Food Inspection Analysis using R

A Minor Project Report
submitted in partial fulfillment of the requirements for
the award of the degree of

Bachelor of Engineering

in

Artificial Intelligence and Data Science

By

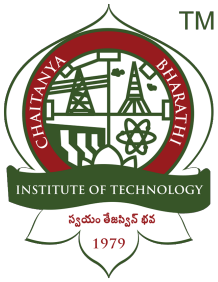
Aayushi Kar, G.Sai Akshitha, Rimsha Fatima

(1601-21-771-001), (1601-21-771-008), (1601-21-771-017)

Under the esteemed guidance of

Smt. T. Satya Kiranmai

Assistant Professor



DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY
HYDERABAD – 500075
MAY 2024

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY
HYDERABAD – 500075

INSTITUTE VISION

"To be the center of excellence in technical education and research".

INSTITUTE MISSION

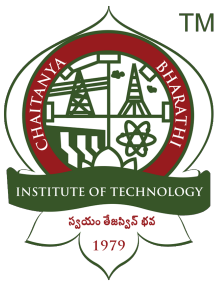
"To address the emerging needs through quality technical education and advanced research".

DEPARTMENT VISION

"To be a globally recognized center of excellence in the field of Artificial Intelligence and Data Science that produces innovative pioneers and research experts capable of addressing complex real-world challenges and contributing to the socio-economic development of the nation."

DEPARTMENT MISSION

1. To provide cutting-edge education in the field of Artificial Intelligence and Data Science that is rooted in ethical and moral values.
2. To establish strong partnerships with industries and research organizations in the field of Artificial Intelligence and Data Science, and to excel in the emerging areas of research by creating innovative solutions.
3. To cultivate a strong sense of social responsibility among students, fostering their inclination to utilize their knowledge and skills for the betterment of society.
4. To motivate and mentor students to become trailblazers in Artificial Intelligence and Data Science, and develop an entrepreneurial mindset that nurtures innovation and creativity.



DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY
HYDERABAD – 500075

DECLARATION CERTIFICATE

We hereby declare that the project titled **Food Inspection Analysis using R** submitted by us to the **Artificial Intelligence and Data Science CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY, HYDERABAD** in partial fulfillment of the requirements for the award of **Bachelor of Engineering** is a bonafide record of the work carried out by us under the supervision of **Smt. T. Satya Kiranmai** . We further declare that the work reported in this project, has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma of this institute or of any other institute or University.

Project Associates

Aayushi Kar, G.Sai Akshitha, Rimsha Fatima

(1601-21-771-001), (1601-21-771-008), (1601-21-771-017)

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY

HYDERABAD – 500075

BONAFIDE CERTIFICATE

This is to certify that the project titled **Food Inspection Analysis using R** is a bonafide record of the work done by

Aayushi Kar, G.Sai Akshitha, Rimsha Fatima

(1601-21-771-001), (1601-21-771-008), (1601-21-771-017)

in partial fulfillment of the requirements for the award of the degree of **Bachelor of Engineering** in **Artificial Intelligence and Data Science** to the **CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY, HYDERABAD** carried out under my guidance and supervision during the year 2023-24. The results presented in this project report have not been submitted to any other university or Institute for the award of any degree.

Smt. T. Satya Kiranmai
Guide

Dr. Kadiyala Ramana
Head of the Department

Submitted for Semester Minor-Project viva-voce examination held on _____

Examiner-1

Examiner-2

ABSTRACT

Food safety is a critical concern in the food industry, particularly with the increasing number of food-serving establishments and the potential risks of food-borne illnesses. This project, 'Food Inspection Analysis using R' employs a data-driven approach to analyze and enhance food inspection practices. Through the analysis of three comprehensive datasets, we focus on different phases of food inspection, including initial food ratings, reports of adverse effects, and final inspection details.

Our analysis includes preprocessing, exploratory analysis, and statistical tests to improve food safety practices and public awareness. By informing stakeholders, policymakers, and the public about food safety issues, we aim to contribute to evidence-based decision-making in the food industry. This project addresses the challenges faced by government organizations in supervising compliance with food safety codes and emphasizes the importance of data analytics in supporting effective food safety practices and public health.

