Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

# Machine learning approach to uncover customer plastic bag usage patterns in a grocery store

**Iman Sudirman[a] and Ivan Diryana Sudirman[b*]**

[a]*Department of Mechanical Engineering, Faculty of Engineering, Pasundan University, Jl. Tamansari No.6, Bandung, 40116, West Java, Indonesia*
[a]*Entrepreneurship Department, BINUS Business School, Undergraduate Program, Bina Nusantara University, Jl. Pasir Kaliki No.25-27, Bandung, 40181, West java, Indonesia*

**C H R O N I C L E**

*Article history*:
Received: February 23, 2023
Received in revised format: April 2, 2023
Accepted: May 11, 2023
Available online: May 11, 2023

*Keywords:*
*Machine learning*
*Decision tree*
*Data Mining*
*Environment*
*Plastic Bag*
*Waste Management*

**A B S T R A C T**

Plastic bags are used by many people because they are inexpensive, lightweight, durable, and waterproof. Plastic bags, on the other hand, do not break down and can pollute the environment if not handled properly. Indonesia produces a lot of plastic waste and is one of the top ten countries that has a problem with plastic waste. In this study, we used three months of data of real transactions from a grocery store. This study shows how the decision tree can identify patterns on plastic bag usage at a small grocery store by using demography and products purchase. The attribute weights showed that in the hometown, the total of several products bought were the factors that affected the use of plastic bags.



## 1. Introduction

The worldwide issue of plastic trash continues to be managed. Pollution caused by plastic trash has the potential to harm wildlife, flora, and people. Since it is not possible to recycle plastic in the environment, it will continue to build up and worsen environmental problems. Animals can become entangled in plastic, and the material can even enter their digestive systems. Indonesia is still working on a plan to control its plastic trash. 11.7 million tons, or 17 percent, of the 68.5 million tons of trash generated in 2021 were in plastic (Indonesia, 2021). According to the ministry of environment, in 2022 as much as 41.3% of waste was a type of food waste, the second largest after that is 18.5%, namely plastic type waste. Food waste is indeed the largest, although it also has a negative impact on the environment, but food waste can still be recycled by nature, unlike plastic materials (2022). According to the World Bank (2021), large quantities of plastic waste are produced annually, and Indonesia's waste management infrastructure is inadequate to deal with this problem. An estimated 7.8 million tons of plastic waste is produced each year in Indonesia, with over half of that amount going to waste due to improper disposal, as reported by the World Bank. Uncollected or “disposed of in open landfills or leakage from improperly managed landfills”, almost 4.8 million metric tons of plastic waste accumulate each year in Indonesia, In Indonesia, there is a significant problem with plastic waste management, with large amounts of plastic waste generated each year and limited capacity to manage it. According to data from the Environment and Forestry Ministry (Post, 2022), of the 68.5 million tons of waste generated in Indonesia in 2021, 17% (11.6 million tons) were plastic waste. However, only a maximum of 9% of this plastic waste can be recycled, according to data from the Indonesian Movement for the Plastic Bag Diet. This has led to the accumulation of waste in landfills in Indonesia’s major cities. Indonesia faces significant challenges in managing plastic waste. These challenges highlight the need for effective strategies to manage and reduce plastic waste in Indonesia (See Fig. 1).

* Corresponding author.
E-mail address: ivan.diryana@binus.ac.id (I. D. Sudirman)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2023.5.011

