Uncovering the Latent Relationships in Food Inspection Records via Unsupervised Clustering

Anonymous Authors[1]

Abstract

Regulatory food premises inspection records encode a structured profile of each establishment's operational characteristics, equipment inventory, and compliance history. However, the administrative risk categories assigned to food premises may not capture the full latent structure present in these records. We present the first machine learning study to leverage open food premises inspection data from Toronto Public Health (TPH), introducing a novel dataset of **20,055** records that has not previously been used in any machine learning or data-driven research. We develop an unsupervised clustering pipeline in which records are projected to 2-dimensions using **UMAP** and partitioned using **Gaussian Mixture Model (GMM)**, with the number of components k selected by minimizing the Bayesian Information Criterion (BIC) over $k \in \{2, \ldots, 10\}$. We selected $k = 5$ to balance statistical fit with interpretability. Cluster quality is assessed against the regulatory three-level risk label (Low, Moderate, High) using Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). Post-hoc analysis reveals five operationally coherent establishment phenotypes, providing a complementary basis for inspection beyond rule-based risk scoring.

1. Introduction

Food safety inspections are a cornerstone of public health infrastructure, implemented by governments worldwide to reduce the incidence of foodborne illness (Barnes et al., 2024). In Ontario, Canada, food establishment inspections fall under the jurisdiction of local public health agencies (LPHAs) according to the Ontario Health Protection and Promotion Act. LPHAs conduct regulatory inspection programs, assessing the compliance of food premises with food safety requirements, including O. Reg. 493/17: Food Premises (Ministry of the Attorney General, 2017). Structured inspection records encode risk categorization, violations, and compliance actions. These records are summarized into a set of ordinal risk categories (e.g., High, Moderate, Low) that determine inspection frequency, mandated by the provincial regulation.

The risk taxonomy is operationally convenient but is determined largely by rule-based assignment procedures developed in 2019 that do not exploit the full breadth of historical inspection records. A natural question follows: *does the full inspection history contain richer latent structure than the official risk taxonomy captures?* Answering this question requires an approach that bypasses the risk categorization labels during learning.

Unsupervised clustering has proven effective at revealing latent structure in public health data, including patient subphenotyping from electronic health records (Birkenbihl et al., 2023; Lu, 2025), behavioural risk profiling of youth substance use (Yang et al., 2022), and cardiovascular disease risk stratification (Khamis et al., 2025). Systematic review by Li et al. documented the rapid growth in machine learning (ML) applications across the food safety domain from 2014 to 2024. Zhang et al. further demonstrated the utility of clustering algorithms and their applications in food safety risk assessment. These precedents motivate our application of unsupervised clustering to food inspection records. Our contributions are as follows:

1. **A standardized preprocessing pipeline** for a structured inspection dataset, comprising feature selection, missingness imputation, and mixed-type data encoding.
2. **A UMAP + GMM clustering framework** with BIC-driven component selection, producing interpretable low-dimensional embeddings and probabilistic cluster assignments.
3. **Post-hoc facility-level analysis** linking cluster membership to official risk categories, quantifying both agreement and divergence, with soft assignment confi-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

dence as an epistemic signal for ambiguous records.

2. Data and Preprocessing

2.1. Data Sources

The TPH Establishment List is a food inspection registry maintained by Toronto Public Health (TPH), comprising 20,055 records across 64 variables obtained under a data-sharing agreement with the City of Toronto. Each record corresponds to an active food premises as of February 6, 2026. Variables span establishment identity, operational characteristics, equipment inventory, staffing, licensing status, and hours of operation.

2.2. Feature Selection and Preprocessing

Three feature types were retained: categorical, binary, and numerical. Variables with greater than 20% missing values were excluded; this threshold eliminated 35 of 64 features, leaving 29 features available for downstream analysis (See Appendix A for details). The `Current Risk` (`Low`, `Moderate`, `High`) is retained solely for post-hoc evaluation and is never seen during training. **Categorical Features.** Variables with more than 15 unique levels were treated as high-cardinality: the 15 most frequent levels were retained and remaining levels were mapped to "Other". All categorical variables were subsequently one-hot encoded with no reference level dropped. **Binary Features**. Free-text "Yes/No" responses were mapped to $\{1, 0\}$. Missing values were imputed as 0 ("No"), consistent with the assumption that absence of a recorded feature indicates its non-presence. **Numerical Features.** Missing values were imputed with the column median, then subsequently standardized to zero mean and unit variance.