Keywords : R, Food Inspection Analysis

ACKNOWLEDGEMENTS

We would like to express our deepest gratitude to the following people for guiding us through this course and without whom this project and the results achieved from it would not have reached completion.

Smt. T. Satya Kiranmai, Assistant Professor, Department of Artificial Intelligence and Data Science, for helping us and guiding us in the course of this project. Without his/her guidance, we would not have been able to successfully complete this project. His/Her patience and genial attitude is and always will be a source of inspiration to us.

Dr. Kadiyala Ramana, the Head of the Department, Department of Artificial Intelligence and Data Science, for allowing us to avail the facilities at the department.

We are also thankful to the faculty and staff members of the Department of Artificial Intelligence and Data Science, our individual parents and our friends for their constant support and help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

This project ”Food Inspection Analysis using R” is dedicated to tackling the critical issue of food safety within the food industry, where the proliferation of food-serving establishments has amplified the risks associated with food-borne illnesses. Our project, adopts a data-driven approach to scrutinize and enhance food inspection practices, with a keen focus on key metrics such as initial food ratings and final inspection details and risk categories. Amidst the challenges posed by urbanization, the supervision of compliance with food safety codes has become increasingly arduous for government organizations, paving the way for potential health and economic repercussions stemming from food-borne illnesses.

In response, our solution approach involves harnessing comprehensive datasets and leveraging advanced analytical techniques in R to unveil insights that can bolster food safety practices. Through meticulous preprocessing, exploratory analysis, and statistical tests, our project aims not only to identify areas of improvement but also to offer actionable recommendations for mitigating risks and enhancing overall food safety standards.

The impact of our endeavor extends beyond the confines of the food industry, as it seeks to inform stakeholders, policymakers, and the public alike about prevailing food safety issues.

1.2 Problem Statement

Performing Food Inspection analysis using R to improve the food safety standards. By fostering evidence-based decision-making and contributing to broader public health efforts, we aspire to catalyze positive change and safeguard the well-being of consumers and communities alike.

1.3 Organization of Project

1) Defining the Problem: The project begins by precisely outlining the problem statement and objectives, focusing on key questions to be addressed and the scope of analysis. This step ensures clarity and direction, preventing unnecessary deviations during the project.

2) Task Allocation: Responsibilities are assigned to team members based on their expertise, covering data collection, preprocessing, exploratory analysis, model development, and visualization. This allocation optimizes productivity and ensures effective contribution from each team member.

3) Technology Stack and Tools: R programming language is utilized for data analysis, while data is sourced from XLS worksheet. Tools such as Git for version control and GitHub for collaboration are selected to facilitate seamless communication and workflow management.

4) Data Quality and Model Performance Checks: Continuous evaluation of data quality and model performance is conducted to ensure accuracy and reliability. Checks are implemented to address issues like missing values and outliers, while model performance metrics are monitored and evaluated.

5) Classification: Classification models such as Random Forest, Support Vector Classifier, Naive Bayes are employed for predicting grade given to restaurant based on inspection type, seating type, risk category its been classified as. This modeling enhances decision-making and strategy formulation.

6) Visualization and Reporting: Visualizations, including charts and graphs, are created to present insights clearly. Comprehensive reports summarizing findings, methodologies, and recommendations are prepared to empower stakeholders with informed decision-making capabilities.

CHAPTER 2

LITERATURE SURVEY

[1] Association between food control inspection grades and regional incidence of infectious foodborne diseases in Finland

- **Year: 2023**
- **Publication: International Journal of Environmental Health Research**
- **Tasks Performed:** Lower food control inspection grades correlated with higher rates of Salmonella infections. This was especially true for inspections that focused on cleanliness of facilities and equipment.Similar trends were observed for Campylobacter infections, with lower grades on food storage and hygiene practices linked to increased illness.
- **Results Obtained:** The results of this study provide evidence for an association between food control inspection grades and foodborne diseases, especially Campylobacter and Salmonella infections. An increase in disease incidence was observed when food control grades were inferior, indicating that food control recognizes non-compliances that may predispose to foodborne diseases. Means of intervention for non-compliances to reduce the incidence of foodborne diseases should be developed.
- **Gaps Reported:** Conclusions about causality cannot be made, a retrospective case-control study comparing FBOs from which foodborne diseases originated and those from which there were not must be done to understand that.

