Chicago's Culinary Compliance
A review of restaurant and food service inspections

Authors:

Vikranth Reddy Chapaala

Venkat Sai Phanindra Anagam

Ritesh Venkata Sai Vesalapu

Oklahoma State University

Master of Science Business Analytics and Data Science
Spears School of Business

December 2023

Table of Contents

Abstract:

This project delves into the Chicago Food Inspection dataset to explore the relationship between the attributes of food establishments and the outcomes of their inspections. By examining the kind of facility, its geographical location, and the various types of inspections conducted, the analysis aims to shed light on how these elements might influence inspection outcomes. Additionally, the project seeks to discover if there are any noticeable trends in inspection outcomes over time. The hypothesis is that the type of facility, its location, and the nature of the inspection significantly influence the inspection outcomes. This project aims to analyze the data to validate or refute this hypothesis.

Introduction:

In the realm of public health and safety, the inspection of food establishments plays a crucial role in ensuring the wellbeing of consumers. The city of Chicago, under the aegis of the Chicago Department of Public Health’s Food Protection Program, has been rigorously conducting inspections of restaurants and other food establishments since January 1, 2010. This continuous and systematic effort is not just a routine procedure but a cornerstone in safeguarding public health standards in one of the United States' largest urban centers.

The inspections, carried out by trained and certified staff, adhere to a standardized procedure to guarantee consistency and reliability in the assessment of these establishments. The results of these inspections are meticulously recorded and reviewed by a State of Illinois Licensed Environmental Health Practitioner (LEHP), ensuring the adherence to high standards of accuracy and professionalism.

The comprehensive data collected from these inspections is more than a record of compliance; it serves as an invaluable resource for understanding the landscape of food safety in Chicago. This dataset, which is publicly available in a simplified form through a dedicated data portal, includes a variety of crucial elements that paint a detailed picture of each establishment. These elements range from the legal and common names of the establishments (DBA and AKA), their unique license numbers, types of facilities, risk categories, and full addresses to more specific data such as the dates of inspections, types of inspections, inspection results, and detailed accounts of any violations found.

This report aims to delve into this rich dataset to uncover trends, patterns, and insights that can inform policy makers, health officials, and the general public. By analyzing the data from January 1, 2010, to the present, we seek to provide a comprehensive overview of the state of food safety in Chicago, identifying areas of success and highlighting avenues for improvement. The ultimate goal is to contribute to the ongoing effort to enhance food safety standards, reduce health risks, and ensure that the dining experience in Chicago remains both enjoyable and safe for all.

Literature Review:

The importance of food safety in public health has been extensively documented, highlighting the critical role of regular inspections in maintaining standards. This literature review focuses on existing research and findings relevant to the inspection of food establishments, the factors influencing inspection outcomes, and the trends observed in these practices, particularly within urban environments like Chicago.

In the first key work, Barnes, J., Whiley, H., Ross, K., & Smith, J. (2022) provide a comprehensive examination of food safety inspections. Their study focuses on the effectiveness and methodologies of these inspections, particularly in the prevention of foodborne illnesses. This research is pivotal in understanding the varying approaches used in food safety inspections and their effectiveness in different types of food establishments. Their findings offer valuable insights into the correlations between inspection methods and outcomes, providing a solid foundation for further analysis in similar contexts, such as the Chicago Food Inspection dataset.

Another significant contribution to this field is the work by Singh, S., Shah, B., Kanich, C., & Kash, I. A. (2022), which delves into the use of algorithms and predictive analytics in food safety inspections. Their research explores the implementation of fair decision-making processes in food inspection procedures, emphasizing the role of data and algorithms. This study is particularly relevant for analyzing temporal trends and changes in food inspection outcomes. It sheds light on how modern technological approaches can be integrated into traditional inspection processes, potentially leading to more efficient and effective food safety practices.

Lastly, the impact of big data in enhancing food safety measures is thoroughly reviewed by Marvin, H. J. P. (2017). This review opens up a discussion on the broader application of data analytics in food safety, especially concerning the geographical impact on inspection outcomes. It provides a comprehensive overview of how big data can be leveraged to improve food safety standards and practices. The insights from this review are crucial for understanding how the analysis of large datasets, such as the Chicago Food Inspection data, can be used to identify trends and patterns that are not immediately apparent, thereby contributing to more targeted and effective food safety measures.

This literature review underscores the complexity and multidimensional nature of food safety inspections. The varied factors impacting inspection outcomes, from the type of facility and location to the specific inspection practices, are crucial for understanding and improving food safety standards. The current project aims to contribute to this body of knowledge by analyzing the Chicago Food Inspection dataset, testing the hypothesis regarding the influence of these factors on inspection outcomes, and identifying any significant trends or patterns.

Data Description:

The provided data originates from inspections conducted on restaurants and other food establishments in Chicago, spanning from January 1, 2010, up to the middle of 2021. These inspections are carried out by the staff of the Chicago Department of Public Health's Food Protection Program, following a standardized procedure. The inspection results are recorded in a database, subsequently reviewed, and approved by a State of Illinois Licensed Environmental Health Practitioner (LEHP). ("Food Inspections - Map | City of Chicago | Data Portal") The data has the fields:

DBA (Doing Business As): This denotes the legal name of the establishment.

AKA (Also Known As): This is the commonly recognized name by which the establishment is known to the public.

License Number: Each establishment is assigned a unique license number by the Department of Business Affairs and Consumer Protection.

Type of Facility: Establishments are categorized into various types, including bakery, banquet hall, candy store, caterer, coffee shop, day care center (for different age groups), gas station, Golden Diner, grocery store, hospital, long-term care center (nursing home), liquor store, mobile food dispenser, restaurant, school, shelter, tavern, social club, wholesaler, or Wrigley Field Rooftop.

Risk Category of Facility: Establishments are classified based on their risk level concerning public health, ranging from 1 (highest risk) to 3 (lowest risk). Inspection frequency is determined by these risk categories, with risk 1 establishments subject to the most frequent inspections and risk 3 establishments subject to less frequent inspections.

Address Details: This encompasses the complete address, including street address, city, state, and zip code, where the establishment is located.

Inspection Date: The date on which a specific inspection occurred. It is common for an establishment to undergo multiple inspections, each marked by a different inspection date.

