



Smart Cafeteria

Food Waste Prediction and Risk Categorization Using Random Forest and Support Vector Machines

Swaroop Raparathi

This project report is submitted to the Department of Mathematics and Natural Sciences at Blekinge Institute of Technology in partial fulfilment of the requirements for the course ET1556.

Contact Information:

Author(s):

Swaroop Raparathi

E-mail: swra25@student.bth.se

University advisor:

Irina Gertsoyich

Department of Mathematics and Natural Sciences

Dept. of Mathematics and Natural Sciences	Internet	:	www.bth.se
Blekinge Institute of Technology	Phone	:	+46 455 38 50 00
SE-371 79 Karlskrona, Sweden	Fax	:	+46 455 38 50 57

Abstract

Context: Food waste is a big problem worldwide. It has serious economic, social, and environmental effects. In institutional cafeterias, making too much food and poor demand forecasting often result in excessive waste. Traditional methods mainly use food log data, but these methods do not predict future needs and ignore environmental factors like temperature, season, and day of the week. This shows we need a solution to help manage resources sustainably.

Research Method: This study introduces a Smart Cafeteria system that combines cafeteria food log data, including attendance, food items, and amounts of waste, with environmental information such as temperature, season, and day of the week. The research method used is a formal experimental approach that combines cafeteria food log data with environmental factors like temperature, season, and weekday. The dataset was preprocessed through feature extraction and encoding. Random Forest and SVM models were trained and predict food waste and classify risk levels based on probabilities.

Objectives:

Collect and integrate cafeteria food log data and environmental data.

Preprocess and prepare data for analysis.

Develop predictive models for food waste and risk categorization.

Train and evaluate model performance using appropriate metrics.

Provide a framework to help institutional cafeterias improve forecasting and minimize waste.

Result and Analysis: The results show that adding environmental factors greatly improves prediction accuracy. The Random Forest model used with environmental data achieved better results, with an MSE of 0.2699 and an R^2 of 0.9749, which indicates high reliability in predictions. For risk categorization, RF reached an accuracy of 0.8372, a precision of 0.8396, and a recall of 0.8372, outperforming the SVM model. Adding environmental features improved the RF's performance by about 24 percent across important metrics.

Conclusion: The study shows that integrating environmental data with cafeteria records significantly enhances machine learning accuracy in predicting and classifying food waste. The Random Forest model performed best, effectively capturing complex relationships and supporting the development of smart systems for waste reduction and sustainability.

Keywords: Smart Cafeteria, Food Waste Prediction, Risk Categorization, Environmental Data, Random Forest

Contents

Abstract	iii
List of Figures	iv
List of Tables	iv
1 Introduction	2
1.1 Background	2
1.2 The scope	2
1.3 Research Questions	3
1.4 Outline	3
1.5 Ethical, societal and sustainability aspects	4
2 Related Work	5
2.1 Research gap	6
3 Method	7
3.1 Evaluation Methodology	8
3.1.1 Regression Evaluation Metrics	8
3.1.2 Classification Evaluation Metrics	9
4 Results and Analysis	10
5 Discussion	12
5.1 Discussion of research questions	12
5.2 Relation to expectations and literature	13
5.3 Defending and questioning the findings	13
5.4 Limitations and their implications	13
6 Conclusions and Future Work	14
6.1 Conclusion	14
6.2 Future work	14
References	15
A Supplemental Information	16

List of Figures

4.1	Impact of environmental data in food waste prediction	10
4.2	Impact of environmental data for risk categorization	10
4.3	Food preparation recommendation output showing predicted waste and risk level	11

List of Tables

Food waste is a serious global issue that has significant economic, environmental, and social effects. In institutional cafeterias, like those in universities, hospitals, and corporate offices, overproduction and inaccurate demand forecasting are major causes of waste. Traditional systems mainly rely on past consumption records, but they often neglect environmental factors like temperature, season, or day of the week. This oversight makes forecasting unreliable, resulting in unnecessary waste, wasted resources, and higher costs.