[2] Predicting Food Safety Compliance for Informed Food Outlet Inspections: A Machine Learning Approach

- **Year: 2021**
- **Publication: International Journal of Environmental Health Research**
- **Tasks Performed:**Food businesses are opening and people eat out more, but inspections haven't kept up, leaving consumers at risk. AI predicts risky food outlets using neighborhood data and machine learning to identify areas with a higher chance of having non-compliant food businesses, allowing inspectors to focus their limited resources.

- **Results Obtained:** The results of this study provide evidence for an association between food control inspection grades and foodborne diseases, especially Campylobacter and Salmonella infections. An increase in disease incidence was observed when food control grades were inferior, indicating that food control recognizes non-compliances that may predispose to foodborne diseases. Means of intervention for non-compliances to reduce the incidence of foodborne diseases should be developed.

- **Gaps Reported:** The model doesn't consider factors that change over time, which could improve its accuracy. It doesn't use past inspection results (unavailable in their data), which could be helpful predictors.

[3] Understanding the Relationships Between Inspection Results and Risk of Foodborne Illness in Restaurants

- **Year: 2016**

- **Publication: PubMed**

- **Tasks Performed:** The study investigated the relationship between restaurant inspection results and the risk of foodborne disease outbreaks, specifically focusing on a large Salmonella outbreak in Illinois linked to a chain of restaurants (Chain A). Inspection data were collected from 106 Chain A establishments, with 46 outbreak cases linked to 23 of these restaurants.

- **Results Obtained:** The analysis found no significant differences in overall demerit points or points for hand washing and cross-contamination between outbreak and non-outbreak restaurants.

- **Gaps Reported:** The study concluded that the outbreak was likely due to a contaminated fresh produce item from a commercial source, which routine inspections are not designed to detect. This suggests that inspection results alone may not predict or prevent outbreaks, emphasizing the need to consider specific pathogen and food item pairings and transmission routes.

[4] Results of routine inspections in restaurants and institutional catering establishments associated with foodborne outbreaks in Finland

- **Year: 2022**

- **Publication: International Journal of Environmental Health Research**

- **Tasks Performed: The study analyzed inspection results for institutional catering and restaurants to identify differences between outbreak and control establishments.**

- **Results Obtained:** In institutional catering, outbreak establishments had significantly poorer grades in cleanliness, shelf-life management, and hygiene proficiency compared to controls. For restaurants, no significant differences were found in most items, though outbreak establishments showed poorer sampling practices and better scores for work clothes and hand hygiene when including weaker evidence outbreaks.
- **Gaps Reported:** Overall, the ratio of outbreak establishments was higher in restaurants than in institutional catering.

[5] Relationship Between Food Safety and Critical Violations on Restaurant Inspections: An Empirical Investigation of Bacterial Pathogen Content

- **Year: 2013**
- **Publication: Journal Environmental Health**
- **Tasks Performed:** The study examined the relationship between critical restaurant inspection violations and food safety by measuring bacterial pathogens in foods from poorly and well-performing restaurants in Jefferson County, Alabama.
- **Results Obtained:** Results showed 35.7 percent of food samples had detectable Staphylococcus aureus, with no difference between the two groups, and 45.2 percent were received outside recommended temperatures. This highlights the need for improved temperature control and hygienic practices, especially handwashing. The presence of S. aureus suggests its commonality and underscores the importance of proper hygiene.
- **Gaps Reported:** Future research should focus on characteristics linked to critical hygiene violations to enhance inspection and educational practices.

CHAPTER 3

METHODOLOGY

3.1 Software Requirements

- Operating System: Windows 7 or above, Mac OS, Linux
- Visual Studio Code
- RStudio

3.2 Hardware Requirements

- x86 64-bit CPU (Intel / AMD architecture)
- 4 GB RAM
- 5 GB free disk space

3.3 Workflow

1) Download libraries/packages involved:

- dplyr: Streamlines data manipulation using verbs (like filter, mutate) for a tidy workflow.
- tidyr: Reshapes data from wide to long format and vice versa for easier analysis.
- ggplot2: Creates various types of elegant and customizable visualizations.
- stats: Provides a wide range of statistical functions for data analysis.
- utils: Offers general utility functions for R environment management.
- methods: Provides generic function methods for R objects.
- caret: A suite of tools for classification, regression, and other machine learning tasks.
- e1071: Offers various machine learning algorithms including support vector machines.

