Analyzing FDA Food Recall Patterns Using Machine Learning: Predicting Outcomes and Revealing Insights

Olapeju Esuola, MSBA

University of Louisville

May 01, 2025

Abstract

Food is essential for human survival, and the global food industry plays a critical role in ensuring its availability. However, with this responsibility comes the need for rigorous regulation to ensure that harmful or contaminated food products do not reach consumers. Regulatory bodies like the U.S. Food and Drug Administration (FDA) must maintain strong oversight to prevent unsafe products from entering the market and when failures occur, to ensure that recalls are carried out swiftly and effectively.

The objective of this project is to analyze FDA food recall data from 2017 to 2025, investigate the reasons behind recalls, identify patterns in terminated recalls, and develop predictive models to estimate the likelihood of a recall being terminated. The dataset, obtained from the official FDA website, includes recall dates, product descriptions, recall reasons, company names, and recall statuses.

The project employs exploratory data analysis (EDA) to uncover patterns, clustering techniques to group similar recall reasons, forecasting methods to predict future recall trends, and machine learning classification to predict recall termination outcomes. Key findings reveal that the most common recall reasons include Listeria contamination, undeclared allergens, and other contamination issues. Additionally, company name and year were identified as the most influential predictors of recall termination.

This work provides actionable insights for manufacturers and regulators, offering a data-driven foundation for improving food safety oversight, enhancing recall management, and ultimately protecting public health.

Keywords: FDA food recalls; machine learning; predictive modeling; recall termination; clustering; forecasting; food safety; regulatory compliance; data analysis; public health

Introduction

Food safety is a cornerstone of public health, and when food products pose a risk to consumers, regulatory agencies must act swiftly to mitigate harm. In the United States, food recalls are managed under the oversight of both the Food and Drug Administration (FDA) and the Food Safety and Inspection Service (FSIS). Companies may voluntarily recall products from the market when they identify issues, especially for FSIS-regulated foods, which are typically inspected both during production and labeling. As the National Agricultural Law Center (NALC) notes, the FSIS voluntary recall process serves largely as a backup to catch food safety hazards that slip through pre-market inspections. For FDA-regulated food, however, the voluntary recall process plays a primary role, as the FDA does not routinely conduct pre-market inspections of every product or label. According to the FDA, common reasons for food recalls include contamination with harmful microorganisms (such as bacteria, viruses, or parasites), the presence of foreign objects like glass or metal fragments, and failure to properly declare major allergens like peanuts or shellfish on product labels. FSIS similarly lists illness outbreaks, undeclared allergens, uninspected production or imports, drug residues, Shiga Toxin-Producing Escherichia coli (STEC), Listeria monocytogenes, Salmonella contamination, and foreign matter as primary triggers for recalls (Phoenix, 2025).

Understanding why and how recalls happen is critical not only for regulatory compliance but also for improving food industry practices and safeguarding public health. Recalls have profound impacts, ranging from direct health consequences for consumers to reputational and financial damage for manufacturers. Yet, despite the availability of detailed recall records, relatively few studies have applied machine learning approaches to systematically analyze patterns and predict outcomes within FDA recall data.

This study aims to analyze FDA food recall data to identify key patterns and develop predictive models that estimate the likelihood of a recall being terminated. By examining recall reasons, product descriptions, company profiles, and time trends, the study seeks to improve understanding of the recall landscape and support more proactive risk management.

Contributions

This research makes three key contributions to the field:

- It provides a comprehensive exploratory data analysis (EDA) of FDA food recalls between 2017 and 2025.
- It uses clustering techniques to group similar recall reasons, offering a clearer view of risk categories.
- It incorporates forecasting techniques to estimate future recall trends.
- It applies predictive modeling to estimate recall termination outcomes, highlighting influential features such as company name and year, and offering actionable insights for regulators and manufacturers.

Literature Review

Food recalls are critical interventions to protect public health, often prompted by contamination, mislabeling, or the presence of allergens. Understanding and predicting these events can enhance food safety measures and regulatory response. Recent advancements in machine learning (ML) and artificial intelligence (AI) have opened new avenues for analyzing recall data and forecasting potential risks.

One study developed an ML algorithm to predict FDA medical device recalls by analyzing publicly available data sources, achieving high sensitivity and specificity up to 12 months before actual recalls occurred (National Library of Medicine, 2023). While this research focused on medical devices, the methodology demonstrates the potential of ML in anticipating regulatory actions across sectors.

In the realm of food safety, researchers applied deep learning techniques to forecast food recalls, using time series data to predict future incidents (ResearchGate, 2023). This approach underscores the applicability of advanced analytics in preempting food safety issues and improving recall management.

Additionally, another study explored the use of consumer product reviews to detect unsafe food products. By applying ML to Amazon.com reviews, the study identified reports of unsafe foods, suggesting that consumer feedback can serve as an early warning system for potential recalls (National Library of Medicine, 2020).

Alongside these academic contributions, regulatory reports highlight key drivers of food recalls. According to the U.S. Food and Drug Administration (FDA, 2015), common causes include contamination by microorganisms such as Listeria monocytogenes, Salmonella, and Escherichia coli, the presence of foreign objects, and undeclared allergens like peanuts and shellfish. Similarly, FSIS lists illness outbreaks, uninspected production, drug residues, STECs, and foreign matter as major reasons for recalls (Phoenix, 2025).

