Classifying Garbage Images: A Comparative Study of Naive Bayes, SVM, and Logistic Regression Models

Ma. Cristina M. Pasague
Bachelor of Science in Computer Science
Negros Oriental State University
cristinapasague27@gmail.com

Abstract—**The escalating global challenge of waste management necessitates innovative solutions to enhance waste sorting processes. Thus, this comparative study was conducted evaluating the effectiveness of three machine learning models – Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression – in classifying garbage images into five categories: cardboard, glass, metal, paper, and plastic. Leveraging a dataset sourced from Kaggle, consisting of 284 garbage images, the models are evaluated using a 5-fold stratified cross-validation. The objective was to assess the accuracy of each model in classifying garbage images, determine the proportion of instances between the predicted and the actual class, and identify misclassified garbage instances. All models achieved high accuracy, with Logistic Regression exceeding the others achieving a remarkable AUC of 0.999, Classification Accuracy (CA) of 0.979, F1 Score of 0.979, Precision of 0.979, and Recall of 0.979, a near-perfect performance. However, all models struggled with the distinction between visually similar materials, resulting in misclassification. The study concluded that while the models demonstrated strong performance in classifying garbage images, there is room for improvement, particularly in distinguishing between the 'glass' and 'metal' categories. This study will provide valuable insights for optimizing future algorithms and improving the overall accuracy of automated waste management systems. This, in turn, can lead to significant environmental and economic benefits, paving the way for a more sustainable and efficient future for waste management.**

Keywords—Garbage image classification, Naïve Bayes, Support Vector Machine, Logistic Regression, waste management, sustainability.

I. Introduction

The exponential rise in global waste generation presents a consequential environmental challenge. Inefficient waste segregation practices can exacerbate environmental issues, leading to pollution, resource depletion, or public health risks. Pollution from overflowing landfills and improper disposal poisons air, soil, and water, harming ecosystems. Resource depletion intensifies as valuable materials are buried instead of recycled. Quality of life plummets in communities burdened by poor waste management systems [1].

Current waste sorting practices often rely on manual labor, which is time-consuming and prone to human error [2]. On the other hand, the rapid advancement in technology has led to the development of numerous applications that enhance our lives. In the wake of increasing waste concerns, the development of accurate automated systems has become in demand [3]. To address such concerns, here comes Machine Learning models, with their ability to learn from data and recognize patterns. They could offer promising solutions for automated waste management systems thus, streamlining waste sorting and processing [4]. However, the effectiveness of different Machine Learning models in classifying specific types of garbage remains a subject of ongoing investigation.

Henceforth, this study presents a comprehensive analysis of the effectiveness of machine learning classification models namely Naive Bayes, SVM, and Logistic Regression for categorizing garbage images into five distinct classes: cardboard, glass, metal, paper, and plastics. In particular, these algorithms are utilized to categorize images of garbage in an effort to promote a more sustainable waste management approach. Such an approach makes it easier for individuals and municipalities to manage their waste effectively [5].

The primary objective of this study is to evaluate the effectiveness of various machine learning classification models in the task of categorizing images of garbage to contribute to the development of accurate and reliable automated systems for efficient waste management. And by that, the research addresses the following key aspects: first is assessing the accuracy of each machine learning classification model in classifying garbage images; second would be determining the proportion of instances between the predicted and the actual class and; lastly, identify the misclassified garbage instances.

II. Literature Review

2.1. Naïve Bayes

The Naïve Bayes algorithm, renowned for its simplicity and efficiency, was proven effective in image classification tasks by a study that tackles the challenge of classifying images of thin metal plates based on their deformation level. The researchers gathered both undeformed and deformed plates and then preprocessed every image as the input for the Naïve Bayes classifier. Trained on the labeled dataset, the algorithm learned all feature values occurring in each deformation class (e.g. safe, slightly deformed, excessively deformed). Based on the learned features, the model classified the new and unseen images into their corresponding deformation levels. The study then evaluated the accuracy of these classifications using metrics like precision, recall, and F1-score. The outcome was that Naïve Bayes achieved an accuracy of over 85% in classifying the deformation levels of the metal plates. This case study serves as a testament to the model's potential for image classification particularly when dealing with well-defined categories [6].