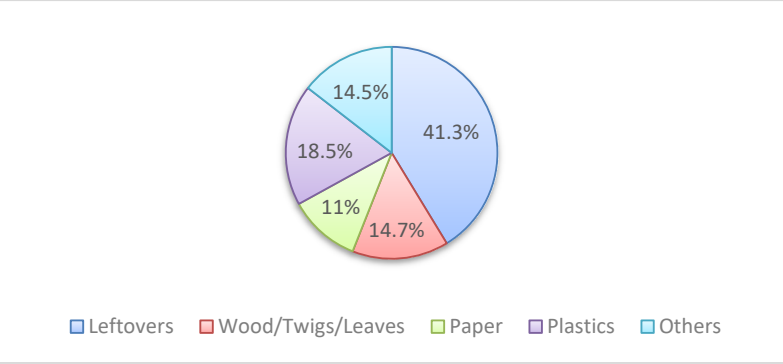


**Fig. 1.** Indonesian Waste Percentage by Type

According to the Plastic Management Index's (VOA, 2022), plastic management in 25 nations reveals that Indonesia lags behind Vietnam, Thailand, and Malaysia in terms of plastic management. In contrast, Japan, Australia, and China claim the top three spots in the world for plastic management in the Asia-Pacific region. It was reported that the Plastics Management Index (PMI) is measured using three pillars: the governance structure, current systematized management capacity, and stakeholder participation. Twelve indicators and forty-four sub-indicators were compiled for these three. Making Oceans Plastic Free (Free, 2017) reports that around 182.7 billion plastic bags are used every year in Indonesia. A total of 1,278,900 tons of plastic bags are discarded each year in Indonesia. A 2015 study by Jenna R. Jambeck and coworkers found that Indonesia contributes nearly as much plastic waste to the ocean as China does. Indonesia is responsible for contributing at least 16% of the plastic trash found in the ocean. According to a study conducted by Schirinzi, Pomeda, Scanchis, Rossini, Farre, and Barcelo (2017), microplastics in seafood and beverages are hazardous to human cells. The Indonesian government has taken several steps to cut down on plastic waste, one of which is requiring shoppers to pay for the plastic bags they use when grocery shopping. This approach is less efficient, however, because high prices for consumers would have a chilling effect on the economy, public approval, and so on. One strategy for reducing plastic waste is to analyze the frequency of plastic bag use. Foreseeing plastic bag use would allow for the detection of causes based on the pattern of use. Assuming this trend is recognized, appropriate measures can be taken. While it is possible that this effort won't be able to eliminate plastic bag use, it should at least reduce the annual plastic bag usage and, in turn, the environmental impact. Accordingly, this study uses machine learning techniques, namely decision trees, to attempt to discern the behavior of grocery store shoppers in a city in West Java, Indonesia. Because clients must pay to use plastic bags after completing their purchase, the cashier tracks the number of plastic bags utilized. This study aims to provide an overview of how to utilize machine learning to disclose plastic bag use patterns from August to October 2022 using a three-month sales data set. Therefore, the purpose of this research is to use machine learning to identify patterns of plastic bag usage by analyzing demographic information (such as age and gender), decision trees, and purchase data (including the number of items purchased by each customer). Predictive models for plastic bag usage can be built using the data trends and the preexisting variables. It is hoped that this study will aid in resolving current environmental issues by shedding light on how shoppers use or buy plastic bags.

## 2. Related Study

Several researchers have dedicated their efforts to study the prediction of the behavior pattern of plastic bags or plastic waste. One of them is a study that studied the patterns and determinants of consumer plastic bag use when shopping in Da Nang, Vietnam, using an original home survey, interviews with key informants, and behavioral theory and machine learning approaches. Two sociodemographic variables and seven socio-psychological variables were found to be as important in predicting the total use of plastic bags in Vietnam. The results showed that the use of plastic bags is widespread and has become a habit frequently (Makarchev et al., 2022). Another related research analyzed transaction data from a high-street health and beauty business to evaluate plastic bag purchasing habits of more than 12,000 people in the United Kingdom and revealed statistical disparities in plastic bag purchasing between areas. A subsample of frequent and infrequent plastic bag buyers was selected and used for predictive modeling and an experimental machine learning technique was utilized to study demographic and psychological correlates of frequent plastic bag use. The findings revealed that frequent plastic bag buyers were younger, more likely to be male, spent more money in the store, were less thrifty, open to new experiences, and unhappy with their physical appearance (Lavelle-Hill et al., 2020). Despite efforts to limit it through levies, the use of plastic bags in South Africa remains ubiquitous. Another study used an online survey to analyze plastic bag usage in South Africa and the factors that influence it. Most of the respondents believed that there was an issue with the use of plastic bags in the country, but regularly used plastic bags due to their convenience. People's willingness to pay for plastic bags was affected by sex, age, education, and environmental awareness, although the connections were often modest. Certain treatments may be successful in developing environmentally conscious behavior, according to the study O'Brien and Thondhlana (2019). Plastic bags are widely used for their low cost, lightweight, durability, and waterproof properties, but they are also nonbiodegradable and can pollute the environment if not managed properly. Indonesia is a significant contributor to plastic waste and is one of the top ten countries