Collectively, these studies and reports highlight the growing role of predictive analytics in food safety. By leveraging ML and AI, stakeholders can enhance their ability to foresee and mitigate risks associated with food products, ultimately safeguarding public health.

Methodology

Research Design

This project uses a quantitative exploratory research design applying supervised machine learning and unsupervised learning to analyze and predict outcomes in FDA food recalls. The aim is twofold:
(1) uncover hidden patterns in recall reasons using clustering, and
(2) predict recall termination status using a classification model.

The structured pipeline followed:

- Data Preparation: Cleaning, standardization, feature engineering
- Data Analysis: Exploratory analysis and visual discovery
- Model Development: Clustering, forecasting, and classification
- Interpretation: Translating model outputs into meaningful insights

Data Collection

The dataset was sourced from the FDA website and includes records from 2017 to 2025. Each record represents a recall event with these fields:

- Date – Recall announcement date
- Product Description – Name/details of product
- Brand Name – Associated brand
- Recall Reason Description – Why it was recalled

- Company Name – Recall-issuing company
- Recall Status – Active or terminated status

Data Preparation

Data Cleaning:
Removed duplicates, handled missing values, corrected inconsistent text.

Data Standardization:
Converted dates to datetime, standardized column names, cleaned text fields.

Feature Engineering:
Derived Year, Month, Description Length, and grouped recall reasons into high-level categories (e.g., Salmonella, Listeria, Allergen, Contaminant, Other).

Data Analysis

Exploratory Data Analysis (EDA):