Recent progress in machine learning offers ways to solve this problem. Models like Random Forest (RF) and Support Vector Machine (SVM) are well-known for effectively handling complex classification and regression tasks. By combining cafeteria food log data with environmental variables, a Smart Cafeteria system can more accurately predict food demand, identify risk levels, and promote sustainable resource management.

1.1 Background

Previous studies have addressed food waste measurement and reduction through different approaches. Beery et al. (2024) developed a smart compost bin that uses sensors to quantify waste in real time, emphasizing accurate measurement and visualization but lacking predictive capabilities [2]. Similarly, Krupp et al. (2022) introduced Campus Plate, a smartphone-based system for redistributing surplus food and raising awareness on college campuses [4]. While effective in promoting waste reduction, both approaches remain largely reactive and do not provide proactive forecasting or risk categorization.

1.2 The scope

The scope includes preparing and processing various data sources by creating features suited for machine learning models. This ensures effective dataset handling and analysis. It also evaluates the performance of the RF and SVM models using important metrics like accuracy, precision, recall, R^2 , and Mean Squared Error (MSE), with cross-validation to ensure results are reliable [1]. This phase also involves providing practical recommendations for cafeteria managers. These recommendations aim to improve menu planning and cut down on waste, offering useful insights that balance

operational efficiency with environmental concerns, although the study remains a proof-of-concept within an academic timeline.

The research does not include real-time IoT integration, recognizing this as a possible area for future work. It focuses on data-driven predictive modeling without hardware implementation. Challenges such as limited data set diversity and synchronization issues are managed to keep the focus on developing and evaluating the model.

1.3 Research Questions

- **RQ1:** How can machine learning models (Random Forest and Support Vector Machines) be used to predict food waste in cafeterias?
- **RQ2:** Does incorporating environmental data improve prediction accuracy compared to using only food-related data?
- **RQ3:** Which model performs better for food-waste prediction and risk categorization: Random Forest or Support Vector Machine?

1.4 Outline

The Introduction describes the overview of the global food waste issue, existing solutions, the Smart Cafeteria concept, and provides a brief mention of ethical, societal, and sustainability considerations.

The Related Work section reviews previous studies on food waste prediction techniques and explores SVM and RF algorithms.

The Method chapter explains the Smart Cafeteria System’s data collection, preprocessing, and the use of Random Forest and SVM models for predicting and classifying food waste.

The Results chapter presents the outcomes, showing that adding environmental data significantly improves the accuracy of food waste prediction and risk classification, with Random Forest performing best.

The Discussion chapter interprets the findings, answers the research questions, relates them to previous studies, and highlights limitations and implications for future improvement.

The Conclusion and Future Work describes Combining environmental data with cafeteria records improves food waste prediction, especially with Random Forest; future work includes IoT sensors, larger datasets, and advanced algorithms.

1.5 Ethical, societal and sustainability aspects

Ethically, the focus is on fair predictions. If the dataset lacks diversity, it could bias results toward specific foods, so a varied dataset is key for equity. Plus, accurate and up-to-date data ensures trustworthy decisions. They should promote fairness and workforce support, avoiding biased predictions and enabling reskilling rather than replacing staff.

Societally, accessibility matters; big institutions might benefit more, so affordable options and training can help smaller ones. Automation could cut jobs, but it should support workers by allowing reskilling and improving efficiency.

For sustainability, better forecasting reduces food, water, and energy waste while promoting eco-friendly habits. Using seasonal data helps cafeterias adjust to climate change for long-term resilience.

Chapter 2

Related Work

Beery et al. (2024) proposed a smart compost bin system that measures food waste in real-time at the consumer level. It uses sensors for multimodal interaction to quantify waste in institutional settings. The system offers data visualization to promote waste reduction and underscores the importance of accurately measuring waste [2]. However, it does not include predictive modeling or risk categorization. In contrast, our Smart Cafeteria project builds on this work by integrating cafeteria and environmental data into Random Forest (RF) and Support Vector Machine (SVM) models [5]. This allows us to predict food waste and categorize risk levels, enabling proactive waste reduction strategies rather than a reactive measurement approach. Additionally, we factor in environmental elements like temperature and season, adding context that Beery et al. overlooked, which makes our system more applicable to the changing conditions of cafeteria environments.