Inspection Type: The type of inspection conducted, which can fall into various categories, such as canvass (routine inspection), consultation (pre-opening inspection), complaint-based inspection, license inspection, suspect food poisoning inspection, or task-force inspection. Re-inspections may occur for most of these inspection types and are duly noted.

Inspection Results: Inspection outcomes are classified into one of three categories: "pass," "pass with conditions," or "fail." A "pass" indicates that no critical or serious violations were identified during the inspection. A "pass with conditions" implies the presence of critical or serious violations, which were addressed and corrected during the inspection. A "fail" indicates the presence of critical or serious violations that could not be rectified during the inspection. It is essential to note that a "fail" does not automatically lead to the suspension of the establishment's license. The data also includes information on establishments that are out of business or not found.

Violations: Establishments may receive one or more of 45 distinct. For each violation number listed, the specific requirement that the establishment must meet to avoid the violation is indicated, followed by a description of the findings that led to the issuance of the violation.

Methodologies

Libraries Used:

NumPy and Pandas:

- Utilized for core data manipulation and numerical computing in Python.
- Pandas enabled seamless management of missing data, data operations, grouping, and aggregation.

Matplotlib and Plotly:

- Matplotlib was used for preliminary data exploration and visualizing analysis results.
- Plotly enhanced the interactivity of visualizations, important for time-series forecasts and ROC curves.

Fuzzywuzzy:

- Implemented for cleaning and consolidating text data by matching and correcting similar strings.

Folium:

- Used for geospatial data visualization and interactive mapping to represent data distribution.

Regular Expressions (re):

- Deployed for searching and parsing complex string patterns within text data.

Scikit-learn:

- The primary library for preprocessing, cross-validation, and machine learning.
- Enabled encoding of categorical variables and supported a variety of predictive models.
- Provided tools for robust model performance evaluation.

Facebook's "Prophet":

- Employed for its user-friendly and flexible approach to time-series forecasting.
- Helped capture trends and seasonality in inspection outcomes, aiding planning and decision-making.

Data Cleaning:

Comprehensive Data Structuring and Preliminary Cleaning: Our data cleaning journey began with the fundamental structuring of the dataset using Pandas. This initial step involved meticulous inspection and cleansing of raw data, from addressing missing values to rectifying format inconsistencies. By executing strategic subsetting and grouping operations, we laid a solid foundation, allowing us to quickly identify and address any discrepancies within our data.

Textual Data Consistency and Normalization: Given the critical nature of textual accuracy, we deployed Fuzzywuzzy's string matching algorithms to standardize and unify our text data. This process was pivotal in harmonizing diverse textual entries, from correcting common misspellings to consolidating varied nomenclatures for similar concepts, thereby ensuring that our dataset reflected a consistent and standardized language, especially in context of maintaining entries in the "facilitytype" and "city" columns.

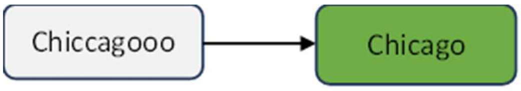


Figure 1. Reference Example

Regex-Powered Data Extraction: With the aim of transforming verbose text into discrete, analyzable units, we harnessed the power of Regular Expressions, collections and itertools to extract specific violation codes. These codes, buried within extensive textual descriptions, were meticulously isolated and categorized, converting complex narratives into structured data points ready for quantitative analysis.

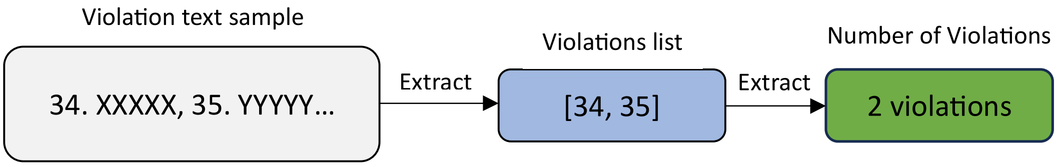


Figure 2. Regex Extraction

Geospatial Data Integrity with Web Scraping: Recognizing the imperative of geospatial accuracy, we leveraged BeautifulSoup4 to enhance the integrity of our location data. By scraping verified geographic information from the web, we were able to authenticate and rectify our city and suburb entries, infusing our dataset with an additional layer of precision.

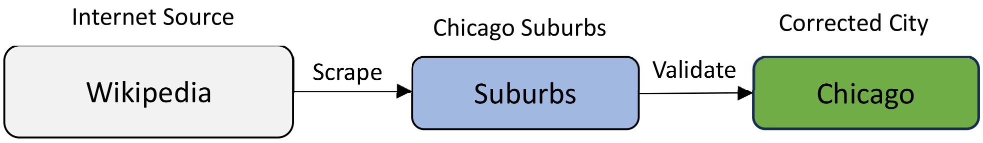


Figure 3. Web Scraping and city validation

Interactive Geospatial Visualization: With clean and verified location data, we employed Folium to create interactive maps that provided a spatial perspective of inspection data points. This visualization not only served as a tool for pattern recognition but also acted as a quality check, revealing any outliers or anomalies in the geographic distribution of the data.

Categorical Data Encoding: Advancing towards the integration of machine learning algorithms, we utilized Scikit-learn's preprocessing capabilities to encode categorical data. This encoding process was vital for translating textual labels into a numerical format, a prerequisite for the algorithms we were poised to implement.