Summary statistics → number of recalls per year, terminated vs. active breakdown

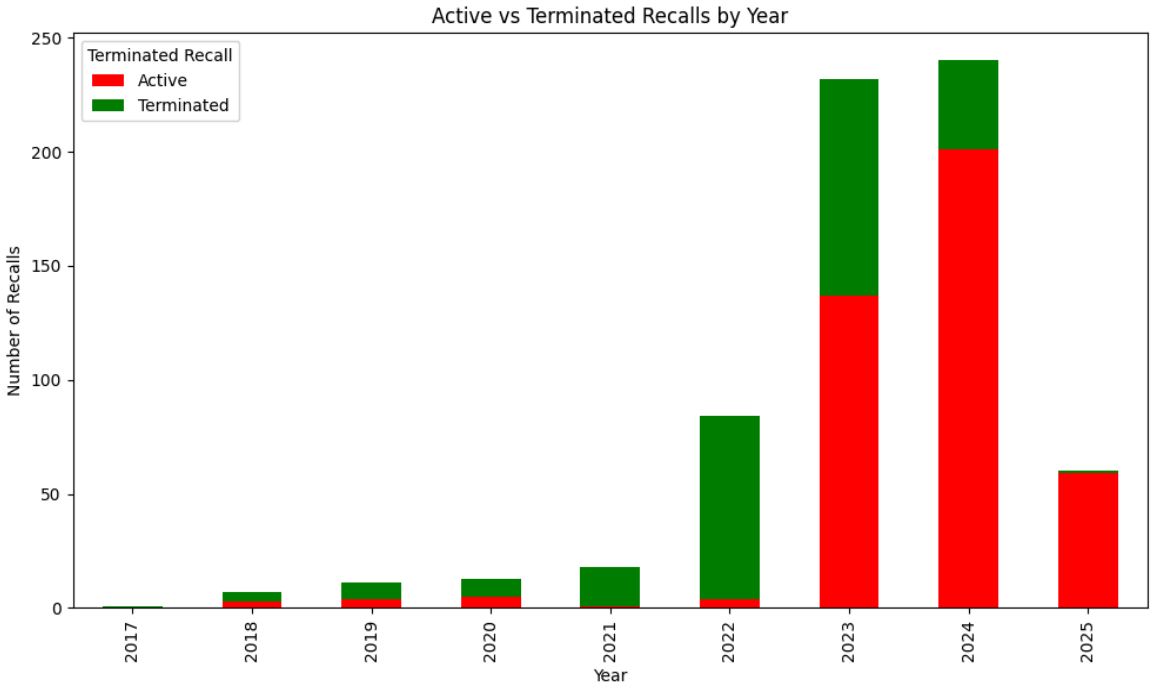


Figure 1*: Active vs Terminated Recalls by Year*

Top recall reasons

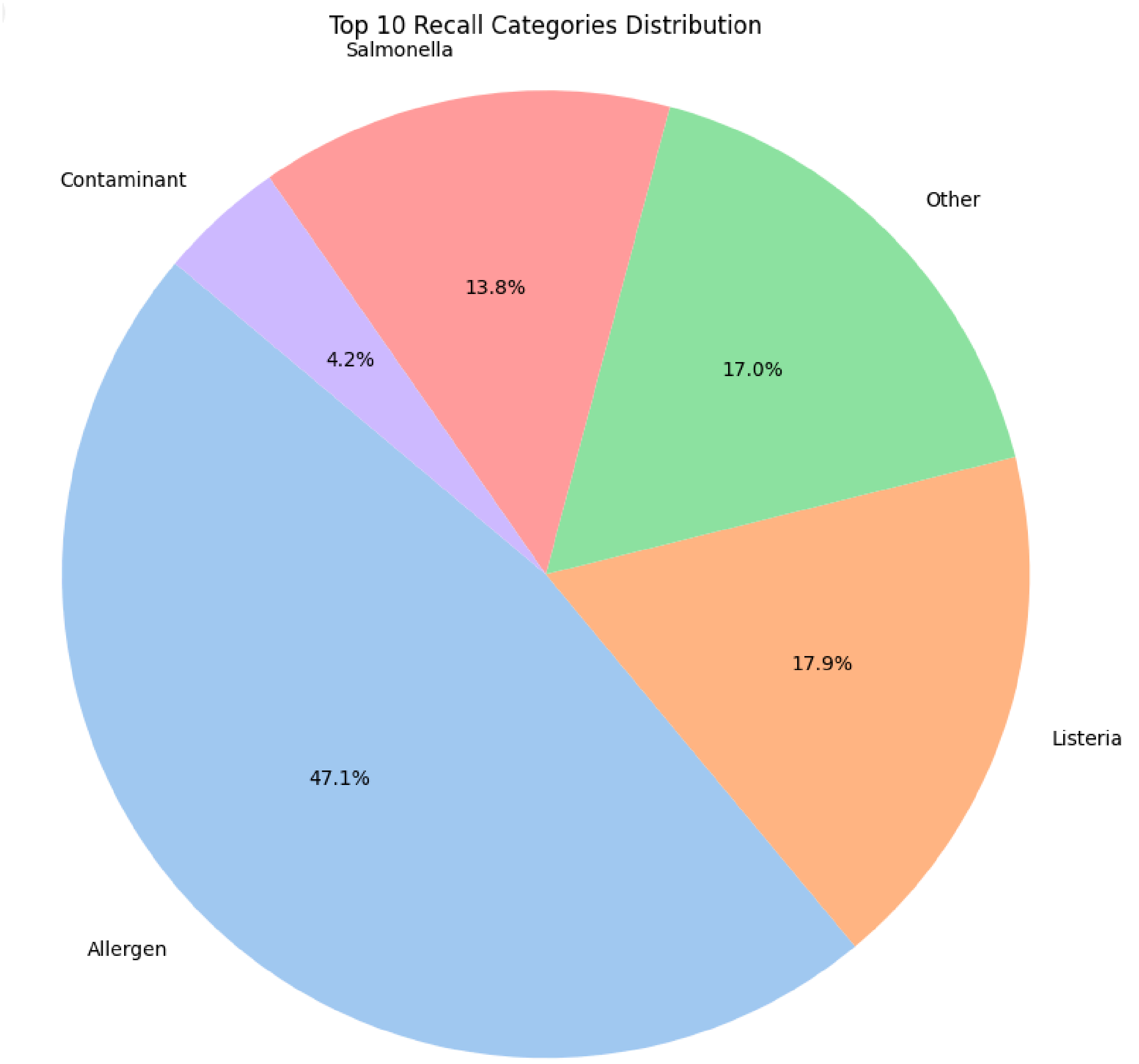


Figure 2: *Pie chart → percentage share of top recall reasons (Listeria, Salmonella, allergens) Explanation: Highlights the dominant causes driving recalls.*