2.2. Support Vector Machine

Another study investigates the use of a machine learning algorithm, Support Vector Machine or SVM, for identifying diseases in plant leaves. The study begins with the image enhancement techniques and extracted features from the images. The features were the morphology of leaf spots which provide crucial information about the visual depiction of the disease. The features are then matched to the variance of the gray level of the red, green, and blue channels of the spots. The study evaluated the performance of SVM classifiers using these features. The results confirmed the effectiveness of the model in identifying plant leaf diseases as the analysis has shown 91.71% accuracy of the model in the field of plant disease identification through image analysis [7].

2.3. Logistc Regression

A previous study explored logistic regression for facial expression recognition, evaluating emotion recognition from facial images. The dataset used contains seven basic emotions: anger, contempt, disgust, fear, happiness, sadness, and surprise. Each image comes with facial landmark annotations (wrinkles, furrows, angles, etc.). The study built logistic regression models to predict the probability of an image belonging to a specific emotion. The model achieved an average accuracy of 82.2% across all seven emotions. Happiness (90.1%), surprise (87.5%), and sadness (84.56%) had the highest accuracy while contempt (76.12%) and fear (74.2%) had the lowest accuracy. Logistic regression provides insights into which facial features contribute most to each emotion classification. Overall, the study suggests that logistic regression can be a viable option for image recognition especially when interpretability and efficiency are high priorities [8].

III. Methodology

Below are the outlines of the approach taken and the details of the steps followed to achieve the research objectives.

Data Overview

3.1.1 Data Collection

Data collection for this study involved extracting relevant information containing images of garbage. The dataset utilized was obtained from a dataset repository about garbage images available on a public platform known as Kaggle. This dataset was carefully chosen to ensure a diverse representation of garbage images, encompassing a decent range of waste materials. To adhere to ethical considerations, any personally identifiable information within the images was removed. This anonymized garbage image dataset could provide a rich and varied disposal scenario, allowing for a robust analysis of machine learning classification models without compromising the privacy of individuals or organizations associated with the disposal of these materials.

3.1.2. Data Characteristics

The dataset was a collection of 284 unique images in a JPEG format, distributed among five distinct categories namely cardboard, glass, metal, paper, and plastic. There are 31 cardboard images, 60 glass images, 57 metal images, 83 paper images, and 53 plastic images. The dimension of every single image was 512x384 pixels. Different angles and parts of the waste materials were taken into consideration.

Orange Data Mining Tools and Algorithms

The research centers on using machine learning techniques to categorize garbage images. With that, orange data mining software stands as an optimal choice for building classification models.

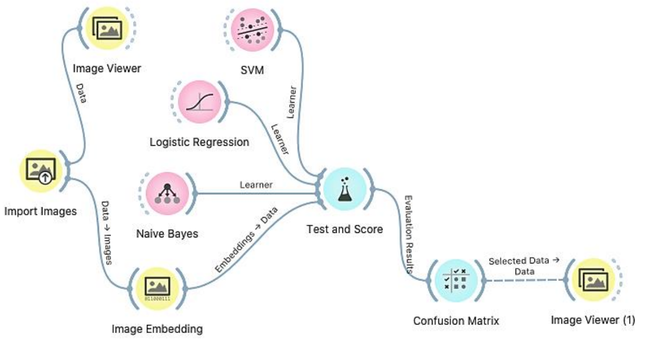

Fig. 1. Orange Data Mining Workflow

3.2.1. Naïve Bayes

Naïve Bayes makes predictions based on assuming features are independent. It utilizes Bayes' Theorem, which assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors [9]:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \qquad (1)$$

- $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(x|c)$ is the likelihood which is the probability of the *predictor* given *class.*
- $P(c)$ is the prior probability of *class.*
- $P(x)$ is the prior probability of the *predictor*.