Krupp et al. (2022) introduced Campus Plate, a smartphone-based system designed to cut down on food waste and tackle food insecurity on college campuses. By connecting students with surplus food, it raises awareness and helps with redistribution, depending on user-reported data for managing waste after consumption [4]. While Campus Plate aims to reduce food waste in university cafeterias like we do, it takes a different path. Smart Cafeteria uses machine learning to predict waste before it happens, which allows for preventive actions like better menu planning. Our proactive, data-driven method uses environmental factors to improve prediction accuracy, which Campus Plate does not address. However, our system could support Campus Plate by offering predictive insights to direct its redistribution efforts, creating a more complete waste management strategy.

Cervantes et al. (2020) conducted a survey of SVM classification, highlighting its applications, drawbacks, and trends. The study shows that SVM works well in high-dimensional spaces because of its principle of margin maximization [3]. This makes it suitable for tasks like risk categorization, though it also points out issues like computational complexity and the need for careful tuning of hyperparameters. This survey informs our use of SVM to predict food waste and risk categorization in the Smart Cafeteria. Unlike Cervantes et al., who discuss SVM's general applications, our project specifically applies SVM to cafeteria data together with environmental variables. We tackle SVM's computational challenges by using modern resources and cross-validation to optimize performance. The insights from the survey guide our model selection, but adding RF for comparison brings a new angle not covered

in their work.

Andersen (2021) outlined essential performance metrics for evaluating machine learning models, including accuracy, precision, recall, Mean Squared Error (MSE), and R^2 [1]. These metrics are regarded as critical for assessing both regression and classification models. In the Smart Cafeteria project, the evaluation stage is guided by Andersen’s framework, through which Random Forest (RF) and Support Vector Machine (SVM) models are compared on a standardized basis. While a general overview of evaluation techniques is provided by Andersen, the metrics are applied in this project specifically to the prediction of food waste (regression) and the categorization of risk levels (classification). In this way, RQ3, which asks which model performs better for prediction and categorization, is rigorously addressed. By Andersen’s standardized evaluation approach, the reliability of the results is strengthened, and alignment with the project’s goal of ensuring robust model accuracy is achieved.

2.1 Research gap

Although several studies have looked at measuring food waste and raising awareness, important gaps still exist in predictive modeling and risk categorization. Current systems like Campus Plate mainly focus on showing waste and increasing awareness, but they do not provide real-time predictions or risk categorization [4]. Additionally, most research relies heavily on cafeteria food log data and often overlooks the impact of environmental factors such as season, temperature, and day of the week. These factors greatly influence consumption patterns and levels of food waste. By addressing these gaps and incorporating predictive analytics, risk categorization, and environmental data, we can achieve better forecasting and more sustainable resource management in institutional cafeterias.

This study presents a Smart Cafeteria System that integrates cafeteria food log data such as daily attendance, prepared food items, and waste quantities with environmental factors including temperature, season, and day of the week. The datasets were collected from publicly available sources: the Kaggle dataset Stockholm Cafeteria Dataset and the GitHub repository Stockholm Environmental Data 2020. The research uses a formal experimental method to evaluate how well machine learning algorithms can estimate food waste and categorize risk levels. The dataset was cleaned and prepared for model training. Temporal features, such as day of the year and week of the year, were extracted from the date. Categorical variables, like day and season, were converted into labels. Seasonal indicators were created by grouping food and waste amounts for winter and summer, allowing the models to capture seasonal consumption and waste patterns.