with problems in the management of plastic waste. Another study aims to determine the factors influencing behavior related to reducing the use of plastic bags in Indonesia, using the theory of planned behavior and partial least squares for statistical analysis and hypothesis testing (Nabila et al., 2020).

In the context of data mining, this technique is frequently employed to identify trends in large datasets. Therefore, to identify trends in this dataset observed, this study employs this methodology. Patterns that have been effectively uncovered by data mining are then used to produce prediction that can help to aid in decision-making, comprehension of complicated events, and so on. Due to the rapid development of information technology, data mining is utilized in a variety of industries, including business, finance, research, health, and others. Data mining employs a variety of prevalent models, including clustering, classification, regression, association, and so forth (Agarwal, 2013; Han et al., 2000; Witten & Frank, 2000).

Data mining and machine learning are closely related and frequently overlap fields. Machine learning is an area of artificial intelligence that focuses on the development of unprogrammed systems that can learn from data. The performance of these algorithms is intended to automatically improve as they are exposed to more data. The decision tree is a prevalent example of a machine learning algorithm used in data mining. Decision trees are a form of supervised learning technique that can be used to classify data based on properties or characteristics. They function by constructing a decision tree based on the values of the data's features. Each branch of the tree represents a distinct choice, while the tree's terminal nodes denote the projected classifications. In data mining, decision trees are frequently used to categorize data into several categories or to create predictions based on aspects of the data. Clustering, classification, regression, and association rule learning are other approaches frequently employed in data mining (Gorunescu, 2011; Jensen, 2006; Wang, 2013)

The decision tree is just one example of the many machine learning models used in machine learning. Because of its popularity and many benefits, the decision tree was chosen as the model for this research. The study's goal is to shed light on consumers' plastic bag use patterns, and the Decision Tree has several advantages over alternative models in this regard. Decision trees are simple to implement, flexible, and applicable to a wide variety of data types because of their ability to process both numerical and categorical information. The decision tree is easy to implement and train, which means it can speed up the analysis of large datasets. The decision tree can also deal with missing values and corrupted data without the need for any preprocessing, as it is resistant to noise and missing values (Abdelhalim & Traore, 2009; Jensen, 2006; Yang, 2019).