3.2.2. Logistic Regression

Logistic Regression is a method for binary classification, predicting outcomes like “yes/no”. It examines the relationship between independent variables and assigns data into two classes. Often employed when dealing with binary outcomes like negative (0) or positive (1) classes [10]. The algorithm uses the formula:

$$f(x) = \frac{1}{1+e^{-x}} \qquad (2)$$

- e is the base of natural logarithms.
- value is the numerical value one wishes to transform.

3.2.3. Support Vector Machine

SVM aims to find a hyperplane in a high-dimensional space to separate data points of different classes. The key objective is to maximize the margin, which is the distance between the hyperplane and the nearest data points of each class [11]. The equation of a hyperplane is given by:

$$f(x) = w \cdot x + b \qquad (3)$$

- w is the weight vector.
- x is the input feature vector.
- b is the bias term.

For non-linearly separable data, SVM uses a kernel trick to map features into a higher-dimensional space where a hyperplane can separate the classes [11]. The decision function becomes:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \qquad (4)$$

- α_i are the Lagrange multipliers.
- y_i is the class label
- X_i is a support vector and x is the input feature vector
- K is the chosen kernel function
- b is the bias term.

IV. Results and Discussions

4.1 The accuracy of each machine learning classification model in classifying garbage images

TABLE I. PERFORMANCES OF THE ALGORITHMS

Test And Score					
Model	***AUC***	***CA***	***F1***	***Precision***	***Recall***
Naïve Bayes	0.989	0.908	0.908	0.911	0.908
SVM	0.998	0.965	0.965	0.966	0.965
Logistic Regression	0.999	0.979	0.979	0.979	0.979

In the evaluation of three machine learning models using a 5-fold stratified cross-validation, the Logistic Regression model emerged as a strong classifier and exhibited outstanding near-perfect performance across the board with a remarkable AUC of 0.999, Classification Accuracy (CA) of 0.979, F1 Score of 0.979, Precision of 0.979, and Recall of 0.979. The Support Vector Machine (SVM) model also demonstrated great results, with an AUC of 0.998, CA of 0.965, F1 Score of 0.965, Precision of 0.966, and Recall of 0.965. The Naive Bayes model, while performing well, showed slightly lower accuracy compared to the other two models, with an AUC of 0.989, CA of 0.908, F1 Score of 0.908, Precision of 0.911, and Recall of 0.908. These findings suggest that all models are capable of accurately categorizing garbage images, with Logistic Regression leading in terms of overall accuracy and precision-recall balance, closely followed by SVM, while Naive Bayes remains a competent performer even with slightly lower accuracy. The findings highlighted the effectiveness of all three models and emphasized their potential for robust applications in waste management systems [12].

4.2 Proportion of instances between the predicted and the actual class

4.2a Naïve Bayes Model

TABLE II. MODEL EVALUATION METRIC OF NAÏVE BAYES

Predicted

		cardboard	glass	metal	paper	plastic	
Actual	**cboard**	31	0	0	0	0	31
	glass	0	48	9	0	3	60
	metal	0	2	51	1	3	57
	paper	3	0	0	79	1	83
	plastic	0	2	1	1	49	53
		34	52	61	81	56	284

The confusion matrix for the Naive Bayes model in a multi-class garbage image classification task breaks down its performance across five classes: cardboard, glass, metal, paper, and plastic. Notably, the model demonstrates strong accuracy with an overall accuracy rate of approximately 90.84%. Delving into specific classes, the model had high precision for cardboard, paper, and plastic, correctly identifying instances with only a few misclassifications ranging from 1 to 3. However, misclassifications are observed particularly between metal and glass, which is 9 misclassified images. Naïve Bayes correctly predicted all 31 cardboards, 48 out of 60 glass type, 51 over 57 in metal type, 79 over 83 in paper type, and 49 out of 53 plastics. These findings from the confusion matrix offer insights into the model's effectiveness, identifying areas for potential improvement and optimization, especially in distinguishing between visually similar materials such as metal-glass, glass-plastic, metal-plastic, and then cardboard-paper [6].