Two supervised machine learning algorithms, Random Forest (RF) and Support Vector Machine (SVM), were used for both regression and classification tasks. The regression models estimated total food waste in kilograms, while the classification models assigned waste risk levels as high or low. Each model was trained with an 80/20 train-test split and validated through cross-validation to ensure reliability and avoid overfitting. Feature scaling was done using standard normalization to keep input ranges consistent across all models. Model performance was assessed using metrics such as Mean Squared Error, R^2 , accuracy, precision, and recall, providing a quantitative evaluation of each model's prediction ability.

The experiment was performed under two scenarios: one that excluded environmental variables and another that included them. The environmental features covered average temperature, day of the week, season, day of the year, and week of the year. By comparing model performance in these two scenarios, the study looked at how environmental information affected prediction accuracy and model reliability. Both RF and SVM algorithms were tested under the same conditions to ensure a fair comparison. Their outputs were examined using statistical measures and cross-validation results to find which model had greater predictive power and generalization ability. Additionally, a seasonal impact analysis was conducted, calculating the average waste per food item across spring, summer, and winter to identify categories with larger seasonal variations. This blend of predictive modeling and seasonal analysis lays the groundwork for creating data-driven recommendations to improve food preparation and reduce waste in institutional cafeterias.

To ensure the research is valid and reliable, The experiment used consistent pre-processing steps, model parameters, and evaluation metrics for both Random Forest and SVM methods. We applied cross-validation to reduce random variation and to confirm that the models performed well on new data. Including both environmental and non-environmental datasets improved internal validity by enabling direct comparison under controlled conditions. However, the study recognizes that results might differ with larger or more varied datasets. Future work could improve external validity by broadening data collection across different cafeteria settings.

3.1 Evaluation Methodology

To assess the performance of the proposed models, both regression and classification evaluation metrics were employed. Regression metrics were used to evaluate food-waste prediction accuracy, while classification metrics were used to evaluate risk-level categorization. The following equations describe each metric.

3.1.1 Regression Evaluation Metrics

1. Mean Squared Error (MSE): Mean Squared Error measures the average squared difference between the predicted values and the actual (true) values.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.1)$$

where:

- n = number of observations
- y_i = actual (true) value of food waste for the i^{th} instance
- \hat{y}_i = predicted value of food waste for the i^{th} instance

2. Coefficient of Determination (R^2 Score): The R^2 score evaluates how well the predicted values approximate the actual values. A higher R^2 indicates better model performance.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.2)$$

where:

- \bar{y} = mean of the actual values
- $(y_i - \bar{y})^2$ = total variance in the data

3.1.2 Classification Evaluation Metrics

1. Accuracy: Accuracy measures the proportion of correctly classified instances over the total number of instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.3)$$

where:

- TP = True Positives (correctly predicted high-risk cases)
- TN = True Negatives (correctly predicted low-risk cases)
- FP = False Positives (incorrectly predicted high-risk cases)
- FN = False Negatives (low-risk cases incorrectly predicted)

2. Precision: Precision evaluates the proportion of correctly predicted positive (high-risk) cases among all predicted positive cases.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.4)$$

3. Recall: Recall (also known as Sensitivity) measures the proportion of actual positive (high-risk) cases that were correctly identified by the model.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.5)$$

Chapter 4

Results and Analysis

The following results are obtained by our experiment:

```
=== Regression Performance: for food waste prediction ===
      Model    MSE    R2
rf_reg_NoEnv 2.7979 0.7400
svm_reg_NoEnv 4.2188 0.6080
rf_reg_Env   0.2699 0.9749
svm_reg_Env   0.9673 0.9101
```

Figure 4.1: Impact of environmental data in food waste prediction

Figure 4.1 shows that environmental factors strongly affect how accurately we can predict food waste in the smart cafeteria system. Models that did not include environmental data produced decent results but were less accurate, with higher MSE and lower R^2 values. When environmental data were included, both models performed much better. The random forest model with environmental data achieved the best results, with a very low MSE of **0.2699** and a high R^2 of **0.9749**. This means that it can predict food waste very accurately. The SVM model also showed significant improvement with environmental data. This demonstrates that environmental data are crucial for predicting food waste accurately and that the random forest model excels in this area.

```
=== Classification Performance: for risk categorization ===
      Model  Accuracy  Precision  Recall
rf_clf_NoEnv    0.6744    0.6744    0.6744
svm_clf_NoEnv    0.6047    0.3656    0.6047
rf_clf_Env      0.8372    0.8396    0.8372
svm_clf_Env      0.6047    0.3656    0.6047
```

