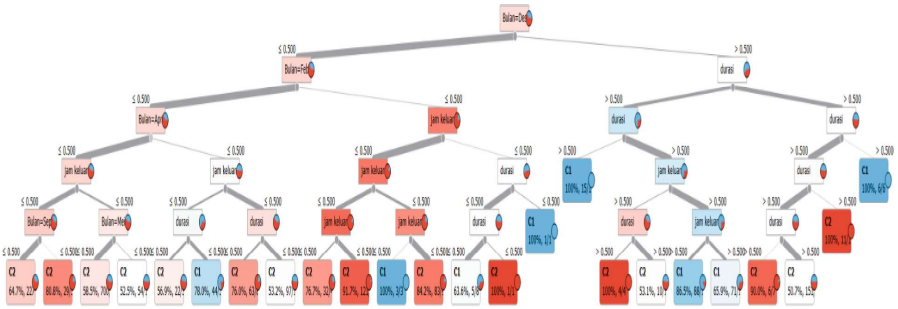

Fig. 4. Random forest decision tree

4 Conclusion

This study analyzes three data attributes related to SBM loss during transportation from the port to the factory: time traveled, port departure time, and month. Time-traveled and port departure time are not significantly correlated to the SBM loss, but the months significantly correlate with the SBM loss. Additionally, two clusters are discovered to classify the SBM loss as a standard loss with a mean percentage of loss −0.083% of total SBM delivered and a non-standard loss with a mean percentage of loss −0.302%. Finally, the random forest is used to predict whether particular features will cause the SBM loss to be standard or non-standard. It is found that the wet seasons will cause the SBM loss to be severed. The AUC is still under 70%. In future work, we need to elaborate on other variables that influence the soya bean material loss during transportation from the port to the factory.

References

1. Al-Dairi, M., Pathare, P.B, Al-Yahyai, R.: Mechaninal damage of fresh produce in postharvest transportation: current status and future prospect. Trends Food Sci. Technol. **124**, 195–207 (2022). https://doi.org/10.1016/j.tifs.2022.04.018
2. Medeiros, P.O., Naas, I., Vendrametto, O., Soares, M.: Post-harvest soybean loss during truck transport: a case study of Piaui State, Brazil. In: Advances in Production Management Systems, Initiative for a Sustainable World. APMS 2019. IFIP Advances in Information and Communication Technology, vol. 488 (2019). https://doi.org/10.1007/978-3-319-51133-7_72
3. Wang, X., Shi, H.: Research on the function optimization of the bulk grain transportation central control system. IOP Conf. Ser. Earth Environ. Sci. **512**(1), 012163 (2020). https://doi.org/10.1088/1755-1315/512/1/012163.(2020)
4. Iordăchescu, G., Ploscutanu, G., Pricop, E.M., Baston, O., Barna, O.: Postharvest Losses in transportation and storage for fresh fruits and vegetables sector. Agric. Food **7**, 244–249 (2019)
5. Jia, F., Peng, S., Green, J., Koh, L., Chen, X.: Soybean supply chain management and sustainability: a systematic literature review. J. Clean Prod. **255**, 120254 (2020). https://doi.org/10.1016/j.jclepro.2020.120254

6. Borse, K., Agnihotri, P.G.: Prediction of crop yields based on fuzzy rule-based system (FRBS) using the Takagi Sugeno-Kang approach. In: Vasant, P., Zelinka, I., Weber, G.W. (eds.) Intelligent Computing & Optimization. ICO 2018. Advances in Intelligent Systems and Computing, vol. 866, pp. 438–447. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-00979-3_46
7. Vasilyev, A.A., Samarin, G.N., Vasilyev, A.N.: Processing plants for post-harvest disinfection of grain. In: Vasant, P., Zelinka, I., Weber, GW. (eds) Intelligent Computing and Optimization. ICO 2019. Advances in Intelligent Systems and Computing, vol. 1072, pp. 501–505, Springer, Cham (2020). https://doi.org/10.1007/978-3-030-33585-4_49
8. Aggarwal, C.C.: Data Mining: The Textbook. Springer International Publishing (2015)
9. Bramer, M.: Principles of Data Mining. Springer London (2016)
10. Schober, P., Boer, C., Schwarte, L.: Correlation coefficients: appropriate use and interpretation. Anesth. Analg. **126**(5), 1763–1768 (2018)
11. Zhu, Y., Kuhn, T., Mayo, P., Hinds, W.C.: Comparison of daytime and nighttime concentration profiles and size distributions of ultrafine particles near a major highway. Environ. Sci. Technol. **40**(8), 2531–2536 (2006)
12. Geographic of Surabaya: http://dpm-ptsp.surabaya.go.id/v3/pages/geografis. Last access 3 Jan 2023
13. Weather Atlas in Surabaya: https://www.weather-atlas.com/en/indonesia/surabaya-climate. Last access 3 Jan 2023
14. Orange Data Mining: https://orangedatamining.com/download/#windows (2022)
15. Wu, J.: Advances in K-Means Clustering: A Data Mining Thinking. Springer, Berlin Heidelberg (2012)
16. De Amorim, R.C., Hennig, C.: Recovering the number of clusters in datasets with noise features using feature rescaling factors. Inf. Sci. **324**, 145 (2015)
17. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput. Appl. Math. **20**, 53–65 (1987)
18. Tanjung, S.Y., Yahya, K., Halim, S.: Predicting the readiness of indonesia manufacturing companies toward industry 4.0: a machine learning approach. Jurnal Teknik Industri, Ind. Eng. J. Res. Appl. **23**(1), 1–10 (2021). https://doi.org/10.9744/jti.23.1.1-10
19. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is Nearest Neighbor Meaningful? Computer Sciences Department, University of Wisconsin, Technical Report #1377 (1998). https://minds.wisconsin.edu/bitstream/handle/1793/60174/TR1377.pdf?sequence=1
20. Rennie, J.D.M, Shih, L., Teevan, J., Karger, D.R.: Tacking the Poor Assumptions of Naïve Bayes Text Classifiers (2003). http://people.csail.mit.edu/jrennie/papers/icml03-nb.pdf
21. Rokach, L., Maimon, O.: Top-down induction of decision tress classifiers- a survey. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **35**(4), 476–487 (2005)
22. Shi, T., Horvath, S.: Unsupervised learning with random forest predictors. J. Comput. Graph. Stat. **15**(1), 118–138 (2006)
23. Jeatrakul, P., Wong, K.W.: Comparing the performance of different neural networks for binary classifications problems. In: Proceeding of the Eighth International Symposium on Natural Language Processing (2009). https://doi.org/10.1109/snlp15315.2009
24. Faraggi, D., Reiser, B.: Estimation of the area under the ROC curve. Stat. Med. **21**, 3093–3106 (2002)