- nnet: Implements neural network models for machine learning.
- randomforest: Provides functions for creating and analyzing random forest models, a popular machine learning technique.

2) Data Preprocessing: We remove the attributes that are not significant for reaching a logical conclusion.

3) Model Building: We have developed Random Forest, SVC, Neural Networks and Naive Bayes models.

4) Statistical Analysis of Models

CHAPTER 4

IMPLEMENTATION

4.1 Data Preprocessing

The data set contains 271645 observations and 14 variables. Following is a table that explains the data variables of the dataset.

Variable	Value
Name	Name of the restaurant
Description	Seating type and Risk category
Inspection-date	Date of inspection DD/MM/YYYY
Inspection-type	Type of inspection. for example, Routine
Inspection-result	Satisfactory, unsatisfactory, complete
Inspection-score	Numeric value
Inspection-closed-business	true or false
Violation Type	Blue, Red
Violation points	Numeric value
Violation-description	description of violation
City	City name
Longitude	longitude
Latitude	latitude
Grade	Numeric

Table 4.1: Variable Table

	Name	Inspection.Date	Type	Risk_Category	City	Longitude	Latitude
1	#7064 ARCO AM/PM	02/08/2018	Seating 0-12	Risk Category III	AUBURN	-122.2216	47.33971
2	@ THE SHACK, LLC	10/03/2018	Seating 0-12	Risk Category III	Seattle	-122.3709	47.57043
3	@ THE SHACK, LLC	10/03/2018	Seating 0-12	Risk Category III	Seattle	-122.3709	47.57043
4	@ THE SHACK, LLC	04/02/2018	Seating 0-12	Risk Category III	Seattle	-122.3709	47.57043
5	@ THE SHACK, LLC	07/27/2017	Seating 0-12	Risk Category III	Seattle	-122.3709	47.57043
6	@ THE SHACK, LLC	06/16/2017	Seating 0-12	Risk Category III	Seattle	-122.3709	47.57043

Figure 4.1: Data Head in R[**data-head1**]

	Inspection.Type	Inspection.Score	Inspection.Result	Inspection.Closed.Business	Violation.Type
1	Routine Inspection/Field Review	0	Satisfactory	false	
2	Routine Inspection/Field Review	10	Satisfactory	false	BLUE
3	Routine Inspection/Field Review	10	Satisfactory	false	BLUE
4	Consultation/Education - Field	0	Complete	false	
5	Routine Inspection/Field Review	0	Satisfactory	false	
6	Routine Inspection/Field Review	0	Satisfactory	false	

Figure 4.2: Data Head in R[**data-head2**]

	Violation.Description	Violation.Points	Grade
1		0	1
2	4100 - Warewashing facilities properly installed,...	5	1
3	3400 - Wiping cloths properly used, stored, proper sanitizer	5	1
4		0	1
5		0	1
6		0	1

Figure 4.3: Data Head in R[**data-head3**]

To begin with, lets look at the city column. Here the cities attribute had to be modified some of the cities were in lower case some were in upper case and some had random spacing between the letters for them . So we changed all the cities to the same case and also removed all the spacing between them so it could be recognized well.We used the function gsub and trimnws.

gsub():- is a versatile function in R that stands for ”global substitution”. It is primarily used to find and replace patterns within strings.

Here’s how `gsub()` works: gsub(pattern, replacement, x)

1. Pattern Matching: First, you provide a pattern that you want to search for within a string. This pattern can be a regular expression, providing flexible matching options.

2. Replacement: You specify what you want to replace the matched pattern with.

3. Global Search and Replace: Unlike the `sub()` function, which only replaces the first occurrence of the pattern, `gsub()` replaces all occurrences of the pattern within the string.

`gsub()` is particularly useful for cleaning and manipulating text data, especially when dealing with large datasets or when you need to make multiple replacements within a single string.