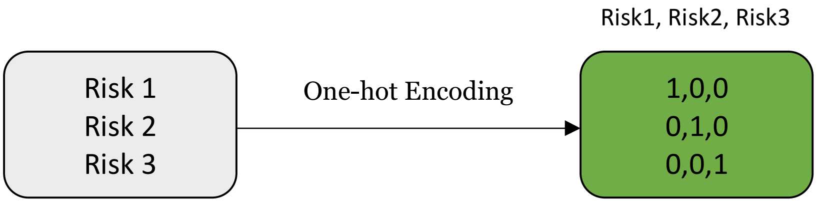
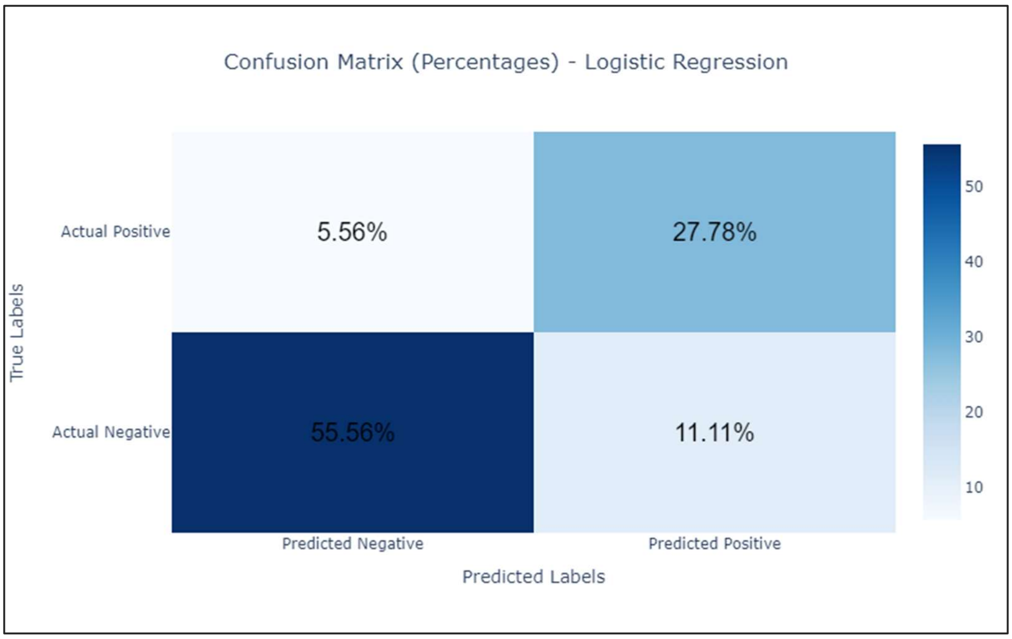


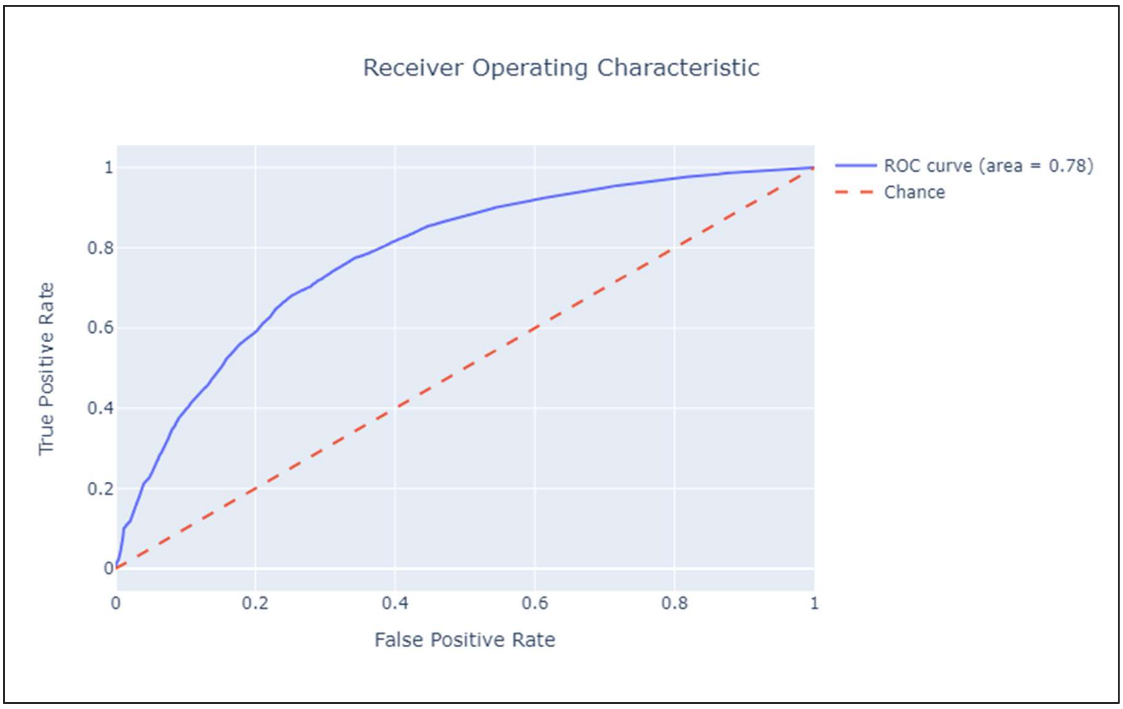
Figure 4. One-hot encoding

Predictive Modelling:

Logistic Regression:

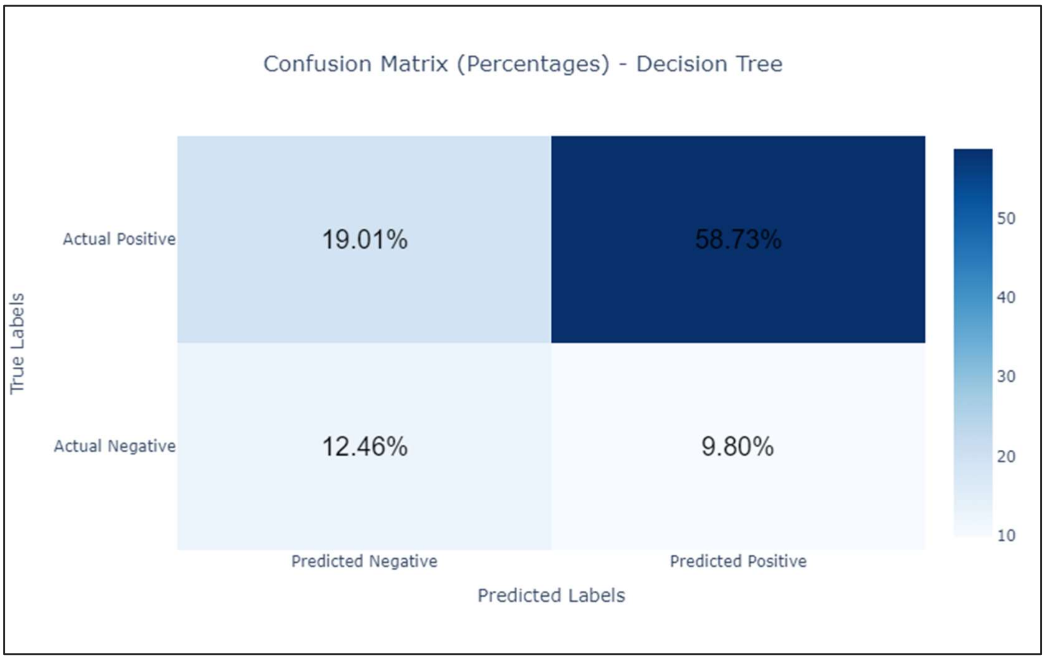
Logistic Regression is a statistical method used for binary classification, predicting one of two outcomes based on input variables. It models the probability of an event occurring.