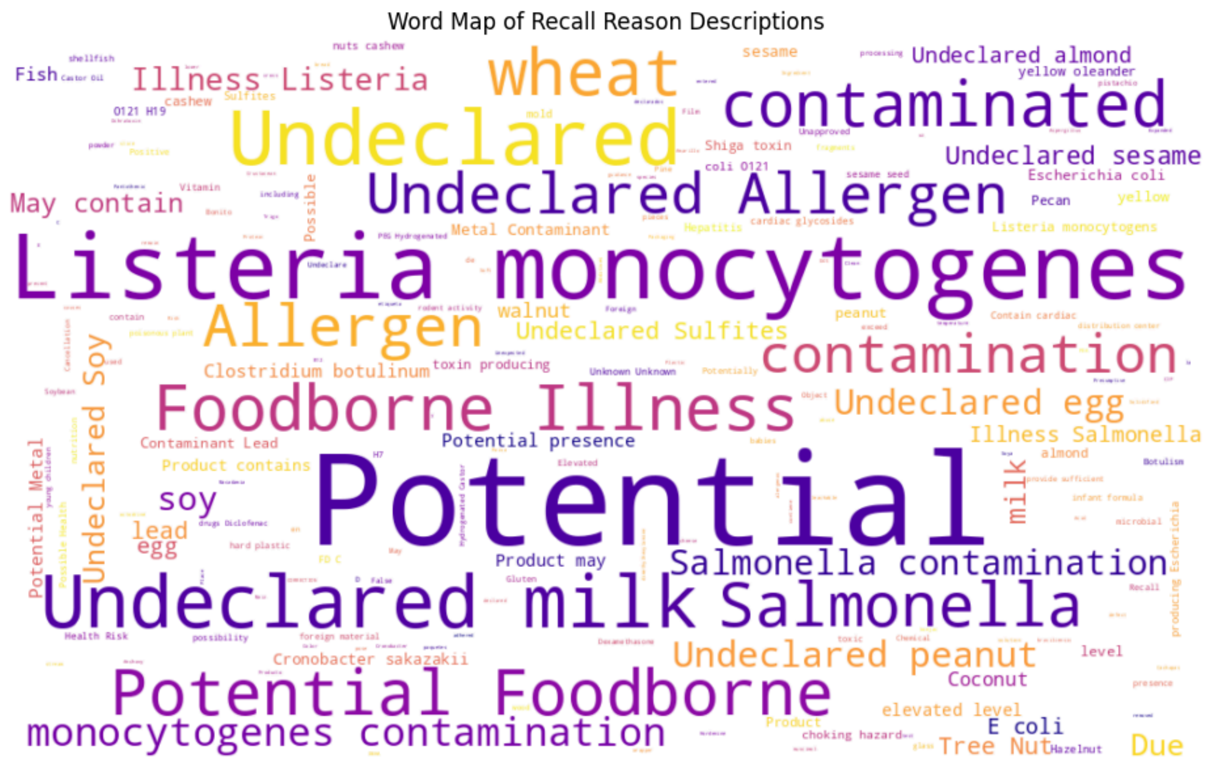


Figure 3: *Word map of recall reason descriptions.*
Explanation: Highlights most used words in recall reason description.

Top 20 food products with highest recalls

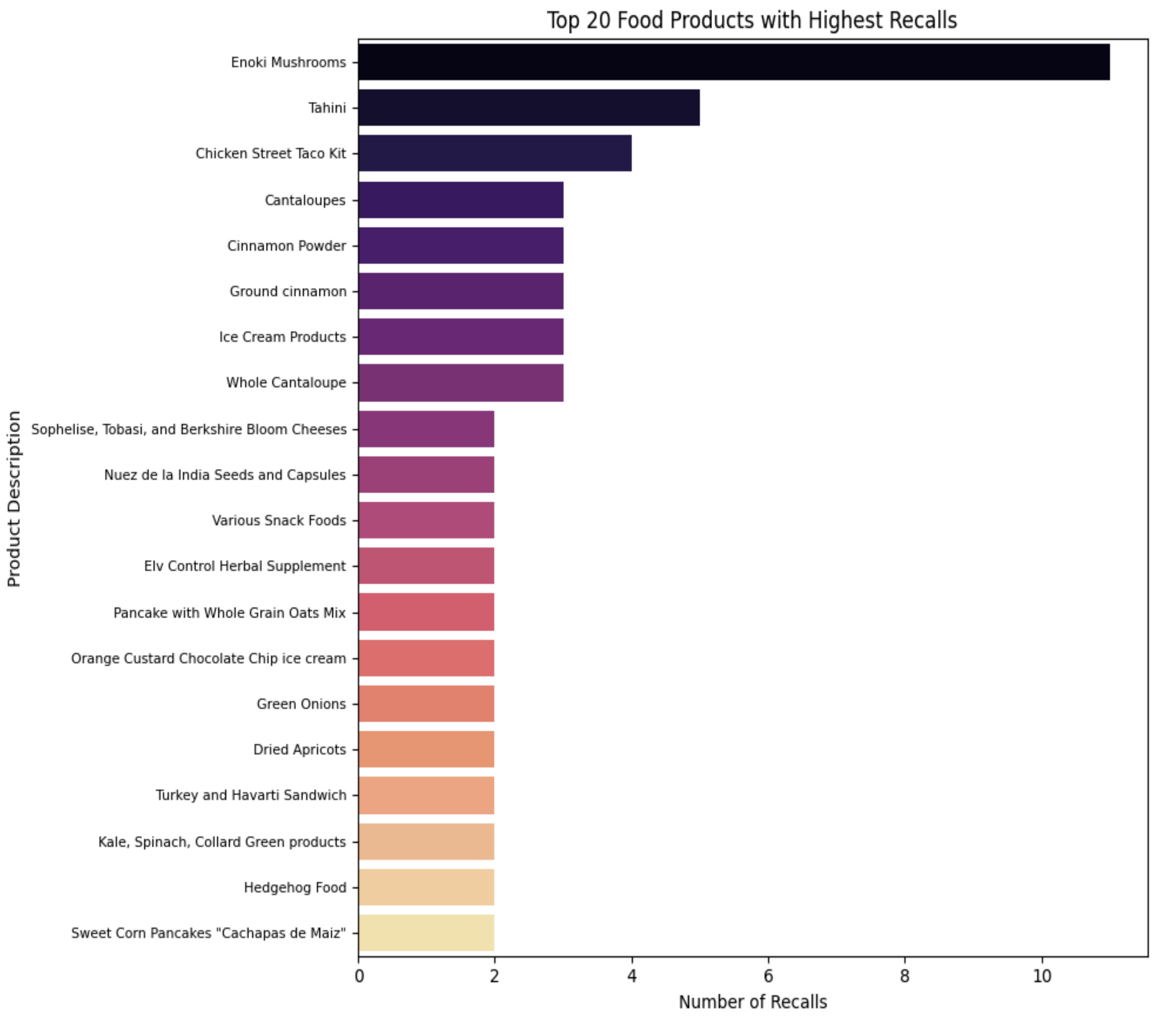


Figure 4: *Horizontal bar chart → top products*
Explanation: High-risk products.

Model Development

1. Clustering (Unsupervised Learning)

- **KMeans clustering** was applied to encoded recall reason descriptions to uncover natural groupings.

Dimensionality reduction using **t-SNE** visualized clusters in 2D.