`trimws()` is a handy function in R used for trimming whitespace (spaces, tabs, and newlines) from the beginning, end, or both ends of a character string.

Here’s a breakdown of how `trimws()` works: trimws(x, which = c(”both”, ”left”, ”right”), whitespace = ”[\t\r\n]”)

1. Trimming Whitespace: `trimws()` removes leading (at the beginning), trailing (at the end), or both leading and trailing whitespace from a character string.

2. Handling Other White Characters: Apart from spaces, `trimws()` also trims tabs (`\t`) and newlines (`\n`) by default.

3. Flexible Usage: `trimws()` is particularly useful for cleaning and preparing textual data, especially when dealing with messy or unformatted data. It helps ensure consistency and readability.

`trimws()` is a convenient function for quickly cleaning up strings in R, making them more suitable for further analysis or presentation.

There were no null values in the cities attribute. However there was a ”(none)” value which was replaced with ”Unknown” to indicate an unknown city in the dataset.

```
 [1] "AUBURN"          "Seattle"         "SEATTLE"         "KENT"            "BELLEVUE"        "KENMORE"
 [7] "Bellevue"        "Issaquah"        "Bothell"         "ALGONA"          "WOODINVILLE"     "Kent"
[13] "DES MOINES"      "Renton"          "ISSAQUAH"        "Auburn"          "KIRKLAND"        "Kirkland"
[19] "BURIEN"          "FEDERAL WAY"     "TUKWILA"         "RENTON"          "SHORELINE"       "Redmond"
[25] "Woodinville"     "SAMMAMISH"       "Federal Way"     "BLACK DIAMOND"   "REDMOND"         "SNOQUALMIE"
[31] "COVINGTON"       "LAKE FOREST PARK" "ENUMCLAW"       "MAPLE VALLEY"    "BOTHELL"         "NORMANDY PARK"
[37] "MERCER ISLAND"   "NORTH BEND"      "DUVALL"          "NEWCASTLE"       "SEA TAC"         "VASHON ISLAND"
[43] "SEATAC"          "Pacific"         "Mercer Island"   "North Bend"      "CLYDE HILL"      "SNOQUALMIE PASS"
[49] "PACIFIC"         "Black Diamond"   "Snoqualmie"      "Carnation"       "CARNATION"       "Enumclaw"
[55] "Duvall"          "Vashon"          "Skykomish"       "SNOHOMISH"       "Maple Valley"    "KING COUNTY"
[61] "Fall City"       "PRESTON"         "LYNNWOOD"        "HOBART"          "TUKWIULA"        "EVERETT"
[67] "RAVENSDALE"      "FALL CITY"       "MEDINA"          "Medina"          "Ravensdale"      "MUKILTEO"
[73] "TACOMA"          "SKYKOMISH"       "(none)"          "NORTHBEND"
```

Figure 4.4: Unique cities before pre-processing[**cities˙bp**]

```
      algona         auburn       bellevue  blackdiamond        bothell         burien      carnation      clydehill
         212           7503          19257           307           2328           3999            524             76
   covington      desmoines         duvall      enumclaw       everett       fallcity    federalway         hobart
        1729           2230            963          2352            26            409          11101             15
    issaquah       kenmore           kent     kingcounty      kirkland  lakeforestpark       lynnwood    maplevalley
        5880           1635          13482             2         10626            510             7           2126
     medina  mercerisland       mukilteo      newcastle  normandypark      northbend       pacific        preston
         135           1320             11            801           438           1616           334            70
 ravensdale       redmond         renton      sammamish        seatac        seattle      shoreline      skykomish
          54          10774          11181           2152          2469         137057          4788           121
    snohomish    snoqualmie snoqualmiepass         tacoma        tukwila       tukwiula       Unknown         vashon
            3          1090             38            411          4556              1            35           231
vashonisland    woodinville
          964          3213
```

Figure 4.5: Unique cities with counts after pre-processing[**cities˙ap**]

For the attribute inspection result we categorised the it to only only four results namely, satisfactory , unsatisfactory, complete and incomplete by considering their corresponding results.

Following snippet shows the values in Inspection-result attribute and count of observations for each grade.