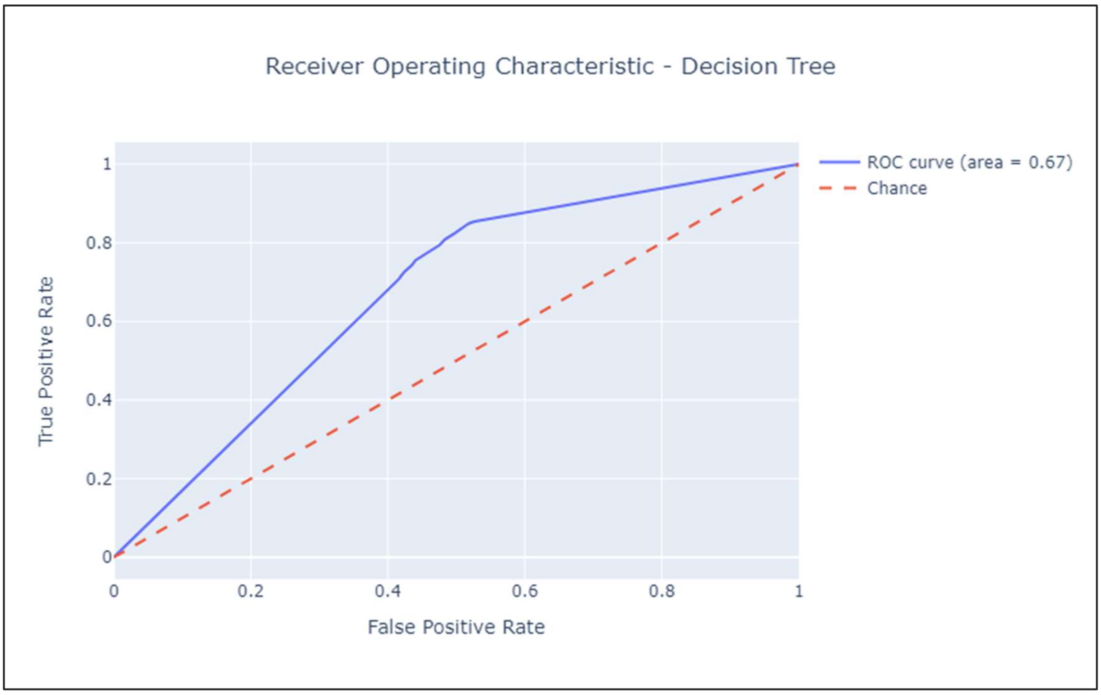




Decision Tree:

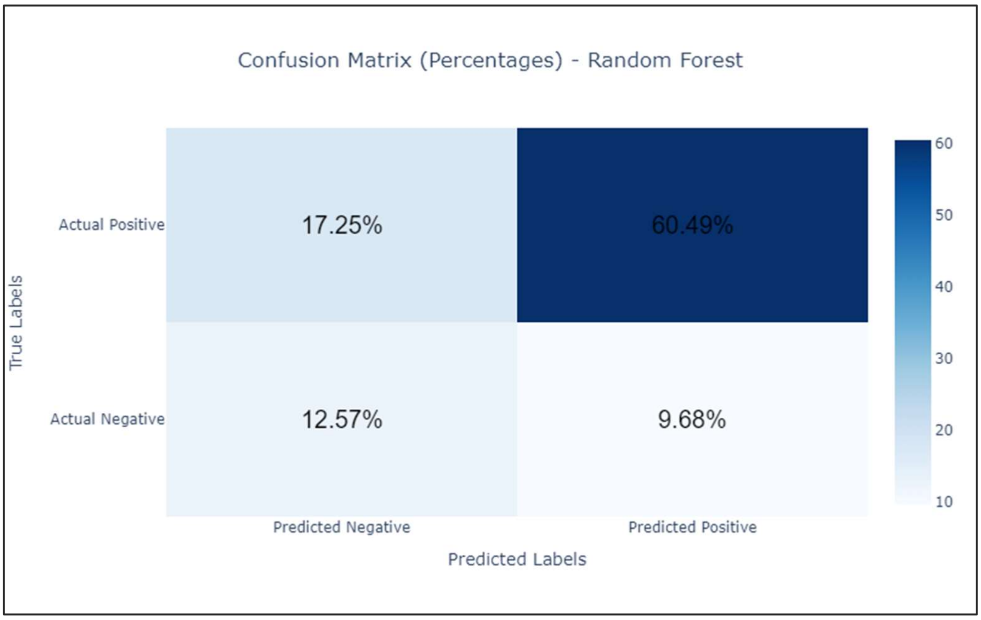
A Decision Tree is a tree-like model used in machine learning for classification and regression. It makes decisions based on input features, breaking down data into smaller subsets.