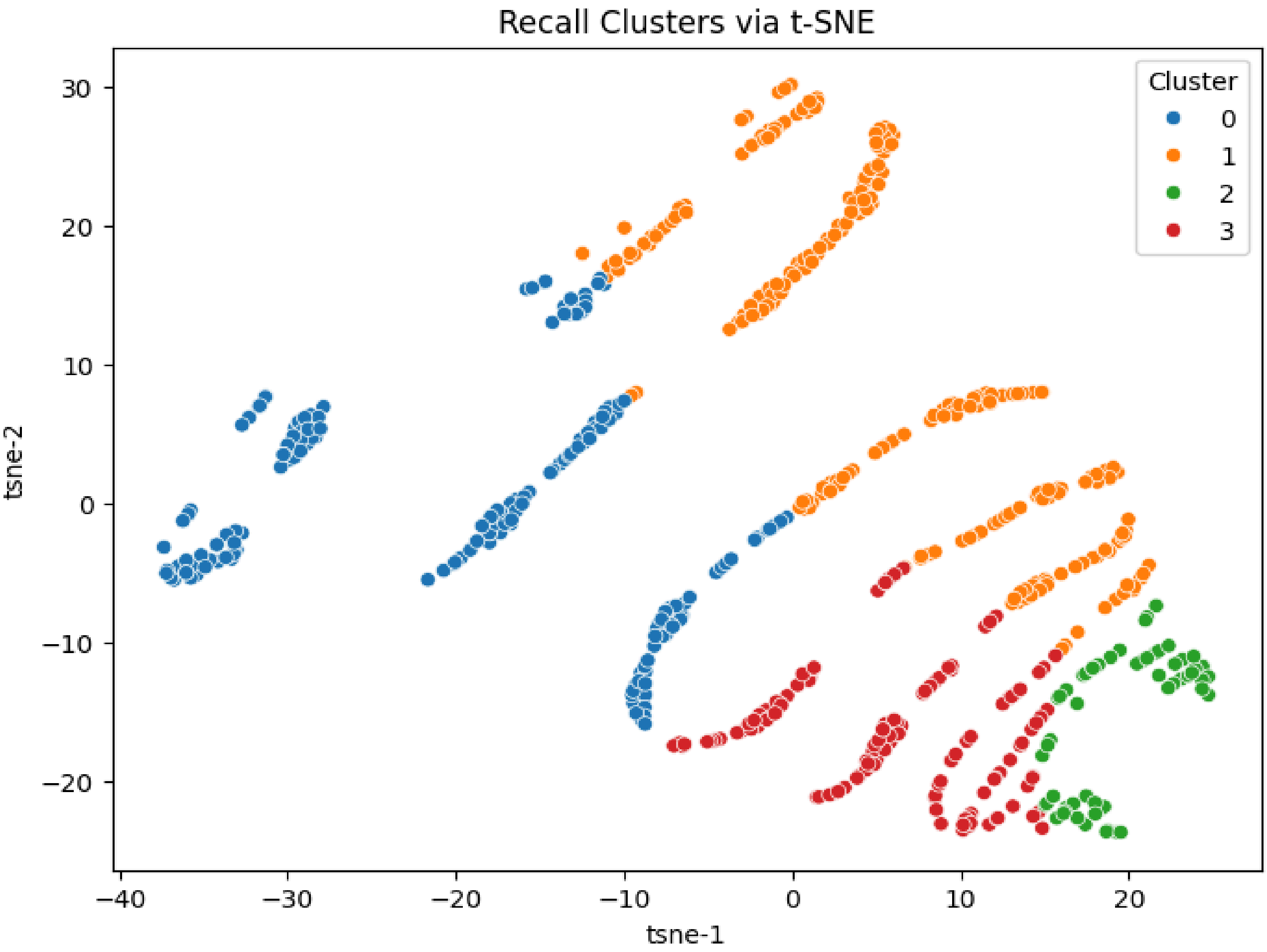
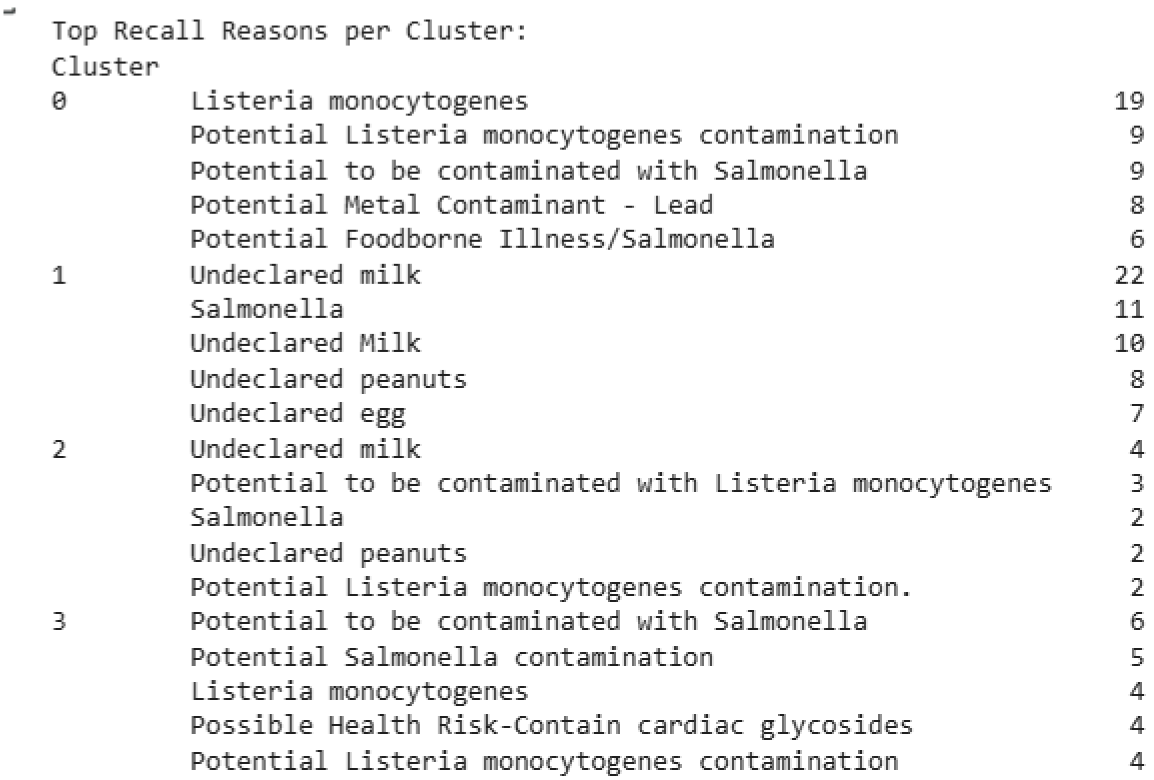