3. Methodology

3.1. UMAP Dimensionality Reduction

Uniform Manifold Approximation and Projection (McInnes et al., 2020) is applied to project the 68 dimensional feature matrix X to 2 dimensions using hyperparameters $n\_neighbors = 30$ and $min\_dist = 0.1$. UMAP is preferred over PCA for this application because it preserves local topological structure while allowing a non-linear manifold discovery, which is well-suited to heterogeneous mixed-type data after one-hot encoding.

3.2. Gaussian Mixture Model Clustering

A Gaussian Mixture Model (Reynolds, 2009) with full covariance is fitted on the 2-D UMAP embedding for each $k \in \{2, 3, \ldots, 10\}$. The optimal k^* is selected by balancing the Bayesian Information Criterion (BIC) score and interpretability:

$$k^* = \arg\min_k \ \mathrm{BIC}(k) = \arg\min_k \left[-2 \ln \hat{L}_k + p_k \ln N\right], \tag{1}$$

where $\hat{L}_k$ is the maximized likelihood, p_k the number of free parameters, and N the sample size. GMMs are preferred over k-means because they produce soft probabilistic assignments, accommodate ellipsoidal clusters of varying shape and scale, and support principled model selection via BIC.

3.3. Evaluation

Cluster assignments are compared with the official risk label via:

- **Adjusted Rand Index (ARI)** (Santos & Embrechts, 2009) : measures pairwise agreement, corrected for chance; $\in [-1, 1]$.
- **Normalized Mutual Information (NMI)** (Strehl & Ghosh, 2002): measures shared information, normalized to $[0, 1]$.

Both metrics treat the risk label as an *external* reference. The clustering is strictly unsupervised (risk labels are never used as input).

3.4. Post-hoc Analysis

For each cluster, we computed the prevalence of each feature, the distribution across establishment groups, and the composition of risk labels. To rank features by discriminative power across clusters, we applied **Pearson Chi-squared test** (χ^2) (Agresti, 2007) for binary and categorical features vs. cluster assignment and **One-way ANOVA F-test** for numerical features versus cluster assignment.