Figure 4.2: Impact of environmental data for risk categorization

Figure 4.2 shows how environmental factors affect the classification performance for risk categorization using Random Forest (RF) and Support Vector Machine (SVM) models. The results clearly show that adding environmental data improves performance across all metrics. The random forest model with environmental data achieved the best results, with an accuracy of **0.8372**, precision of **0.8396**, and recall of **0.8372**. Compared to the model without environmental data, this represents an improvement of about **24%** in accuracy, precision, and recall. This shows that the model can classify risk levels much more effectively when environmental

information is included. In contrast, the SVM model showed no improvement when environmental data were added, indicating that Random Forest benefits more from environmental features. Overall, this analysis highlights that environmental data play an important role in improving classification accuracy, and the random forest model performs best for risk categorization in the smart cafeteria system.

```
=== Food Preparation Recommendation ===  
Enter food item for recommendation (e.g., Rice, Chicken, Vegetables): Rice  
  
Prediction for 30-08-2020:  
Predicted Rice Waste: 12.29 kg  
Average Rice Prepared: 116.74 kg  
Recommended Rice to Prepare: 104.45 kg  
Waste Risk Level: Low (Probability: 83.00%)
```

Figure 4.3: Food preparation recommendation output showing predicted waste and risk level

Figure 4.3 shows an example output of the food preparation recommendation system. The system allows users to input any food item (such as rice, chicken, vegetables, etc.) and provides predictions related to food preparation and waste. It displays the predicted food waste, the average amount typically prepared and the recommended amount to cook to minimize waste. In addition, it estimates the level of waste risk (Low or High) along with a probability score. This enables efficient meal planning and helps reduce food waste.

5.1 Discussion of research questions

RQ1: How can machine learning models predict food waste in cafeterias?

Both the Random Forest (RF) and Support Vector Machine (SVM) models predicted total food waste and classified risk levels within the smart cafeteria system. They captured complex relationships between menu composition, environmental conditions, and seasonal changes. However, their predictive power varied widely. The RF model performed the best, with a very low Mean Squared Error (MSE) of 0.2699 and a high R^2 of 0.9749. This shows its strong ability to model nonlinear interactions and the benefits of ensemble learning. In comparison, the SVM model performed reasonably well but was less accurate. This confirms that tree-based ensemble methods are better for predicting food waste in this dataset.

RQ2: Does adding environmental data improve the prediction compared to using only food data?

Including environmental features, such as temperature, day of the week, and seasons, improved both regression and classification results. When we added environmental data, the RF model's R^2 increased significantly, while the MSE decreased. This shows a better fit and higher predictive accuracy. For risk categorization, the RF model's accuracy, precision, and recall improved by about 24 percent, reaching 0.8372, 0.8396, and 0.8372, respectively. The SVM model also gained in regression, but showed little to no improvement in classification performance. These findings confirm that environmental context strongly influences food waste patterns and improves prediction quality.

RQ3: Which model performs better for food-waste prediction and risk categorization?

The results show that the Random Forest model consistently performed better than SVM in both regression and classification tasks with environmental data included. The RF model's structure helped it manage data variety and nonlinear relationships, leading to improved predictive accuracy and reliability. Although SVM saw some benefits from environmental inputs, it did not generalize as effectively in the classification task because it is highly sensitive to nonlinear and noisy data, requires

extensive hyperparameter tuning, and struggles with mixed feature types.