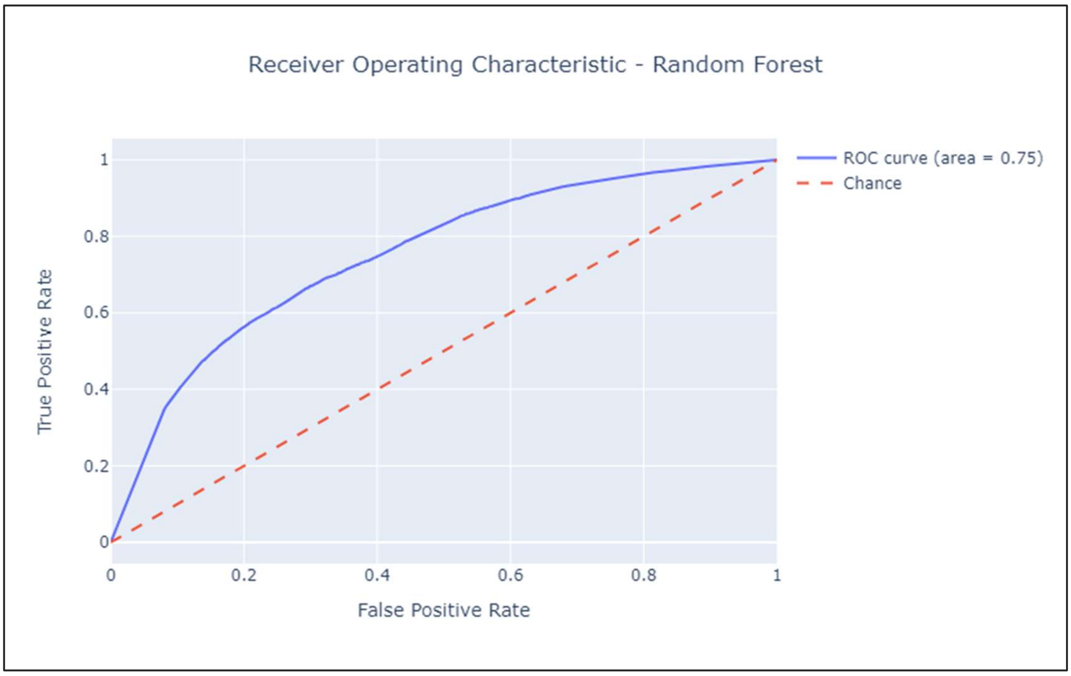




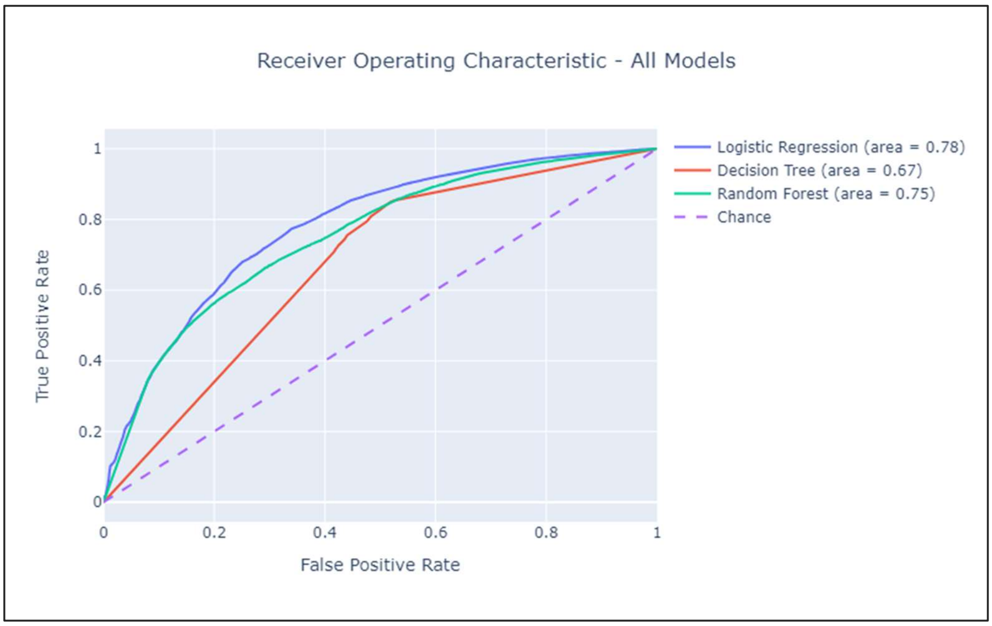
Random Forest:

Random Forest is an ensemble learning technique that combines multiple Decision Trees to improve prediction accuracy. It's known for its robustness and ability to handle complex data.





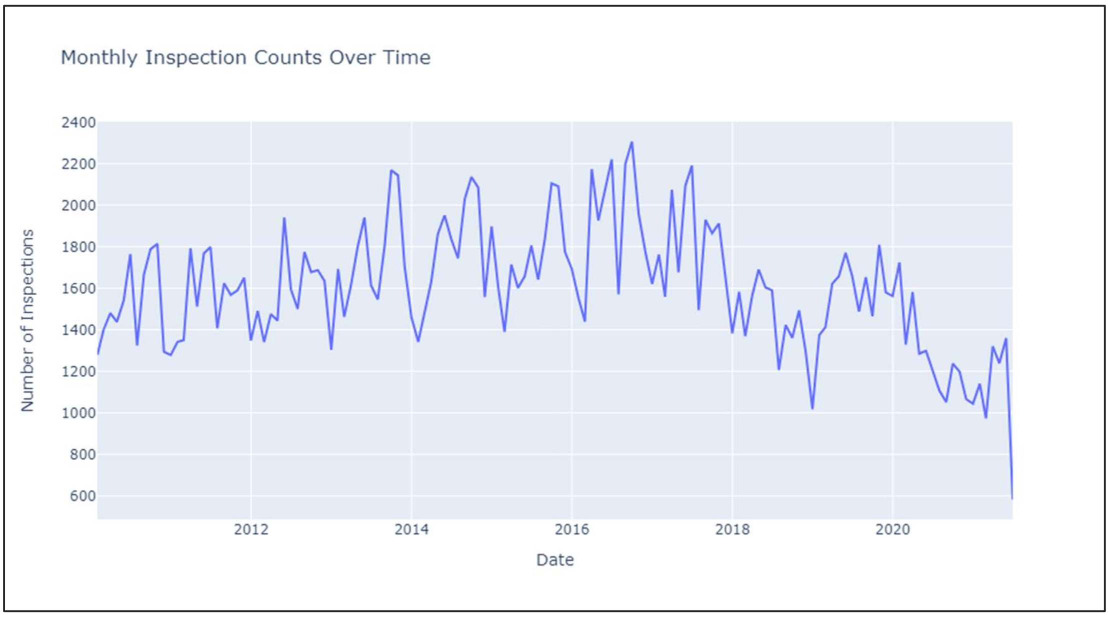
Preferred Model:



In our comparative analysis of classification models, the ROC curves reveal that Logistic Regression achieves the highest classification accuracy with an AUC of 0.78, followed by Random Forest with an AUC of 0.75, and Decision Tree with an AUC of 0.67. Each model's performance surpasses the baseline of random chance, indicating effective predictive capabilities, with Logistic Regression being the most proficient in distinguishing between classes.