	1	2	3	4
Baseline Data	63	39	3	0
Complete	23433	12865	1590	137
Confirmed	1	2	0	0
Exchange information	1	0	0	0
In Compliance	0	3	0	0
Incomplete	252	151	21	2
Increased Knowledge	1	0	0	0
Needs Assessment	0	1	0	0
No Longer At Location	4	1	0	0
Not Accessible	93	68	8	0
Not Applicable	129	62	8	0
Not Confirmed	7	3	0	0
Not In Compliance	1	0	0	0
Not Permitted	0	1	0	0
Not Ready For Inspection	45	18	3	0
Not Tested	1	0	0	0
Out of Business	1	1	0	0
Satisfactory	45005	17887	2019	114
Unsatisfactory	44663	52719	9569	1109

Figure 4.6: No of observations for a inspection result for every grade[**inspection-result˙bp**]

After categorizing the results we get only 4 values in the Inspection-result attribute.

Complete	Incomplete	Satisfactory	Unsatisfactory
46378	1590	96945	126249

Figure 4.7: No of observations per inspection-result[**inspection-result˙ap**]

From the above table we have known that the description variable combines both seating type and risk category. We have separated it to Type and Risk Category variables to perform analysis on it. We have also renamed values in Type attribute for easy understanding.

Following are the before and after results of cleaning the description attribute form the original data set.

Description
Seating 0-12 - Risk Category III
Seating 0-12 - Risk Category III
Seating 0-12 - Risk Category III
Seating 0-12 - Risk Category III
Seating 0-12 - Risk Category III
Seating 0-12 - Risk Category III

Figure 4.8: Description Attribute[**description**]

```
Risk Category I  Risk Category II Risk Category III
          16765             35460            218879
```

Figure 4.9: Risk Category Attribute[**risk category**]

We then created a No Seating value under Type attribute for all those observations that specified ”no-seating-values” for further analysis. We also removed the null values from the Grade column. Following are the plots that display the relationship between Type and Grade attribute.

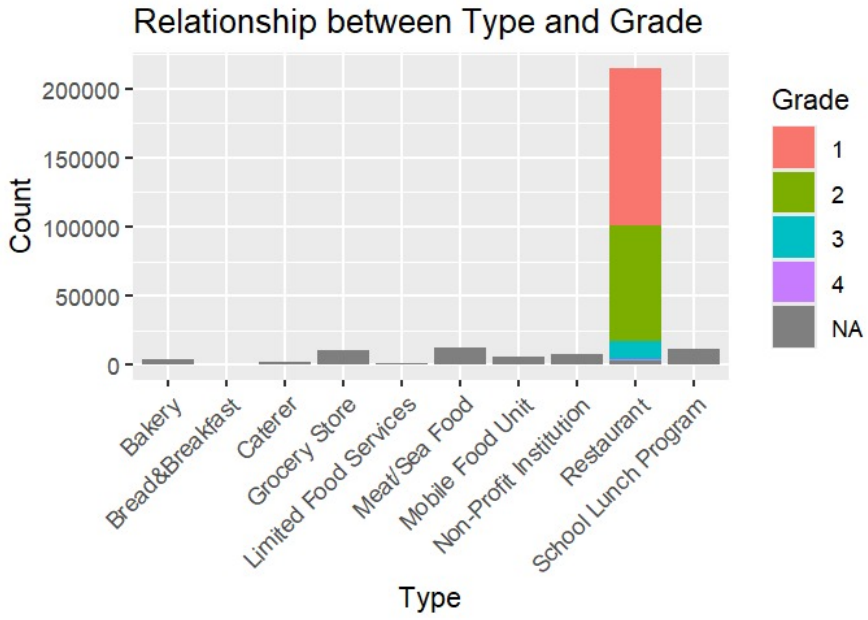


Figure 4.10: Relationship between Type and Grade[**TypeGrade**]