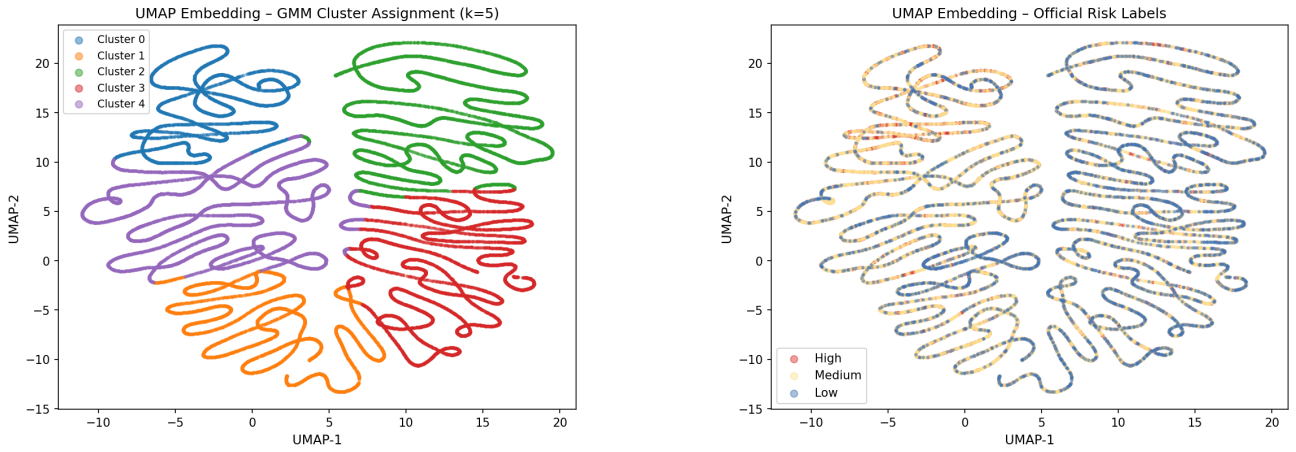
4. Experiments and Results

4.1. BIC Model Selection

BIC decreased monotonically from 275,758 at $k = 2$ to 270,956 at $k = 10$, with no inflection point (Appendix B). In the absence of a clear elbow, we selected $k = 5$ as the configuration as it produced additional cluster information than the regulatory three-level risk categorization, while generating a qualitatively distinct operational interpretation.

4.2. UMAP Visualization

The UMAP projection (Figure 1) of clusters appears as continuously connected, ribbon-like manifolds. This reveals that the underlying high-dimensional feature space has strong continuity. Food premises transition gradually along risk and infraction gradients rather than falling into



(a) GMM cluster assignment ($k = 5$). (b) Official Current Risk label ($High, Medium, Low$).

Figure 1. UMAP embeddings comparison. (a) shows the hard assignment of cluster membership generated by GMM. It applies a winner-takes-all threshold: the point is assigned to whichever cluster achieves the highest posterior probability; (b) shows the current regulatory risk label being assigned to the food premises.

discrete groups. Comparing panel (a) with (b), the official Current Risk labels do not align cleanly with any single GMM cluster. Medium and Low labels are interleaved across multiple clusters. This mismatch suggests the hand-assigned official risk categories may not fully capture the multivariate structure present in the inspection data.

4.3. Cluster Profiles and Risk Alignment

Cross-tabulation of cluster membership against official risk labels (Figure 2) reveals highly differentiated cluster profiles. The alignment metrics yield ARI $= 0.09$ and NMI $= 0.18$, indicating modest but statistically meaningful correspondence. The relatively low values reflect that clustering is driven by operational and equipment features rather than by the risk label directly. The five clusters exhibit distinct profiles:

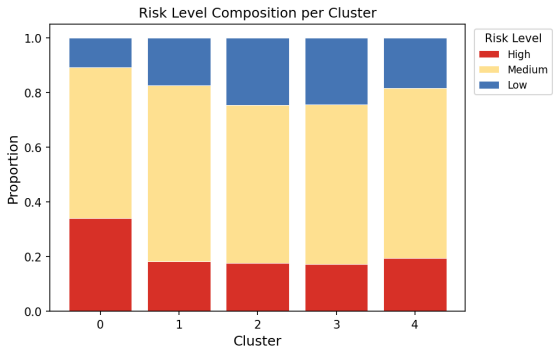


Figure 2. Row-normalized cross-tabulation of cluster membership vs. official risk level.

- **Cluster 0-Full-service restaurants**: High prevalence of commercial dishwashers, 3-compartment sinks, and liquor licences; predominantly full-service restaurants with High/Medium risk.
- **Cluster 1-Institutional kitchens**: Primarily weekday-only operations (Mon–Fri open, Sat–Sun closed); dominated by institutional catering (child care, cafeterias).
- **Cluster 2-Retail convenience stores**: High tobacco vendor prevalence and convenience store features; predominantly Low risk retail establishments.
- **Cluster 3-Take-out only outlets**: High takeout-only prevalence; limited equipment; Medium risk food take-out premises.
- **Cluster 4-Seasonal and mobile vendors**: Mixed profile; seasonal flag elevated; includes outdoor vendors and food carts.

4.4. Feature Discriminability

Figure 3 presents the top-15 features ranked by cluster-discriminating power. Previous Risk level (Chi² statistic) is the single most informative feature, followed by equipment indicators: Commercial Dishwasher, Donair/Shawarma, 3-Compartment Sink, and Barbeque/Rotisserie. Day-of-week features (Saturday, Sunday) also rank highly, reflecting systematic differences in the operational schedules of distinct establishment types.