Figure 5: *t-SNE scatterplot → clusters of recall reasons*
Explanation: Groups similar recall reasons, helping regulators focus on specific risk clusters.



```
Top Recall Reasons per Cluster:
Cluster
0        Listeria monocytogenes                                      19
         Potential Listeria monocytogenes contamination               9
         Potential to be contaminated with Salmonella                 9
         Potential Metal Contaminant - Lead                           8
         Potential Foodborne Illness/Salmonella                       6
1        Undeclared milk                                             22
         Salmonella                                                  11
         Undeclared Milk                                             10
         Undeclared peanuts                                           8
         Undeclared egg                                               7
2        Undeclared milk                                              4
         Potential to be contaminated with Listeria monocytogenes     3
         Salmonella                                                   2
         Undeclared peanuts                                           2
         Potential Listeria monocytogenes contamination.              2
3        Potential to be contaminated with Salmonella                 6
         Potential Salmonella contamination                           5
         Listeria monocytogenes                                       4
         Possible Health Risk-Contain cardiac glycosides              4
         Potential Listeria monocytogenes contamination               4
```

Figure 6: *Top recall reasons per cluster*

2. Forecasting (Time Series Analysis)

- Used **Prophet** to forecast monthly recall volume over the next 6 months.
- Modeled seasonality and trend components.

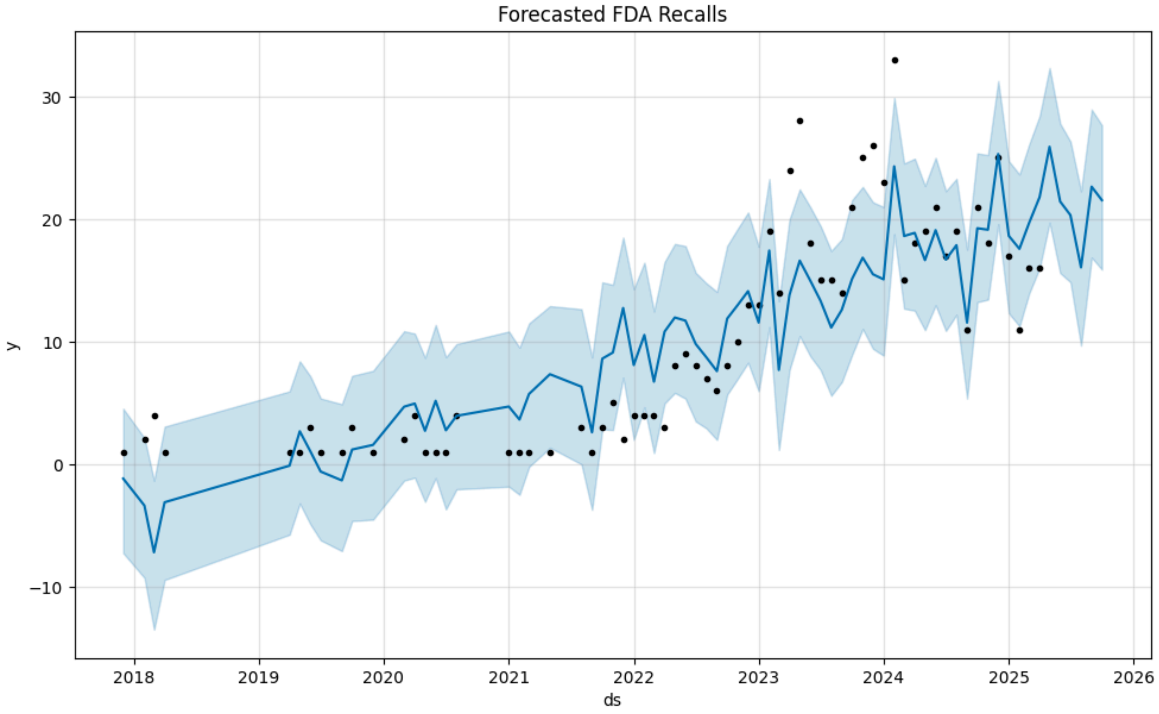
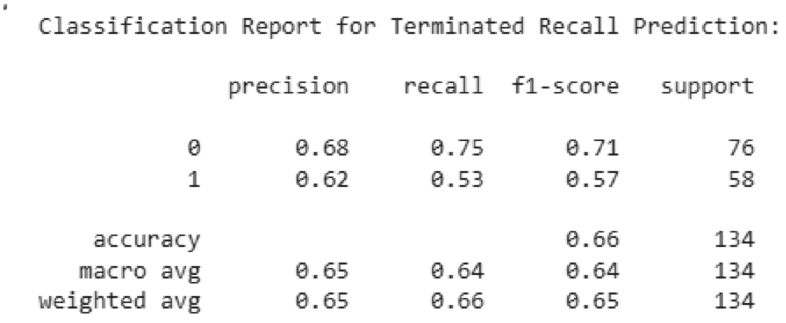