## 3. Method

This research uses the CRISP-DM method, also known as the Cross Industry Standard Process for Data Mining, is a framework for data mining initiatives that is widely adopted. It provides a systematic approach for identifying and resolving the many processes involved in a data mining project, such as defining the business problem and objectives, selecting and preparing the data, constructing and evaluating the model, and implementing the findings. There are six stages in CRISP-DM, namely: business understanding, data understanding, data preparation, modeling, model evaluation, and deployment. Each step consists of several tasks and activities that help guide the data mining procedure and ensure that the results are accurate, reliable, and relevant to the business challenge at hand. The CRISP-DM method is designed to be versatile and adaptable to the requirements of many sectors and organizations, and is frequently combined with other data mining approaches and tools for optimal results (Chapman, 2000).The data set used in this study is sales data for 3 months from a grocery store. The store is in a district near the city of Bandung. Consumers who become members will give their member cards during transactions to get points from the loyalty program. Every transaction using a plastic bag is recorded because there is a fee that must be paid by the consumer, even though it is not expensive. The data set consists of 5,374 examples, representing the total number of member customers who were recorded as having conducted transactions from August through October 2022. The quantity of things sold was drastically reduced based on the total of sales. This is done to conserve computing resources. After partitioning the data, the attributes that remain are hometown, age, sex, member id, plastic bag, net sales, spending, egg, fried indomie, indomie chicken, grapefruit, fortune cooking oil, yakult, and sovia cooking oil. For the hometown attribute, the data were divided into Greater Bandung City which consists of Bandung City, Bandung district, and Cimahi. Sumedang will become its own category then, aside from that, grouped into the label named Outside Bandung. To aid interpretation, the numerical data for the net sales attribute are grouped into three categories. High, which spending is above 3.000.000 Rupiah, Medium, those who spend between 1.000.000 Rupiah and 3.000.000 Rupiah, and Low, which spending is below 1.000.000 Rupiah. The data type for product attributes such as egg, grape, etc. is numeric and indicates the total amount of the item purchased on the period when the dataset was compiled. For plastic bags, it is the number data type that shows the amount of plastic used by customers. Then they were reclassified into High and Low categories. The plastic bag usage data are shown as a histogram to determine the amount threshold for plastic bag usage between High and Low. Upon examination of the histogram, it was determined that a total value below 6 would indicate Low and a total value above 6 would indicate High.

## 4. Result

This study aims to identify patterns within this dataset. The anticipated attribute is a plastic bag that is divided into low- and high-class. As indicated previously, this study predicts plastic bag usage using a decision tree. One of the reasons is because the pattern is easily discernible and apparent. This research uses RapidMiner to analyze the data set. The following operators

are utilized. The operator shown in Fig. 2 is the operator used to equalize samples from the two groups, High and Low, in the use of plastic bags. The number of low instances of plastic usage is greater than the number of high instances. Class samples must be proportional so that the results are not skewed. Consequently, the lower class needs to be under-sampled using stratified sampling.

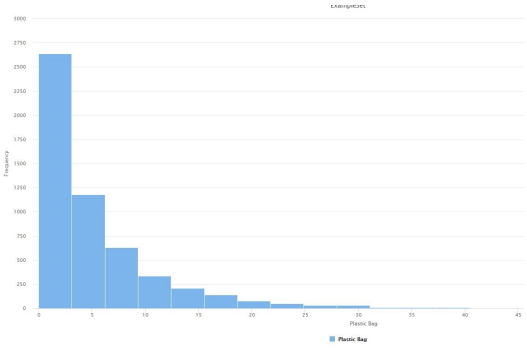

**Fig. 2.** Plastic Bag Usage

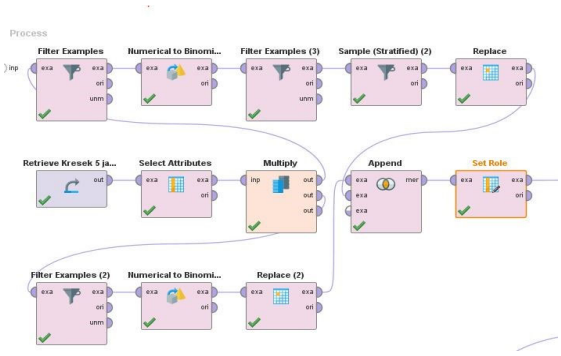


**Fig. 3.** Operator Use for Dataset Preparation

Therefore, each class includes 1558 instances. So, the total samples used are 3116. After ensuring that the data is prepared, the role in the operator role is set. The Optimize Parameter operator, which iterates over the model's parameters until optimal results are obtained, is employed. For validation, this study uses cross-validation and observes several parameters which are precision, recall, and accuracy. In addition, this study also presents the results of the optimal model with simulation operators. Due to computational limitations, operators must be run many times by varying a series of parameters to optimize on the Optimize Parameters operator and find the best accuracy. The table above shows the results of using an optimized decision tree.