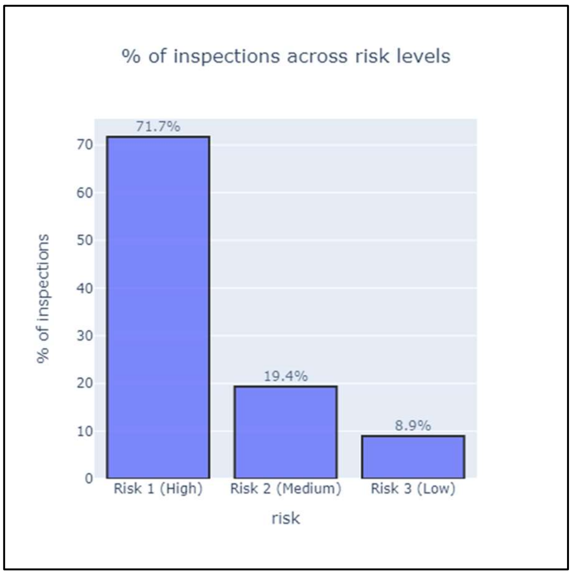
Results:

Inspections Trend:

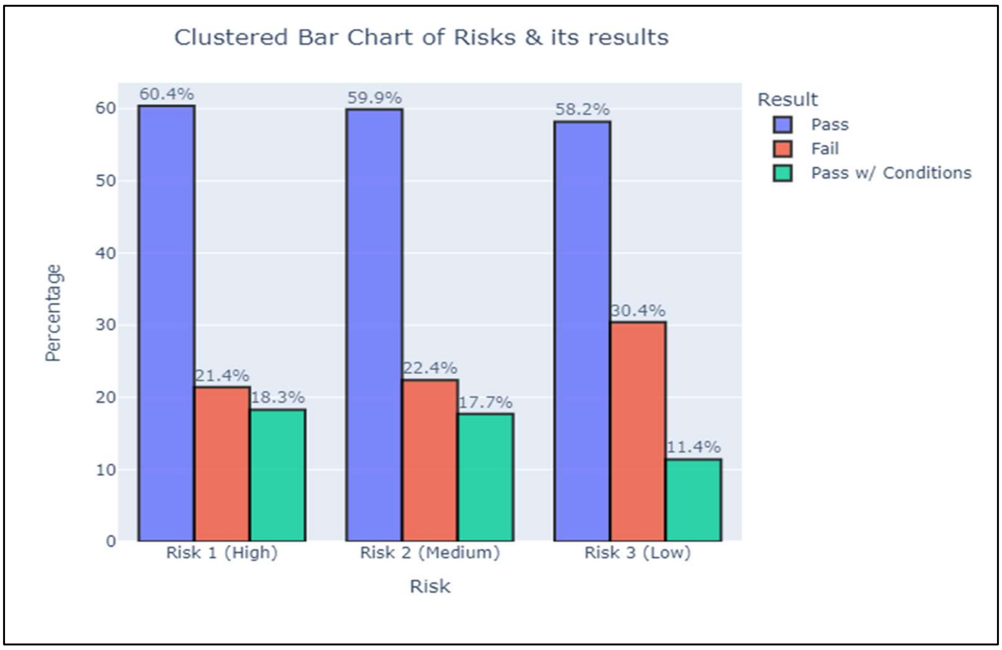


The chart presents a timeline of monthly inspection of food establishments starting from 2010. Initially, the number of inspections fluctuates around 1400 to 1800 per month, with occasional peaks. From around 2016, there's an increase in variability, with counts occasionally surging past 2000 inspections per month. From 2019 onwards, there is a clear downward trend, with inspections sharply decreasing to below 1200.

Risk Analysis:

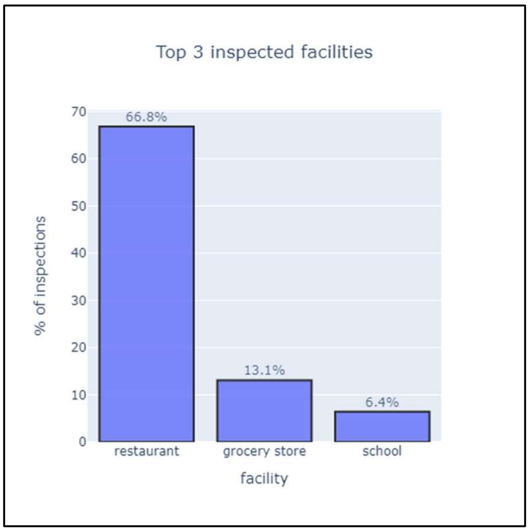


The bar chart illustrates the distribution of inspections across different risk categories assigned to food establishments. The data shows a predominant focus on high-risk establishments, which account for 71.7% of all inspections. This suggests that health inspectors prioritize venues with a greater potential for food safety issues. Medium-risk establishments receive a smaller share of attention, with 19.4% of inspections. Lastly, low-risk establishments, constitute only 8.9% of inspections.