Figure 7: *Line chart → historical + predicted recall trend with confidence intervals*
Explanation: Provides early warning on expected recall surges.

3. Classification (Supervised Learning)

- Random Forest Classifier trained to predict whether a recall would be terminated.
- Features: Year, Month, Company Name, Product Description Length, Recall Reason Category.
- Evaluated using precision, recall, f1-score, accuracy.
- Identified feature importance.



```
Classification Report for Terminated Recall Prediction:

              precision    recall  f1-score   support

           0       0.68      0.75      0.71        76
           1       0.62      0.53      0.57        58

    accuracy                           0.66       134
   macro avg       0.65      0.64      0.64       134
weighted avg       0.65      0.66      0.65       134
```

Figure 8: *Classification report table → precision, recall, f1-score*

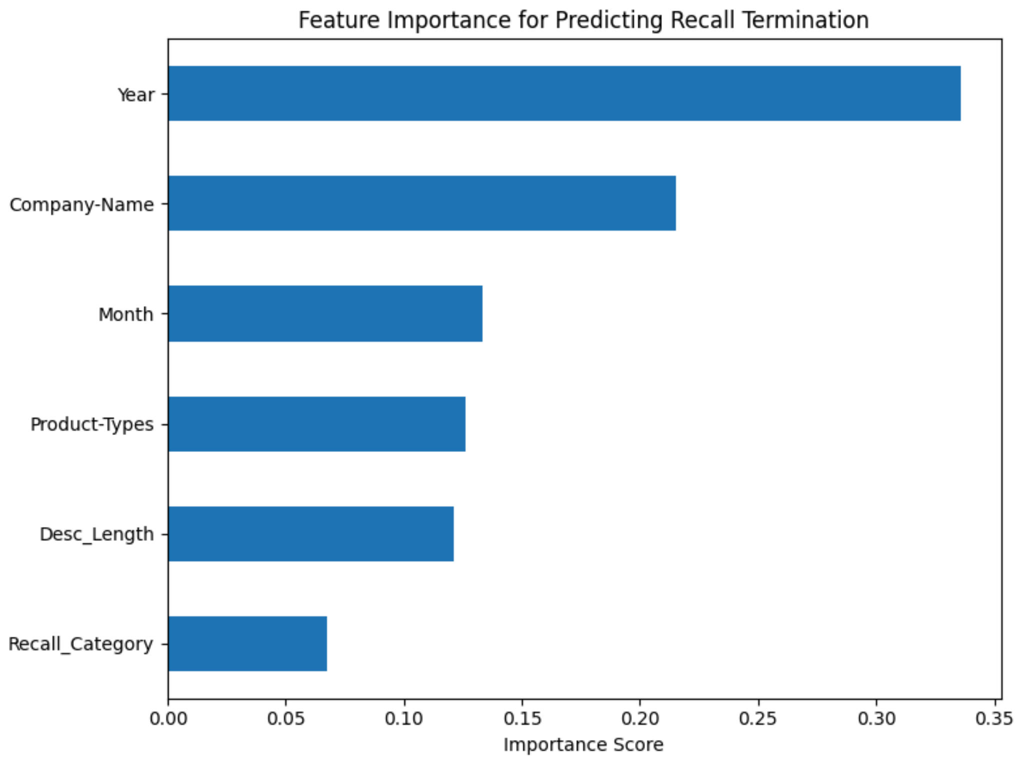


Figure 13: *Feature importance plot → ranking of features*

Explanation: Shows which factors most influence recall termination (company name and year were top drivers).