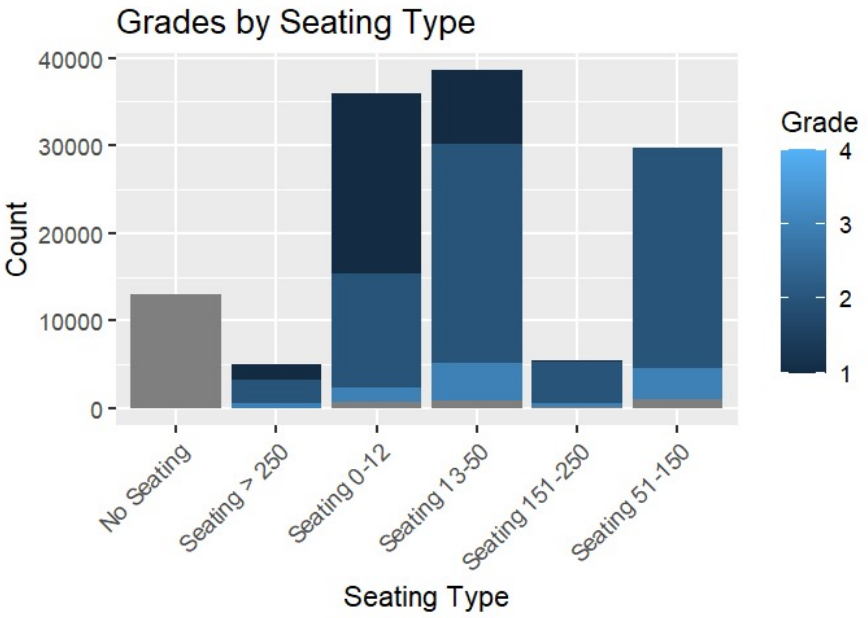


Figure 4.11: Grade by seating type[**grade˙by˙seating˙type**]

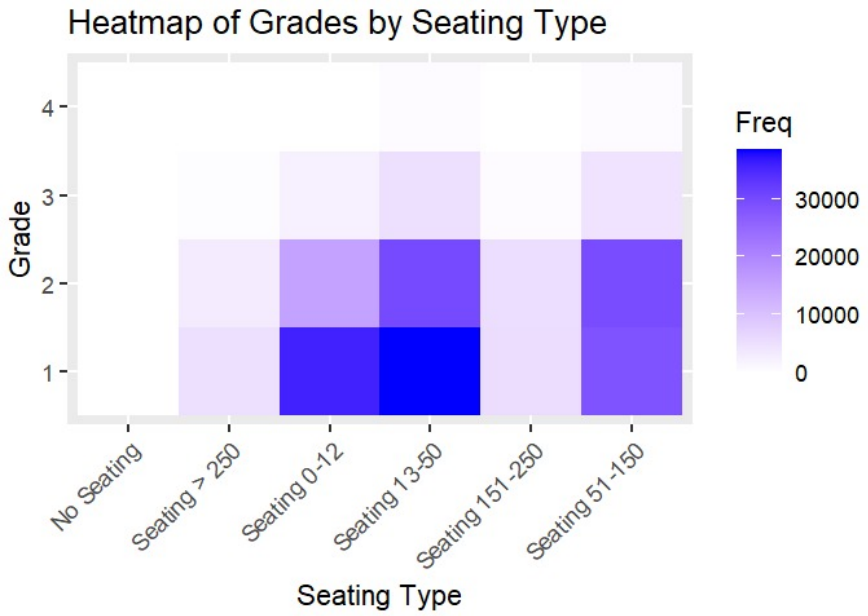


Figure 4.12: Heatmap of Grades by seating type[**heatmap˙grade˙type**]

Additionally, we have removed the observations which had negative numbers and NA in Inspection-score variable.

The violation type variable in the dataset had BLUE and RED values indicating the severity of the violation. However, we found that a significant number of observations did not violate the guidelines and thus did not have any value. These observations were given the value WHITE suggesting that the inspection officer concluded there was no violation made.

We then converted the values in Risk category variable i. e Risk Category I, Risk Category II, Risk Category III to numeric labels 3,2,1 so as to find its relationship with Grade. Similarly, in the values under Inspection Result i.e satisfactory, complete, incomplete, and unsatisfactory, have also been converted to numeric labels 1,2,3,4 respectively. The values in the Inspection-closed-business was also changed to an integer value. These text to numeric conversions were done with the help of as.numeric() and as.factor() methods.