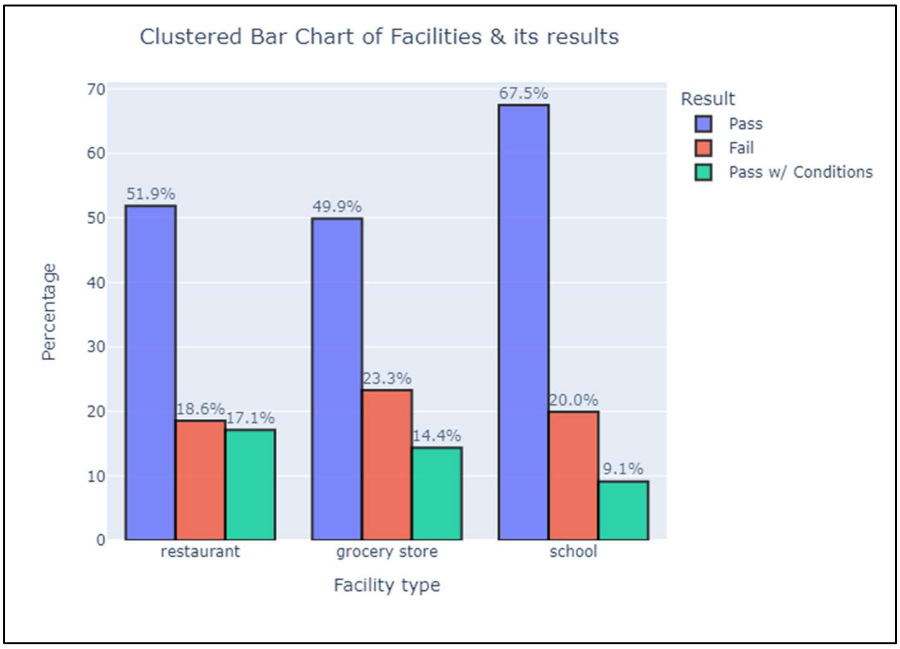


The clustered bar chart presents the outcomes of food establishment inspections categorized by risk level and further divided into three results: pass, fail, and pass with conditions. High-risk establishments have a pass rate of 60.4% but also show a considerable proportion of conditional passes (18.3%) and failures (21.4%), indicating that while many meet the standards, a significant number have issues that need addressing. Medium-risk establishments exhibit a similar pattern, with a slightly lower pass rate of 59.9%, but also a lower failure rate at 17.7%. Low-risk establishments show a different trend, with a pass rate slightly reduced to 58.2%, yet they have a notably higher rate of conditional passes (30.4%) and the lowest failure rate (11.4%).

Facility types:

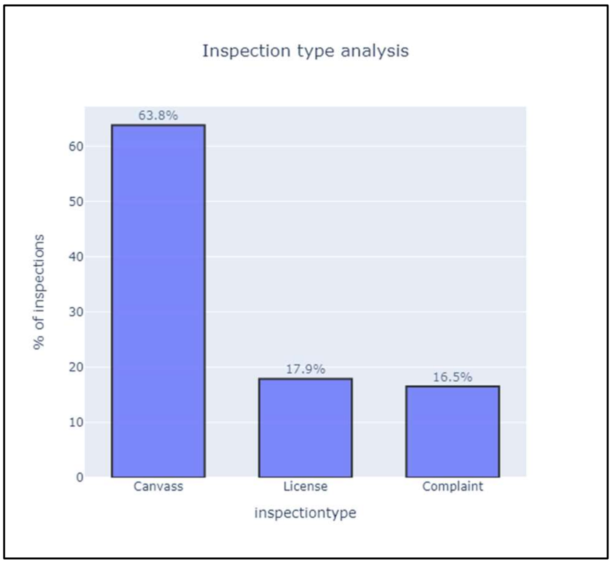


The bar graph showcases the proportional distribution of inspections across the top three types of facilities: restaurants, grocery stores, and schools. Restaurants lead by a significant margin, representing 66.8% of all inspections, grocery stores are the next most inspected facility type, comprising 13.1% of inspections. Schools account for 6.4% of inspections, the lowest among the three. This distribution indicates a prioritized allocation of inspection resources towards restaurants, acknowledging their critical impact on public health.

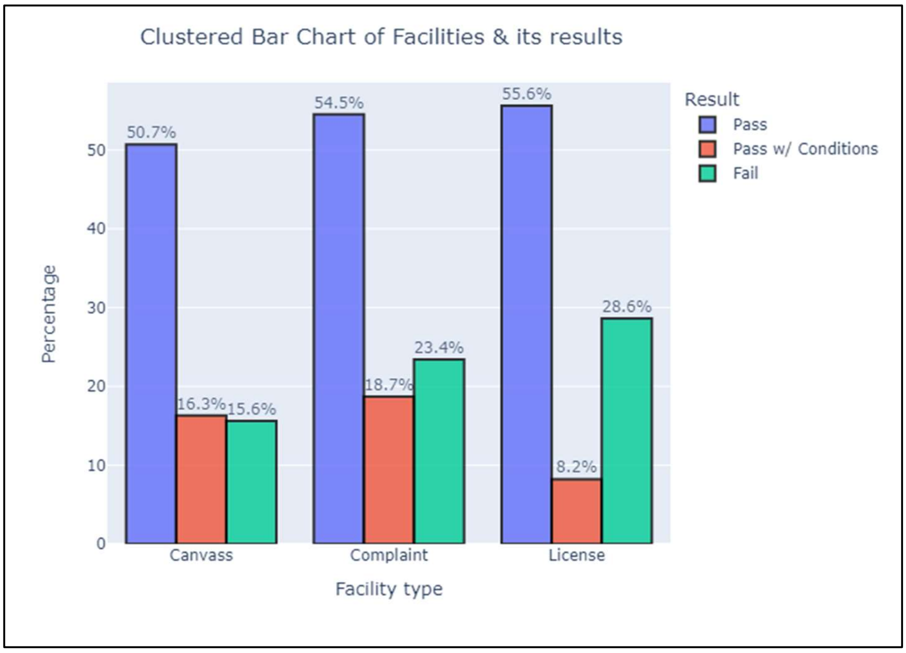


The clustered bar chart provides a comparison of inspection results across three types of facilities: restaurants, grocery stores, and schools. Restaurants show a pass rate of 51.9% but also have a relatively high percentage of both failures (18.6%) and conditional passes (17.1%), suggesting that while more than half meet health standards, a significant portion face challenges in compliance. Grocery stores present a nearly even pass rate at 49.9% but exhibit a lower failure rate (14.4%) and a higher conditional pass rate (23.3%) compared to restaurants. Schools demonstrate the highest pass rate at 67.5%, alongside a lower failure rate (9.1%) and a moderate rate of conditional passes (20.0%)

Inspection types:

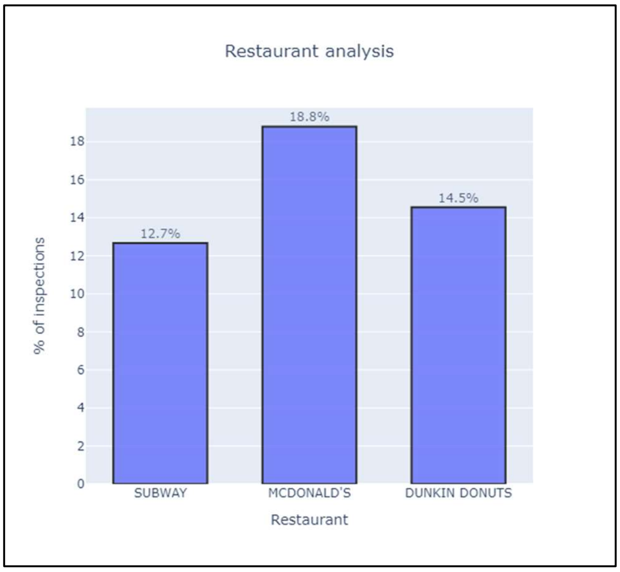


The bar graph shows the top three types of inspections conducted, categorized as Canvass, License, and Complaint. Canvass inspections, which are likely routine and proactive checks, constitute the majority at 63.8%. License inspections. account for 17.9%, suggesting a significant but smaller proportion of the inspection activity. Complaint inspections make up 16.5% of the total.