4.2b Support Vector Machine (SVM) Model

TABLE III. MODEL EVALUATION METRIC OF SVM

Predicted

		cardboard	glass	metal	paper	plastic	
Actual	**cboard**	29	0	0	2	0	31
	glass	0	54	5	1	0	60
	metal	0	0	55	1	1	57
	paper	0	0	0	83	0	83
	plastic	0	0	0	0	53	53
		29	54	60	87	54	284

The confusion matrix for the Support Vector Machine (SVM) model, evaluated in a multi-class garbage image classification scenario, provides a comprehensive overview of the model's classification performance across the five material categories. Evidently, the SVM model achieved an overall accuracy rate of approximately 96.47%. In specific classes, such as cardboard, metal, paper, and plastic, the model showcases great classification, indicating the model's effectiveness in accurately identifying images within the mentioned categories. However, in the glass category, a few misclassifications are observed, with 5 images mistakenly categorized as metal. Additionally, the model incorrectly predicted 3 instances to be paper instead of cardboard, glass, or metal in actuality. Despite this, the SVM model excels in correctly distinguishing between different materials, particularly evident in achieving perfect classification for plastic and paper instances. These findings from the confusion matrix shed light on the model's strengths and areas for potential refinement, contributing valuable insights for further optimization [7].

4.2c Logistic Regression Model

TABLE IV. MODEL EVALUATION METRIC OF LOGISTIC REGRESSION

Predicted

		cardboard	glass	metal	paper	plastic	
Actual	**cboard**	31	0	0	0	0	31
	glass	0	56	4	0	0	60
	metal	0	0	55	1	1	57
	paper	0	0	0	83	0	83
	plastic	0	0	0	0	53	53
		31	56	59	84	54	284

The confusion matrix for the Logistic Regression model in the context of classifying multi-class garbage images, provides a comprehensive overview of the model's classification performance across the five material categories. Very impressively, the model achieves an overall accuracy of approximately 97.88%, signifying its outstanding ability to accurately classify instances within the specified categories. Examining each of the classes, the Logistic Regression model excels in correctly classifying instances for cardboard, paper, and plastic, with no false positives observed in any class. While for the case of glass and metal, 4 glasses were misclassified as metal then 2 metals were misclassified as either paper or plastic. Ultimately, the overall proportion of instances viewed from the Logistic Regression model emphasized the high-performance nature of the model, reflecting its effectiveness in handling the complexities of the multi-class garbage image classification task. These findings from the confusion matrix showcase the model's reliability and suitability for applications in waste management, where accurate classification of diverse materials is crucial for effective automated sorting systems [8].

