Analyzing Food Safety Compliance in Toronto: Identifying Current Hazards and Challenges*

Several Factors Affecting Local Food Safety in Toronto

Pengyu Sui

April 18, 2024

The dataset used in this study comes from the DineSafe program managed by Toronto Public Health, available through Open Data Toronto. We clean it up, generate the model and visualize it.

Extract of the questions from Gebru et al. (2021)

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - This dataset was created to analyze food safety in Toronto. Our task was to create a logistic regression model by cleaning the dataset Dinesefe, which was downloaded from the Open Data Toronto website.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - Created by the DineSafe program managed by Toronto Public Health to provide data on food safety in Toronto.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - No one funded the creaton of the dataset
4. *Any other comments?*
   - NA

*Code and data are available at: https://github.com/simon0202sui/Analyzing-Food-Safety-Compliance-in-Toronto.git.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
    - The entity of the data is a csv file containing information about food safety, such as the number of violations, the minimum number of sampling inspections per year, the name of the violating restaurant and other information.
2. *How many instances are there in total (of each type, if appropriate)?*
    - There are 217 instances.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
    - It is not a sample of a larger set.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
    - such as the number of violations, the minimum number of sampling inspections per year, the name of the violating restaurant and other information.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - The primary target or label in this dataset is the Establishment Status, include Pass and Conditional Pass.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
    - Information on specific personal details or proprietary business information that might compromise privacy or competitive positions may be omitted.
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
    - Relationships between inspections are implicit through the establishment data, showing repeated inspections or follow-ups which may indicate compliance history and trends.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - For model training and testing purposes, the dataset can be split into training and test sets, typically using a 70-30% split to validate the predictive accuracy of the logistic regression models developed.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - NA

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The dataset is self-contained and primarily relies on internal data from Toronto Public Health. External dependencies, if any, include links to additional government or health data portals, which are maintained under governmental data policies ensuring continuity and availability.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - he dataset does not contain confidential data that is protected by legal privilege or confidentiality agreements, such as personal health information or private communications. Instead, it focuses on public health inspection results, which are typically available to the public as part of transparency initiatives by local government bodies.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No, the dataset does not contain information that is offensive, insulting, threatening, or likely to cause anxiety. The data are strictly related to food safety inspections and compliance with public health regulations, which are intended for public dissemination to ensure community health and safety.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

- The dataset does not identify sub-populations by personal demographics such as age or gender. It focuses on establishments and their compliance with health regulations. However, it categorizes establishments by type (e.g., restaurant, cafe, supermarket), which allows for analysis of compliance across different types of food service providers. This categorization can help identify trends or issues specific to certain types of establishments but does not differentiate data based on personal characteristics of individuals.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No this is not possible

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No it doesn't contain any of this.

16. *Any other comments?*

    - NA

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was directly observed and recorded during routine inspections conducted by Toronto Public Health inspectors. Each instance in the dataset represents findings from these inspections, detailing compliance with food safety regulations

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - Data were collected using standardized public health inspection protocols by certified health inspectors. These procedures are regulated and periodically reviewed to ensure they meet the necessary standards of accuracy and thoroughness.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The dataset is not a sample; it encompasses all recorded inspections under the DineSafe program for the specific time frame it covers.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - Toronto Public Health inspectors, who are municipal employees, conducted the inspections. Their compensation is part of their regular duties as public health officials.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The data was collected over the course of routine inspections and is continuously updated. Each data instance corresponds to the actual date of the inspection as recorded in the dataset.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - Given that the dataset involves regulatory compliance data that is collected as part of standard public health operations, specific ethical reviews for the data collection process per se are generally not applicable. However, the processes adhere to municipal and provincial regulations regarding public health data collection.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - The data were collected directly during inspections of the establishments, without intermediation by third parties.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - As the data collection is a routine part of health inspections, establishments are aware that inspections are standard practice and that findings will be recorded. However, specific notifications about data recording for this dataset per se are not typically provided as the data use is understood implicitly.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Consent in the traditional sense is not applicable as the data collection is mandated by law as part of public health and safety regulations. Establishments must comply with these inspections as a condition of their operational licensing.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - NA

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - NA

12. *Any other comments?*

    - NA

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Yes, preprocessing and cleaning involve standardizing entries, correcting any evident errors, and ensuring that all entries are complete where possible. Labeling involves categorizing inspection results and noted violations according to predefined criteria.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - Typically, both raw and processed data are retained to ensure that original records are preserved for verification and validation purposes.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - The preprocessing is generally conducted using statistical software packages like R, with specific scripts developed for regular data cleaning tasks.

4. *Any other comments?*
    - NA

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Yes, the dataset is regularly used for generating compliance reports, identifying trends in food safety, and guiding policy decisions regarding public health practices in Toronto.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - Public health reports and policy documents that utilize these data may be available through Toronto Public Health publications and the Open Data Toronto portal.
3. *What (other) tasks could the dataset be used for?*
    - Beyond regulatory compliance, the dataset could be used for academic research in public health, policy studies, or by private entities seeking to understand food safety trends for business analytics.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - No
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - No
6. *Any other comments?*
    - NA

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

- Yes, the dataset will be made available to third parties. It is part of an open data initiative by the City of Toronto, which aims to promote transparency, enable community engagement, and foster innovation by providing access to city-collected data. As such, it is intended for use by researchers, developers, policymakers, and the general public to enhance understanding and decision-making concerning public health and food safety.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The dataset is distributed via the Open Data Toronto portal, where it can be accessed directly and downloaded in various formats, such as CSV or through an API. The specific distribution method allows for easy access and integration into various applications and research projects. Currently, the dataset does not have a digital object identifier (DOI), as it is maintained as a part of a continuously updated government database rather than a static research dataset.

3. *When will the dataset be distributed?*

    - The dataset is already available and is continuously updated. New data from ongoing inspections are added regularly, ensuring that the dataset remains current and relevant. The frequency of updates may vary, but typically, updates are processed and released in alignment with the city's schedule for public data updates, ensuring that the latest information is always accessible.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - No

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - No

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - No

7. *Any other comments?*

    - NA

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - NA
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - NA
3. *Is there an erratum? If so, please provide a link or other access point.*
    - NA
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - NA
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
    - The dataset primarily focuses on establishments rather than individuals, so typical personal data retention policies do not apply.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
    - Older versions of the dataset are generally not maintained separately once updated, as the dataset is intended to provide the most current and accurate reflection of the health inspection status of establishments.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
    - While external contributions to the dataset itself are not typically accepted because the data is sourced directly from official inspections and must adhere to strict data integrity and verification standards, users are encouraged to use the dataset for their own analyses, applications, or research. If users have suggestions or feedback, they can typically submit these through the Open Data Toronto portal's feedback mechanism.

8. *Any other comments?*
    - NA

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.