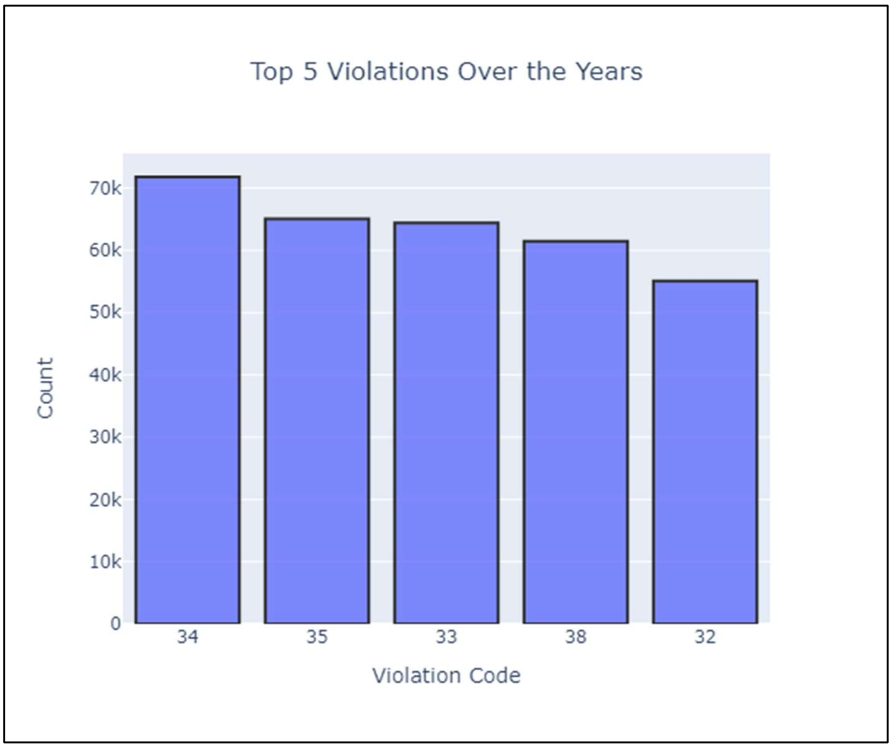
The clustered bar chart compares the results of inspections based on three types: Canvass, Complaint, and License. Canvass inspections result in a 50.7% outright pass rate, with 15.6% failing and a notable 16.3% passing with conditions, implying that while half of the establishments are deemed fully compliant during routine checks, a significant portion requires further action or monitoring. Inspections triggered by complaints have a lower pass rate at 54.5%, and a higher failure rate at 18.7%, indicating that issues raised by the public or employees tend to be valid concerns. License-related inspections, which may be linked to the application or renewal process, show a higher pass rate of 55.6%, but a substantial 28.6% pass with conditions and the lowest failure rate at 8.2%.

Analysis of Restaurants:



The bar chart provides an analysis of the failure percentages during inspections. McDonald's leads with the highest failure rate at 18.8%. Dunkin' Donuts follows with a 14.5% failure rate. Subway has the lowest failure rate among the three at 12.7%, which could indicate better compliance with regulations compared to other two food chains.

Analysis of Violations:



The bar chart illustrates the frequency of the top five most repeated violation codes over a span of years. Violation code 34 is the most frequent, with occurrences nearing 70,000, suggesting it might represent a common issue that establishments consistently struggle with. Codes 35 and 33 follow closely, indicating that they are also prevalent issues within the industry. Code 38 shows slightly fewer instances but remains among the top violations, while code 32 has the least occurrences among the five but is significant enough to be a leading concern.

Conclusions:

In conclusion, our analysis has provided valuable insights into the factors influencing the outcomes of food establishment inspections. It is evident that the number of violations, citations, and the type of inspection conducted play significant roles in determining the inspection outcomes.

Our initial hypothesis, which stated that the attributes of food establishments impact inspection outcomes, has been confirmed as true. These findings emphasize the importance of maintaining high standards and compliance within food establishments to ensure positive inspection outcomes and, ultimately, the safety of consumers. Further research and data-driven strategies may be explored to enhance inspection processes and promote food safety across the industry.

Teamwork:

Vikranth Reddy - Responsible for data cleansing, employing various Python libraries to ensure data consistency and accuracy. He also played a key role in preparing the presentation and final reports.

Phanindra Anagam – Performed a comprehensive Literature review and conducted Exploratory Data Analysis (EDA) and Descriptive Analytics, examining various dataset columns in relation to the outcomes. He identified significant trends within the data.

Ritesh Vesalapu – Worked on predictive analytics and developed several machine learning models and evaluated their accuracy by analyzing different parameters.

References:

Barnes, J., Whiley, H., Ross, K., & Smith, J. (2022). Defining Food Safety Inspection. *International journal of environmental research and public health*, *19*(2), 789. https://doi.org/10.3390/ijerph19020789

Singh, Shubham, Shah, Bhuvni, Kanich, Chris, & Kash, Ian A. *Fair Decision-Making for Food Inspections. EAAMO (2022). Equity and Access in Algorithms, Mechanisms, and Optimization, 1-11.* https://doi.org/10.1145/3551624.3555289

Marvin, H. J., Janssen, E. M., Bouzembrak, Y., Hendriksen, P. J., & Staats, M. (2017). ("Sci-Hub | Big data in food safety: An overview. Critical Reviews in ...") Big data in food safety: An overview. *Critical reviews in food science and nutrition*, *57*(11), 2286–2295. https://doi.org/10.1080/10408398.2016.1257481