5. Discussion

The proposed pipeline recovers five interpretable establishment phenotypes from the TPH registry without access to

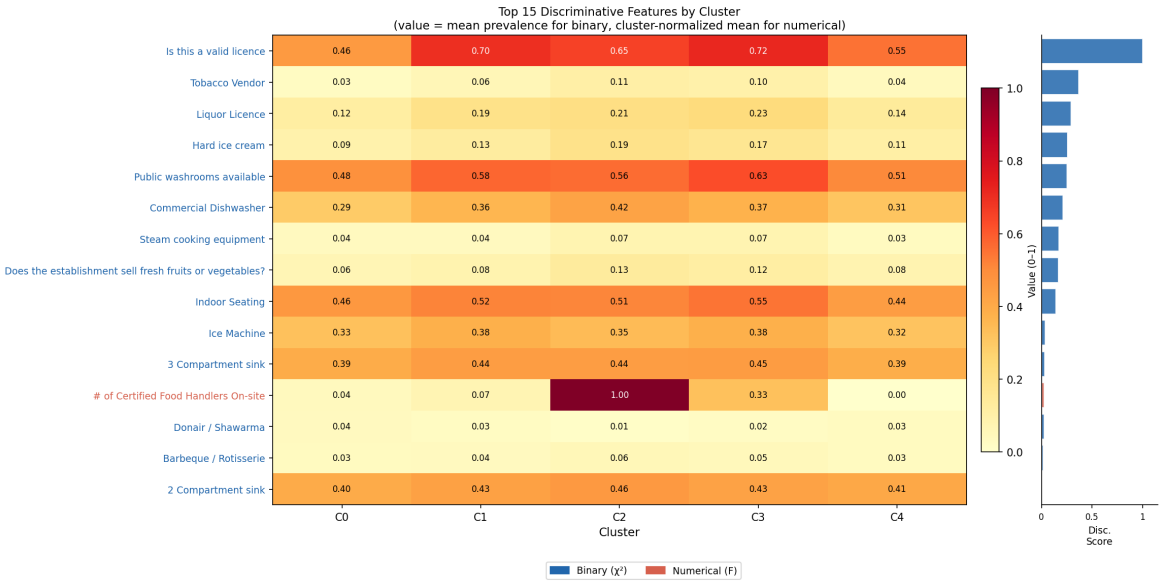


Figure 3. Top 15 discriminative features ranked jointly by normalized discriminability score (sidebar). Binary features (blue labels) are scored by Pearson χ^2 and displayed as cluster mean prevalence; numerical features (red labels) are scored by one-way ANOVA F-statistic and displayed as cluster-normalized means rescaled to $[0, 1]$. Cell values and colour intensity encode the per-cluster profile.

risk labels during training. The moderate alignment with official risk tiers (ARI = 0.09, NMI = 0.18) indicates that operational and equipment profiles encode structural information beyond what the current three-class taxonomy captures. We highlight several actionable implications:

Medium risk is internally heterogeneous. . Multiple clusters with predominantly Medium-risk membership exhibit substantially different equipment profiles. Stratifying the Medium tier by equipment complexity could enable more granular inspection scheduling.

Having a valid license is the strongest individual predictor of cluster membership. This finding validates the use of licence status in risk labelling but also suggests that establishments undergoing licence transitions may be systematically misallocated under the current scheme.

Day-of-week patterns differentiate clusters. Saturday and Sunday operations are statistically significant cluster discriminators, reflecting that weekend-only or daily operators present different food safety risk profiles even within the same risk tier.

5.1. Limitations

The monotonically decreasing BIC without a clear elbow makes k selection sensitive to range assumptions. $k = 5$ was selected for better interoperability. Missingness imputation as 0 for binary features may introduce noise for variables with systematic rather than random absence patterns. Finally, the analysis is cross-sectional. Timestamps and geographical information were excluded in the study. Nevertheless, the clusters has demonstrated meaningful separation. Longitudinal modelling of risk trajectory would be a natural extension to this line of work.