**Table 1**
Model Performance

| | |
|---|---|
| Accuracy | 60.69% |
| Classification Error | 39.31% |
| Kappa | 0.2.14 |
| Recall | 60.69% |
| Precision | 60.79% |

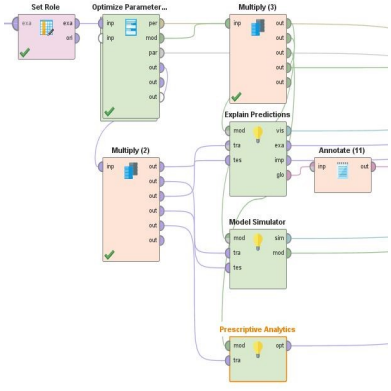


**Fig. 4.** Operator Use to Generate Predictive Results

The accuracy is the percentage of correct predictions made by the model, which in this case is 60.69%. The classification error is the percentage of incorrect predictions made by the model, which is 39.31% in this case. Kappa is a statistical measure that compares the accuracy of the model to the accuracy that would be expected by chance alone, with a value of 0.214 in this case. Recall is the percentage of true cases that were correctly predicted by the model, which is 60.69% in this case. Precision is the percentage of predicted cases that are true, which is 60.79% in this case. The results are quite good considering that this data set is related to human behavior. The optimizeی parameters of the model for this data set are Decision Tree.maximal depth = 40, Decision Tree.criterion = precision, Decision Tree.minimal leaf size = 1, Decision Tree.minimal size for split = 41, Decision Tree.apply prepruning = true. The weight of the attribute in a decision tree is a measure of how important each attribute was in making the tree.

**Table 2**
Attribute Weight

| | |
|---|---|
| Home town | 0.259 |
| Egg | 0.111 |
| Grape | 0.073 |
| Age | 0.032 |
| Fortune | 0.026 |
| Gender | 0.020 |
| Spending | 0.017 |

## 5. Discussion

The attribute with the most weight in this table is “hometown”. This means that it had the greatest effect on how the decision tree was made. On the other hand, “Spending” is the attribute with the least weight, which means that it had the least effect on how the decision tree was made. Based on these results, it can be said that the hometown attribute is the most important in predicting the use of plastic bags, then the total of purchasing eggs, wine, age, Fortune brand cooking oil, gender, and finally spending. As can be observed in the decision tree image, the link between the “hometown” variable and the probability of a person being classed as “Low” or “High” based on their “Age,”, “Grape”, “Spending”, “Fortune,” and “Egg” values. The data are initially divided based on the “hometown” variable. For instance, if the individual is from Bandung, the tree divides based on their “Age”. If a person’s “Age” is larger than or equal to 21.5, they are categorized as “Low”, and if it is less than or equal to 21.5, they are categorized as “High”. Until it reaches a leaf node with a classification of “Low” or “High,” the tree continues to divide and classify the data in this manner. The number contained within the curly brackets at each leaf node indicates the number of “High” and “Low” classifications for the data at that node.