5.2 Relation to expectations and literature

The results align with prior research emphasizing that hybrid machine-learning pipelines can improve sustainability decisions. Similar to campus plate [4], this study demonstrates the utility of integrating digital data for waste reduction but extends it by offering predictive capability. The superiority of RF and SVM is consistent with earlier findings on their robustness for nonlinear, high-dimensional datasets [3]. However, the modest R^2 values suggest that cafeteria data alone cannot capture all determinants of waste, echoing [1] that model evaluation metrics must be interpreted in context rather than in absolute terms.

5.3 Defending and questioning the findings

On one hand, the models' high predictive accuracy and better performance with environmental data strongly support the research hypothesis that contextual factors are important in predicting food waste. On the other hand, the small improvement margin for SVM in classification suggests that not all algorithms gain equally from environmental variables. While the Random Forest model performed well, some limitations, like the lack of real-time attendance data, portion sizes, or special event indicators, may still limit the overall predictive ability. Therefore, future systems could include these additional factors to improve prediction accuracy.

5.4 Limitations and their implications

The study faced several limitations that may have influenced the results. The dataset scope was limited, as it may not fully represent all cafeterias, environmental conditions, or timeframes, affecting the generalization of the models. Synchronizing cafeteria and environmental data proved challenging, and misalignment between datasets could reduce prediction accuracy. Additionally, missing or inconsistent sensor data, due to failures or calibration errors, introduced partial or noisy information that could impair classification and estimation. The availability of labeled risk levels was limited, constraining model training efficiency, while variability in attendance or menu changes could make predictions less precise under unusual or rare conditions. Overall, these limitations highlight areas for improvement in future research.

Chapter 6

Conclusions and Future Work

6.1 Conclusion

Food waste in cafeterias creates serious environmental, economic, and social problems. This study showed that combining environmental data with cafeteria food logs greatly improves the accuracy of food waste prediction and risk classification. Traditional models, which rely only on past consumption data, were not enough. Including environmental factors like temperature and season significantly boosted predictive performance.

Finally, the study shows that using machine learning along with environmental awareness can greatly enhance food demand forecasting and reduce waste.

6.2 Future work

In future work, we can use real-time IoT sensors in cafeterias to monitor food consumption, temperature, and waste. This will help predictive models continuously improve their accuracy. We can also expand datasets to include multiple cafeterias from different regions and seasons to improve model generalization and reliability. Additionally, exploring new machine learning techniques and developing decision-support tools could offer useful insights for menu planning, procurement, and waste reduction. This will further connect research with practical implementation. It can also include usercentered evaluations, such as cafeteria staff surveys and real world pilot testing, to assess practical effectiveness and usability.

References

- [1] G. Andersen. Essential metrics for evaluating machine learning models - what you need to know. Section: ML developers questions. [Online]. Available: <https://moldstud.com/articles/p-essential-metrics-for-evaluating-machine-learning-models-what-you-need-to-know>
- [2] A. J. Beery, D. W. Eastman, J. Enos, W. Richards, and P. J. Donnelly, “Smart compost bin for measurement of consumer food waste,” in *Companion Proceedings of the 26th International Conference on Multimodal Interaction*, ser. ICMI Companion '24. Association for Computing Machinery, pp. 100–107. [Online]. Available: <https://dl.acm.org/doi/10.1145/3686215.3686216>
- [3] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, “A comprehensive survey on support vector machine classification: Applications, challenges and trends,” vol. 408, pp. 189–215. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0925231220307153>
- [4] B. Krupp, J. Gersey, and F. Lebo, “Campus plate: connecting students on college campuses to reduce food waste and food insecurity,” in *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, ser. RACS '22. Association for Computing Machinery, pp. 172–177. [Online]. Available: <https://doi.org/10.1145/3538641.3561506>
- [5] D. Yuan, J. Huang, X. Yang, and J. Cui, “Improved random forest classification approach based on hybrid clustering selection,” in *2020 Chinese Automation Congress (CAC)*, pp. 1559–1563, ISSN: 2688-0938. [Online]. Available: <https://ieeexplore.ieee.org/document/9326711/>