4.3 Misclassified Garbage Instances

4.3a Naïve Bayes model's misclassified instances

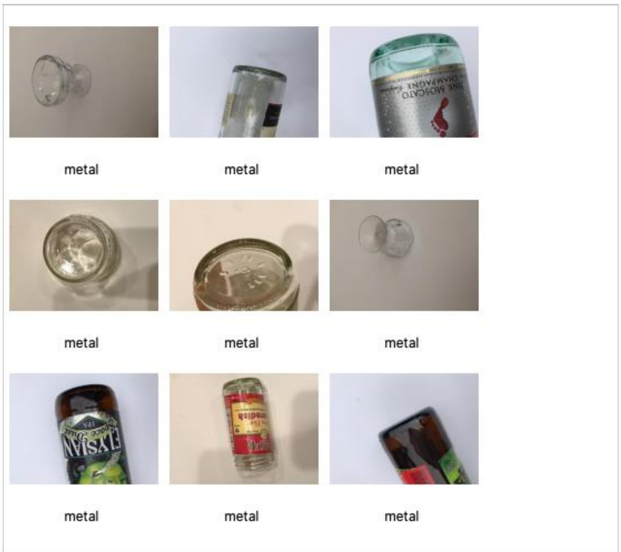

Fig. 2. Glass type of garbage but categorized as Metal

In the evaluation of the Naive Bayes algorithm for garbage image classification, a notable occurrence was the misclassification of nine images originally labeled as "glass" but inaccurately predicted as "metal" by the model. This misclassification suggests potential challenges in distinguishing between the two categories. The images in question, despite being visually identified as glass, led to a conflicting prediction from the Naive Bayes model, indicating a limitation in its ability to discern subtle nuances in the features that differentiate glass and metal materials.

Misclassification can occur for various reasons. One key factor influencing misclassifications is the ambiguity present in the visual characteristics of images. For instance, changes in lighting or shadows can alter the perceived color or visual of an object, making it confusing for the model to distinguish between instances [13].

4.3b SVM model's misclassified instances

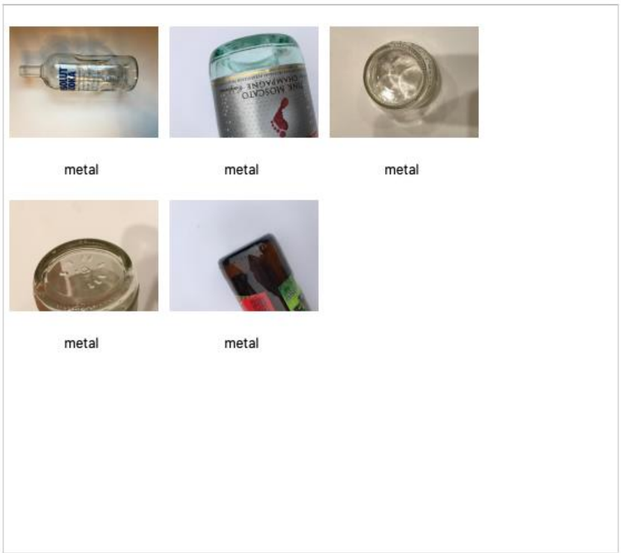

Fig. 3. Glass type of garbage but categorized as Metal

The Naïve Bayes and SVM model both had their highest misclassification in the glass category. SVM also misclassified five glasses as 'metal'. As we can observe, four out of five can also be seen in Figure 2. The repeated misclassification of images indicates a significant challenge in distinguishing between these two categories in the context of garbage image classification. The consistent misclassification suggests potential difficulties arising from shared visual features, ambiguous image characteristics, or limitations in feature representation [13]. The fact that both models exhibit a similar pattern tells the complexity of the task and emphasizes the need for further refinement.

4.3c Logistic Regression model's misclassified instances

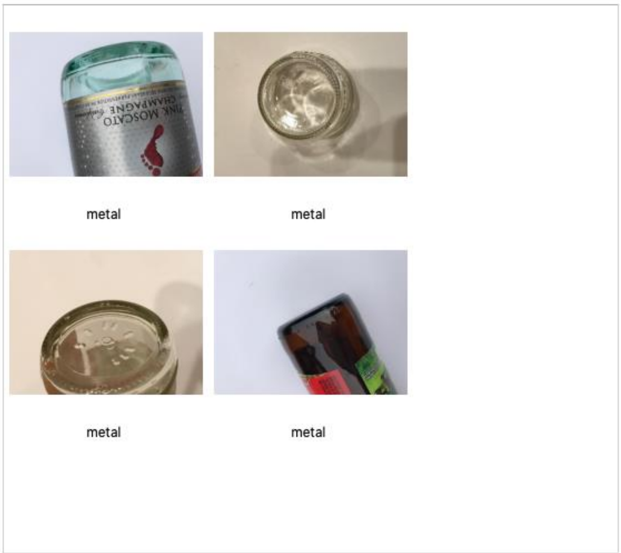

Fig. 4. Glass type of garbage but categorized as Metal

Among the three machine learning classification models, the Logistic regression has the least number of misclassified images in the same category, glass images. Only four glass images were incorrectly categorized as ‘metal’. However, same case with the previous two models discussed, the images were consistently similar across all models. This pattern suggests that there may be specific aspects of the images, such as certain shapes, colors, lighting, reflection, texture, or contextual elements that pose difficulties for the models in distinguishing between glass and metal [13].