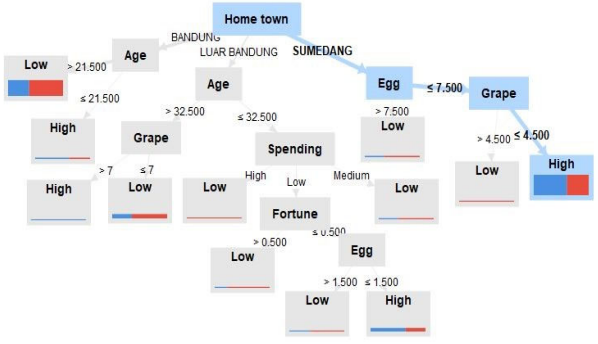


**Fig. 5.** The Decision Tree Result

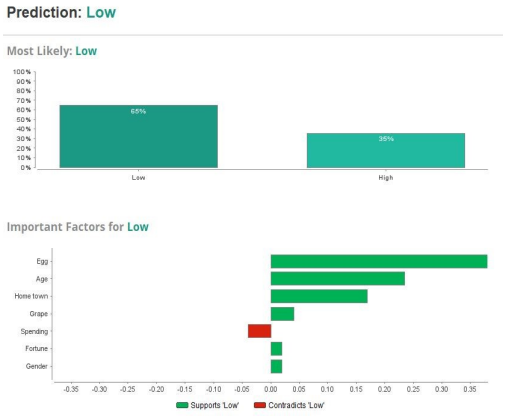


**Fig. 6.** Simulation Result

Since this dataset comes from a store located in Sumedang, let us observe Home Town Sumedang for plastic bag usage is High, which is highlighted in blue path. For the “Home town = SUMEDANG” branch of the decision tree, it appears that the model uses the attribute “Egg” to forecast whether a customer’s plastic bag usage will be high or low. If a consumer has more than 7,5 eggs, the model predicts that they will use fewer plastic bags. If the consumer has fewer than or equal to 7.5 eggs, the model will predict using the attribute “Grape”. The algorithm forecasts low plastic bag usage if the customer consumes more than 4.5 grapes. If the consumer has fewer than or equal to 4.5 grapes, the model predicts that they will use numerous disposable bags. This was interesting to us because the number of plastic bags used depended not only on where the store was, but also on how many eggs and fruits were bought. In this case, the number of grapes purchased was the most important factor. The low number of eggs and fruits purchased during the 3-month period suggests that people will use many plastic bags. This could happen because consumers often buy and need both things, but if they are bought in small amounts, they are bought more often, which means that more plastic bags are used. RapidMining allows model simulation; this study uses simulation to find how the variables interact with the prediction. The results of the weighting and decision tree show that the attribute of the hometown has the greatest weight. So, in this simulation, the hometown is set to Sumedang. After that, the next most influential variable is also set. Fig. 5 shows the result when the hometown is set to be Sumedang, then move the number of eggs right and left until there is a change from High to Low. The result is that if the total purchase of eggs is more than 8 then it is likely that the use of plastic bags will be low, regardless of age, gender, spending or other products.

## 6. Conclusion

This research suggests that the decision trees model was successful in discovering patterns within the provided dataset. Attribute weights and a decision tree model illustrated the interconnectedness of factors that influence consumers' reliance on plastic bags. Evaluation metrics also demonstrated that the decision tree model could predict the target variable in human behavior-related datasets, which are notoriously difficult to forecast. These findings demonstrate the potential utility of data mining and machine learning for gaining insight into and forecasting patterns of plastic bag usage. This study’s attribute weights, and decision tree give useful information about the factors that affect plastic bag use, such as hometown, age, egg, grape, cooking oil brand, gender, and spending. These results can help governments and organizations develop programs and policies to reduce plastic bag use and encourage people to be involved in preserving the environment. One problem with this research is that it is based on a single dataset, which may not be representative of the entire population. In addition, this study only used one method of machine learning, which is the decision tree, which means that other methods might have led to different results. The accuracy of the predictions may be improved by using more complex models such as deep learning.