6. Conclusion

To the best of our knowledge, this is the first application of unsupervised clustering to food premises inspection records, and the first use of TPH's open inspection dataset in any data-driven research context. Prior work on such data has relied primarily on statistical models (Howell et al., 2026; Cho, 2026), and much of the broader food safety ML literature has focused on upstream tasks such as microbial detection and supply chain optimization (Zhong et al., 2023; Revelou et al., 2025; Liu et al., 2024). Inspector-generated records are distinctive in that virtually all covered premises serve consumers directly, making them particularly relevant to public health practice.

This work demonstrates that unsupervised clustering can automatically discover latent structure within structured inspection records, recovering operationally coherent establishment phenotypes that partially diverge from the official risk taxonomy. The identified clusters and their risk compositions may inform new approaches to inspection scheduling and resource allocation. Moreover, the discriminative feature analysis provides a foundation for subsequent predictive risk modelling, with the potential to support more responsive, evidence-based food safety policy.

7. Impact Statement

This paper presents work aimed at advancing the field of food safety inspection. Applications of this methodology could improve the efficiency of regulatory inspection programs and support more equitable allocation of public health resources. No specific ethical concerns are anticipated beyond those applicable to any use of administrative records under data-sharing agreements.

References

Agresti, A. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Hoboken, NJ, 2nd edition, 2007. ISBN 978-0-471-22618-5.

Barnes, J. B., Smith, J. C., Ross, K. E., and Whiley, H. Performing food safety inspections. *Food Control*, 160: 110329, June 2024. ISSN 0956-7135. doi: 10.1016/j.foodcont.2024.110329.

Birkenbihl, C., Ahmad, A., Massat, N. J., Raschka, T., Avbersek, A., Downey, P., Armstrong, M., and Fröhlich, H. Artificial intelligence-based clustering and characterization of Parkinson's disease trajectories. *Scientific Reports*, 13(1):2897, February 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-30038-8.

Cho, S. Y. Enhancing food collection inspection efficiency using a Bayesian network model. *Food Control*, 181: 111786, March 2026. ISSN 0956-7135. doi: 10.1016/j.foodcont.2025.111786.

Foundation, P. S. Python (Version 3.11), 2016.

Howell, A., Evans, A., Jensen, S., Arnold, N. L., and Kowalcyk, B. A generalized linear mixed model approach to assess associations between Certified Food Protection Managers and inspection performance in Franklin County, Ohio. *Food Control*, 182:111802, April 2026. ISSN 0956-7135. doi: 10.1016/j.foodcont.2025.111802.

Khamis, G. S. M., Alqahtani, N. S., Alanazi, S. M., Alruwaili, M. M., Alenazi, M. S., and Alrawaili, M. A. Using Fuzzy C-Means clustering and PCA in public health: A machine learning approach to combat CVD and obesity. *Informatics in Medicine Unlocked*, 57:101666, January 2025. ISSN 2352-9148. doi: 10.1016/j.imu.2025.101666.

Konishi, S. and Kitagawa, G. Bayesian Information Criteria. In *Information Criteria and Statistical Modeling*, pp. 211–237. Springer New York, New York, NY, 2008. ISBN 978-0-387-71887-3. doi: 10.1007/978-0-387-71887-3_9.

Li, R., Yin, C., Yang, S., Qian, B., and Zhang, P. Marrying Medical Domain Knowledge With Deep Learning on Electronic Health Records: A Deep Visual Analytics Approach. *Journal of Medical Internet Research*, 22 (9):e20645, September 2020. ISSN 1439-4456. doi: 10.2196/20645.

Liu, J., Bensimon, J., and Lu, X. Chapter Two - Frontiers of machine learning in smart food safety. In Lu, X. (ed.), *Smart Food Safety*, volume 111 of *Advances in Food and Nutrition Research*, pp. 35–70. Academic Press, 2024. doi: 10.1016/bs.afnr.2024.06.009.