Results & Insights

Clustering Analysis: Top Recall Reasons per Cluster

The clustering analysis grouped similar recall reasons using KMeans and visualized the patterns using t-SNE. The top recall reasons within each cluster are summarized below:

- **Cluster 0:**
    - Listeria monocytogenes (19)
    - Potential Listeria monocytogenes contamination (9)
    - Potential Salmonella contamination (9)
    - Potential metal contaminant — lead (8)
    - Potential foodborne illness — Salmonella (6)
- **Cluster 1:**
    - Undeclared milk (22)
    - Salmonella (11)
    - Undeclared milk (duplicate in different records) (10)
    - Undeclared peanuts (8)
    - Undeclared egg (7)
- **Cluster 2:**
    - Undeclared milk (4)
    - Potential Listeria monocytogenes contamination (3)

  - Salmonella (2)
  - Undeclared peanuts (2)
  - Potential Listeria monocytogenes contamination (2)
- **Cluster 3:**
  - Potential Salmonella contamination (6)
  - Potential Salmonella contamination (5)
  - Listeria monocytogenes (4)
  - Possible health risk — cardiac glycosides (4)
  - Potential Listeria monocytogenes contamination (4)

Explanation: This visualization highlights how certain recall reasons—especially Listeria, Salmonella, and undeclared allergens—naturally group together, suggesting consistent underlying patterns across years and companies.

Forecasting Analysis: Predicted Recall Trends

Explanation of the plot:

- **X-axis:** Time (2018 to 2026)
- **Y-axis:** Number of recalls
- **Blue line:** Predicted trend of recalls
- **Black dots:** Actual historical data points
- **Blue shaded area:** Uncertainty/confidence interval

Insights:

- There is a clear upward trend, indicating that FDA recalls are expected to increase over time.
- The model predicts especially high recall counts after 2023.
- The widening blue band in later years reflects increasing uncertainty in long-range predictions.

Classification Analysis: Terminated Recall Prediction

Model: Random Forest Classifier

Key performance metrics:

- **Class 0 (Active recalls):** precision 0.68, recall 0.75, f1-score 0.71
- **Class 1 (Terminated recalls):** precision 0.62, recall 0.53, f1-score 0.57
- **Overall accuracy:** 66%
- **Macro average:** Equal weighting across classes
- **Weighted average:** Accounts for class imbalance

Interpretation:
The model performs moderately well, with better predictive power for active recalls than for terminated recalls, reflecting an opportunity to improve class-specific recall and precision.

Feature Importance Insights

Top predictors of recall termination:

- **Year:** Indicates regulatory or operational trends over time
- **Company Name:** Suggests some companies have consistently different termination outcomes, possibly due to reputation, resources, or regulatory history

Insight:
The company involved plays a critical role in determining whether a recall is terminated, highlighting the importance of corporate practices and regulatory relationships.

Discussion & Recommendations

1. **Regulatory Recommendations:**
    a. Target companies with repeated terminated recalls for additional oversight.
    b. Prioritize high-risk recall categories (Listeria, Salmonella) in inspections.
    c. Use cluster profiles to tailor communication and prevention strategies.
2. **Industry Recommendations:**
    a. Establish early warning systems leveraging forecast trends.
    b. Analyze company-level patterns to identify process or supplier risks.
    c. Train teams on top predictive features (e.g., timing, product type) to reduce recall likelihood.
3. **Technical Recommendations:**
    a. Expand features with text vectorization (TF-IDF, BERT embeddings) for recall reason descriptions.
    b. Address class imbalance (active vs. terminated recalls) using oversampling or cost-sensitive learning.
    c. Use ensemble models to improve classification accuracy.

Future Development

1. Incorporate geographic data to assess regional risk patterns.

2. Integrate social media and news data to capture public risk signals.

3. Develop a real-time dashboard for FDA and companies to monitor recall forecasts and cluster alerts.

4. Collaborate with manufacturers to implement predictive risk scoring tools within supply chain management.

Conclusion

This project explored FDA food recall data using a combined approach of clustering, forecasting, and supervised classification modeling. Through clustering, we uncovered dominant risk patterns, showing that microbial contamination (especially Listeria and Salmonella) and allergen mislabeling are the most common recall drivers. The forecasting analysis projected a rising trend in FDA recalls, emphasizing the need for proactive resource planning. The classification model, while achieving moderate accuracy, revealed meaningful predictors — notably, the year of recall and the company involved — highlighting how company practices and regulatory shifts shape recall outcomes.

Together, these findings offer both practical and technical insights. For regulators, the study underscores the importance of targeted oversight and early warning systems. For industry, it points to the value of data-driven quality control, risk monitoring, and supplier management. Technically, the project paves the way for richer models using advanced text mining, geographic data, and real-time updates.

While the current models provide useful predictions, future work should expand feature sets, improve classification performance, and integrate additional data sources to build a comprehensive early warning platform for food safety.

Explore the Full Project

The full code, models, and visualizations for this project are available on GitHub:

https://github.com/Lape2/Insightful-Analytics-Hub/blob/main/Food_Recall.ipynb

References

Phoenix. (2025, February 12). *Food recalls: Here's what to know about why they're issued, and what you should do*. USA Today. Retrieved from https://www.usatoday.com/story/news/health/2025/02/12/food-recalls-why-issued-what-to-do/

U.S. Food and Drug Administration. (2015). *FDA enforcement statistics summary, fiscal year 2015*. U.S. Department of Health and Human Services. Retrieved from https://www.fda.gov/media/95539/download

National Library of Medicine. (2023). A machine learning algorithm to predict FDA medical device recalls. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC11908527/

ResearchGate. (2023). Enhanced food safety through deep learning for food recalls prediction. Retrieved from https://www.researchgate.net/publication/345656863_Enhanced_Food_Safety_Through_Deep_Learning_for_Food_Recalls_Prediction

National Library of Medicine. (2020). Detecting reports of unsafe foods in consumer product reviews. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC6951857/