**References**


Abdelhalim, A., & Traore, I. (2009). A New Method for Learning Decision Trees from Rules. *2009 International Conference on Machine Learning and Applications*, 693–698. https://doi.org/10.1109/ICMLA.2009.25

Agarwal, S. (2013). Data Mining: Data Mining Concepts and Techniques. *2013 International Conference on Machine Intelligence and Research Advancement*, 203–207. https://doi.org/10.1109/ICMIRA.2013.45

Chapman, P. (2000). *CRISP-DM.pdf*. http://www.statoo.com/CRISP-DM.pdf

Free, M. O. P. (2017, November 11). *The Hidden Cost of Plastic Bag Use and Pollution in Indonesia • Making Oceans Plastic Free*. Making Oceans Plastic Free. https://makingoceansplasticfree.com/hidden-cost-plastic-bag-use-pollution-indonesia/

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques*. Springer Science & Business Media.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM SIGMOD Record*, *29*(2), 1–12. https://doi.org/10.1145/335191.335372

Indonesia, C. N. N. (2021). *Sampah Plastik 2021 Naik ke 11,6 Juta Ton, KLHK Sindir Belanja Online*. nasional. https://www.cnnindonesia.com/nasional/20220225173203-20-764215/sampah-plastik-2021-naik-ke-116-juta-ton-klhk-sindir-belanja-online

Jensen, W. A. (2006). *Decision Trees for Business Intelligence and Data Mining: Using SAS® Enterprise Miner™: Technometrics: Vol 50, No 3*. SAS Institute, 2006.

Lavelle-Hill, R., Goulding, J., Smith, G., Clarke, D. D., & Bibby, P. A. (2020). Psychological and demographic predictors of plastic bag consumption in transaction data. *Journal of Environmental Psychology*, *72*, 101473. https://doi.org/10.1016/j.jenvp.2020.101473

Makarchev, N., Xiao, C., Yao, B., Zhang, Y., Tao, X., & Le, D. A. (2022). Plastic consumption in urban municipalities: Characteristics and policy implications of Vietnamese consumers' plastic bag use. *Environmental Science & Policy*, *136*, 665–674. https://doi.org/10.1016/j.envsci.2022.07.015

Nabila, Y., Nurcahyo, R., & Farizal. (2020). The Key Factors in Reducing the Use of Plastic Bags. *2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA)*, 197–201. https://doi.org/10.1109/ICIEA49774.2020.9102102

O'Brien, J., & Thondhlana, G. (2019). Plastic bag use in South Africa: Perceptions, practices and potential intervention strategies. *Waste Management*, *84*, 320–328. https://doi.org/10.1016/j.wasman.2018.11.051

Post, T. J. (2022). *Tons of trash, less to recycle: The irony of recycling plastic waste*. The Jakarta Post. https://www.thejakartapost.com/culture/2022/05/24/tons-of-trash-less-to-recycle-the-irony-of-recycling-plastic-waste.html

Schirinzi, G. F., Pérez-Pomeda, I., Sanchís, J., Rossini, C., Farré, M., & Barceló, D. (2017). Cytotoxic effects of commonly used nanomaterials and microplastics on cerebral and epithelial human cells. *Environmental Research*, *159*, 579–587. https://doi.org/10.1016/j.envres.2017.08.043

VOA. (2022). *Mengerikan, Indonesia Sudah Darurat Sampah Plastik: Sehari Mencapai 64 Juta Ton, Nomor Dua Terbesar di Dunia*. VOI - Waktunya Merevolusi Pemberitaan. https://voi.id/bernas/137477/mengerikan-indonesia-sudah-darurat-sampah-plastik-sehari-mencapai-64-juta-ton-nomor-dua-terbesar-di-dunia

Wang, Z. (2013). Application of Decision Making Tree Method in the Real Estate Development Scheme Optimization. *2013 Third International Conference on Intelligent System Design and Engineering Applications*, 365–368. https://doi.org/10.1109/ISDEA.2012.91

Witten, I. H., & Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

World Bank. (2021). *Plastic Waste Discharges from Rivers and Coastlines in Indonesia* [Text/HTML]. World Bank. https://www.worldbank.org/en/country/indonesia/publication/plastic-waste-discharges-from-rivers-and-coastlines-in-indonesia

Yang, F.-J. (2019). An Extended Idea about Decision Trees. *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 349–354. https://doi.org/10.1109/CSCI49370.2019.00068