Lu, Z. Clustering Longitudinal Data: A Review of Methods and Software Packages. *International Statistical Review*, 93(3):425–458, 2025. ISSN 1751-5823. doi: 10.1111/insr.12588.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020.

Ministry of Health and Long-Term Care. Food premises reference document, 2019. 2019.

Ministry of the Attorney General. Food Premises, Ontario Regulation 493/17, 2017.

Revelou, P.-K., Tsakali, E., Batrinou, A., and Strati, I. F. Applications of Machine Learning in Food Safety and HACCP Monitoring of Animal-Source Foods. *Foods*, 14 (6):922, March 2025. ISSN 2304-8158. doi: 10.3390/foods14060922.

Reynolds, D. Gaussian Mixture Models. In *Encyclopedia of Biometrics*, pp. 659–663. Springer, Boston, MA, 2009. ISBN 978-0-387-73003-5. doi: 10.1007/978-0-387-73003-5_196.

Santos, J. M. and Embrechts, M. On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. In Alippi, C., Polycarpou, M., Panayiotou, C., and Ellinas, G. (eds.), *Artificial Neural Networks – ICANN 2009*, pp. 175–184, Berlin, Heidelberg, 2009. Springer. ISBN 978-3-642-04277-5. doi: 10.1007/978-3-642-04277-5_18.

Strehl, A. and Ghosh, J. Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research 3*, 2002.

Yang, Y., Butt, Z., Leatherdale, S. T., Morita, P. P., Wong, A., Rosella, L., and Chen, H. H. Phenotyping Risk Profiles of Substance Use Among Canadian Youth via Cluster Analysis on COMPASS Data, December 2022.

Zanabria, R., Racicot, M., Cormier, M., Arsenault, J., Ferrouillet, C., Letellier, A., Tiwari, A., Mackay, A., Griffiths, M., Holley, R., Gill, T., Charlebois, S., and Quessy,

S. Selection of risk factors to be included in the Canadian Food Inspection Agency risk assessment inspection model for food establishments. *Food Microbiology*, 75:72–81, October 2018. ISSN 0740-0020. doi: 10.1016/j.fm.2017.09.019.

Zhang, Q., Lu, Z., Liu, Z., Li, J., Chang, M., and Zuo, M. Application of Machine Learning in Food Safety Risk Assessment. *Foods*, 14(23):4005, November 2025. ISSN 2304-8158. doi: 10.3390/foods14234005.

Zhong, J., Sun, L., Zuo, E., Chen, C., Chen, C., Jiang, H., Li, H., and Lv, X. An ensemble of AHP-EW and AE-RNN for food safety risk early warning. *PLOS ONE*, 18(4): e0284144, 2023. ISSN 1932-6203. doi: 10.1371/journal.pone.0284144.

A. Dataset Description and Preprocessing

Dataset Description. The dataset used in this study was drawn from the City of Toronto's municipal-level food premises inspection registry, with the original file name "Establishment List.xlsx". Each record within this file corresponds to a unique food premises and includes a curated set of operational and infrastructural attributes collected during routine inspections. For the official assigned risk classification, TPH originally assigned four ordinal levels (High, Medium, Low, and Super Low) instead of three (which is the provincial standard). Thus, for the purpose of standardization, we merged the class "Super Low" with "Low" (having inspection frequency that's lower or equal to once per year). The raw dataset exhibits a moderate class imbalance, with "Medium" being the dominant group. Figure 4 shows the distribution of the risk categorization.