V. Conclusion

The evaluation of the three machine learning models – Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression – in classifying garbage images revealed that all three models performed remarkably well, with the Logistic Regression model emerging as the strongest classifier achieving an accuracy of 97.88%.

The primary challenge for all models was distinguishing between similar waste materials particularly glass and metal, with most misclassifications occurring between those two categories. Potential factors contributing to misclassifications include visual ambiguity (e.g., lighting, shadows, reflections), shared visual features between glass and metal, and limitations in feature representation. The findings suggest the suitability of these models for waste management applications, but further refinement is needed to address the glass-metal confusion and enhance overall performance.

Valuable insights derived from the study include: logistic regression might be the preferred model for waste categorizing tasks due to its overall accuracy and interpretability; data augmentation techniques or feature engineering could be explored to improve model robustness to variations in lighting and other visual characteristics and; incorporating domain knowledge about waste materials and their visual properties could potentially improve feature representation and classification accuracy.

Through this research, the objective which is to contribute to the continuous improvement of classification models for enhanced accuracy in waste categorization was achieved. The findings have implications not only for the field of waste management but also for the broader context of sustainable and technology-driven solutions. Nevertheless, further research based on these insights and the implementation of the used algorithms could yield more effective approaches to address waste management system challenges.

VI. References

[1] World Bank. "What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050," World Bank, 2018.

[2] L. Shen and Y. Guo, "A review of automatic waste sorting systems using computer vision," Waste Management, vol. 109, pp. 113-123, 2020.

[3] A. Smith et al., "Advancements in Machine Learning for Waste Management: A Comprehensive Review," Journal of Environmental Science and Technology, vol. 45, no. 2, pp. 123-140, 2019.

[4] L. Koehn, "Applications of machine learning in waste management," Waste Management, vol. 112, pp. 121-130, 2020.

[5] F. Herbert. "Goal 12: Responsible Consumption and Production," Sustainable Development Goals, United Nations, 2015.

[6] R. Kumar, S. Ayyasamy, and S. K. Natarajan, "Naïve Bayes Machine Learning Model for Image Classification to Assess the Level of Deformation of Thin Components," Pattern Recognition Letters, vol. 203, pp. 312-319, 2022.

[7] S. Sivasakthi, "Plant Leaf Disease Identification Using Image Processing and SVM, ANN Classifier Methods," Pattern Recognition Letters, vol. 203, pp. 312-319, 2022.

[8] C. Zhou, L. Wang, Q. Zhang, and X. Wei, "Face recognition based on principal component analysis and logistic regression analysis," Optik Int. J. Light Electron Opt., vol. 125, no. 23-24, pp. 6069-6075, 2014.

[9] Analytics Vidhya. Naive Bayes Classifier: Understanding How It Works. Analytics Vidhya, Sep. 19, 2017.

[10] What Is Logistic Regression? Spiceworks, 2022.

[11] V. Kanade. "Support Vector Machine Formulation and Derivation," Towards Data Science, 2021.

[12] A. Saki and G. A. Ferns, "A comparative study of Naive Bayes, logistic regression, and SVM for breast cancer prediction," Medical Imaging, 2020.

[13] M. Smith and J. Brown, "Understanding Misclassifications in Image Recognition: A Look at Visual Ambiguity," International Conference on Artificial Intelligence and Machine Learning. 2023.