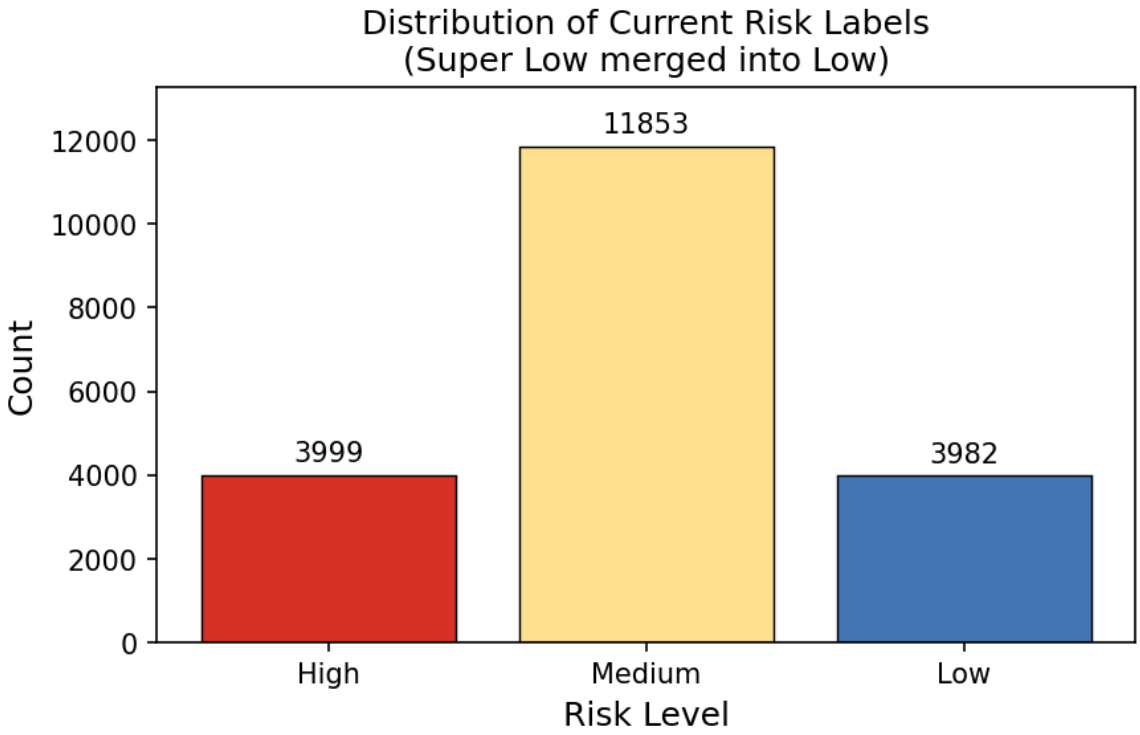


Figure 4. Risk classification distribution of the original file. The majority of food premises are classified as "Medium". The required inspection frequency for "Medium" is twice per year. The required inspection frequency for "High" is three times per year, and for "Low" is once per year.

Missingness Report. Features with higher than 20% missingness were excluded in the downstream analysis. Table 1 summarizes the features being dropped. The potential reasons for having a high value of missingness are 1) the field was not required to be filled during inspection; 2) administrative variables that might not be applicable to the majority of food premises;3) accidental omission during manual data entry and/or collection (human error).

Table 1. Variables excluded due to >20% missingness (top-20 shown).

Variable	**Missing (%)**
Date Permanently Closed	99.8%
Vehicle Plate Number	99.7%
Date Out of Business (OOB)	99.7%
Toronto Vehicle Licence	99.7%
Seasonal Month Open	95.9%
Seasonal Month Closed	95.9%

Continued on next page

Table 1. Variables excluded due to >20% missingness (continued)

Variable	Missing (%)
Food Preparation Location	94.8%
Approved Source of Food	86.0%
Start of Operation Date	77.2%
UNIT	74.4%
Regulated Food Premises	50.7%
Municipal Code Chapter 545 N/A	44.0%
Business Licence N/A	43.1%
Public washrooms location?	40.5%
Toronto Licence Expiry Date	36.8%
Sunday Closing	36.2%
Sunday Opening	36.0%
Number of staff washrooms	32.4%
Saturday Closing	31.8%
Saturday Opening	31.7%

Preprocessing Statistics. After filtering for missingness, the retained variables are presented in Table 2. For each weekday (Sunday through Saturday), a binary *Open* indicator was engineered: if both the opening and closing time fields are non-null and non-zero for a given day, the establishment is coded as open on that day ($= 1$); otherwise closed ($= 0$). This collapses 14 sparse time columns into 7 interpretable binary features. All features were standardized where appropriate prior to dimensionality reduction. UMAP was applied to the full feature matrix to produce a 2-dimensional embedding for visualization and downstream clustering.

Table 2. Descriptive statistics for all retained features (raw, before scaling).

Feature	Type	Missing (%)	Mean	Std
Is this a valid licence	binary	7.950	0.625	0.484
Tobacco Vendor	binary	17.700	0.072	0.258
Seasonal	binary	0.120	0.039	0.193
Commercial Dishwasher	binary	14.330	0.357	0.479
2 Compartment sink	binary	16.040	0.426	0.494
3 Compartment sink	binary	15.760	0.424	0.494
Soft Serve Ice Cream	binary	12.750	0.035	0.182
Hard ice cream	binary	12.890	0.141	0.348
Barbeque / Rotisserie	binary	12.910	0.042	0.201
Ice Machine	binary	11.830	0.350	0.477
Cooking equip. for customers	binary	13.090	0.008	0.091
Steam cooking equipment	binary	13.180	0.054	0.225
Sous Vide Cooking	binary	13.130	0.014	0.118
Donair / Shawarma	binary	13.140	0.026	0.159
Indoor Seating	binary	12.340	0.499	0.500
Takeout only	binary	13.550	0.356	0.479
Public washrooms	binary	14.120	0.556	0.497
Sunday_Open	binary	0.000	0.000	0.000
Monday_Open	binary	0.000	0.000	0.000
Tuesday_Open	binary	0.000	0.000	0.000
Wednesday_Open	binary	0.000	0.000	0.000
Thursday_Open	binary	0.000	0.000	0.000
Friday_Open	binary	0.000	0.000	0.000

Continued on next page

Table 2. Descriptive statistics for all retained features (continued)

Feature	Type	Missing (%)	Mean	Std
Saturday_Open	binary	0.000	0.000	0.000
# of Cert. Food Handlers Req.	numerical	13.320	1.194	0.942
# of Cert. Food Handlers On-site	numerical	13.230	1.320	1.819
ESTABLISHMENT	categorical	0.010	—	—
POSTAL CODE	categorical	1.670	—	—
Establishment Group	categorical	0.120	—	—
Previous Risk	categorical	16.340	—	—

B. Gaussian Mixture Model Clustering

Gaussian Mixture Model (GMM) clustering was performed on the UMAP embeddings. The number of clusters/components was selected through the Bayesian information criterion (BIC) (Konishi & Kitagawa, 2008). We selected $k = 5$ as the operationally interpretable solution. It should be noted that $k = 5$ is not the minimal k within the sweeping range.

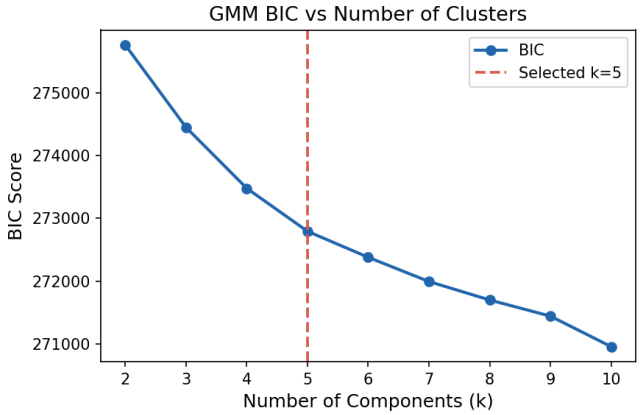


Figure 5. BIC score as a function of GMM components k. BIC is minimized at $k = 10$, the selected k is $k = 5$.

Table 3. GMM BIC and AIC scores for $k = 2$ to 10. Selected $k^* = 5$.

k	**BIC**
2	275,757.7
3	274,446.3
4	273,476.9
5	**272,792.3**
6	272,378.6
7	271,992.6
8	271,698.8
9	271,442.0
10	270,956.4