In R, `as.numeric()` is a function used to coerce or convert a variable into a numeric data type. When you apply `as.numeric()` to a variable, R attempts to coerce the variable into a numeric type. If the variable contains characters representing numeric values, R will convert those characters into numeric values. If the variable contains non-numeric characters, such as letters or special characters, the conversion will result in missing values (`NA`). If the variable is a factor, `as.numeric()` converts the levels of the factor into their corresponding integer codes and then coerces them into numeric values. This can sometimes lead to unexpected results, so it's important to be cautious when converting factors to

numeric.

In R, `as.factor()` is a function used to coerce or convert a variable into a factor data type. Factors are used to represent categorical data in R. When you apply `as.factor()` to a variable, R attempts to coerce the variable into a factor data type. If the variable contains unique values, each unique value will become a level in the factor, and the values in the original variable will be replaced by their corresponding factor levels. Factors have levels, which represent the distinct categories or groups in the data. The levels are ordered based on the order in which they appear in the data, unless otherwise specified.

In the following chapter we will discuss the model building and statistical analysis.

4.2 Model Building

The dataset we are using has 14 variables. We have logically concluded that ”Risk-Category”, ”Inspection.Score”, ”Inspection.Result”, ”Inspection.Closed.Business”, ”Violation.Type”, ”Violation.Points”,”Type” are correlated Grade. The correlation matrix for these values is as follows

```
                                [,1]
Risk_Category             0.15531519
Inspection.Score          0.34466658
Inspection.Result         0.25249933
Inspection.Closed.Business 0.06489072
Violation.Type            0.20836444
Violation.Points          0.19278946
Type                      0.10503237
```

Figure 4.13: Correlation Matrix[**correlation˙matrix**]

From the correlation matrix, its concluded that Inspection.Closed.Business variable does not affect the Grade. Thus, we exclude it for model building.

Since the Grade attribute has 4 values i.e 1, 2, 3, 4, we have selected multi-class classifcation models such as Random Forest, Naive Bayes, SVM. We also built a Neural Network. We have set the training data as 20 percent of the actual dataset. Let’s analyse the results of these models in the next section.

We have used the libraries caret, dplyr, randomForest, e1071 and nnet in R to build RandomForest, NaiveBayes, SVM and Neural Network to train the models.

Upon training the Random Forest model we observed that there were no predictions made for the Grade=4 class since all observations that result to Grade as 4 were included in the test set.

To balance the representation of grade across all classes we have initialised weights to each class.

```
          1           2           3           4
0.536700357 0.394772667 0.061891151 0.006635825
```

Figure 4.14: Class Weights randomly assigned[**Class˙Weights**]

```
         1          2          3          4
0.53670036 0.39477267 0.15000000 0.06635825
```

Figure 4.15: Class Weights adjusted[**Class˙Weights**]

CHAPTER 5

RESULTS

We have used the ”overall” function in ConfusionMatrix object to calculate the total accuracy of the model. However, we iteratively considered each individual class to calculate the Precision, Recall and F1 score for the RandomForest Model.

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2     3     4
         1 18218  9167  1040    55
         2  4251  7391  1414   148
         3   115   203   209    11
         4    18    62    54    22

Overall Statistics

               Accuracy : 0.6098
                 95% CI : (0.6051, 0.6144)
    No Information Rate : 0.5333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2452

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3  Class: 4
Sensitivity            0.8060   0.4393 0.076923 0.0932203
Specificity            0.4811   0.7725 0.991705 0.9968203
Pos Pred Value         0.6397   0.5598 0.388476 0.1410256
Neg Pred Value         0.6846   0.6767 0.940057 0.9949316
Prevalence             0.5333   0.3970 0.064113 0.0055689
Detection Rate         0.4299   0.1744 0.004932 0.0005191
Detection Prevalence   0.6720   0.3116 0.012695 0.0036812
Balanced Accuracy      0.6436   0.6059 0.534314 0.5450203
```

Figure 5.1: Confusion Matrix[**cm**]

```
For Random Forests model
 Class Label    1       2       3       4
 Precision      0.806   0.439   0.077   0.093
 Recall         0.64    0.56    0.388   0.141
 F1 Score       0.713   0.492   0.128   0.112
>
```

Figure 5.2: Random Forest Metrics[**RF˙metrics**]

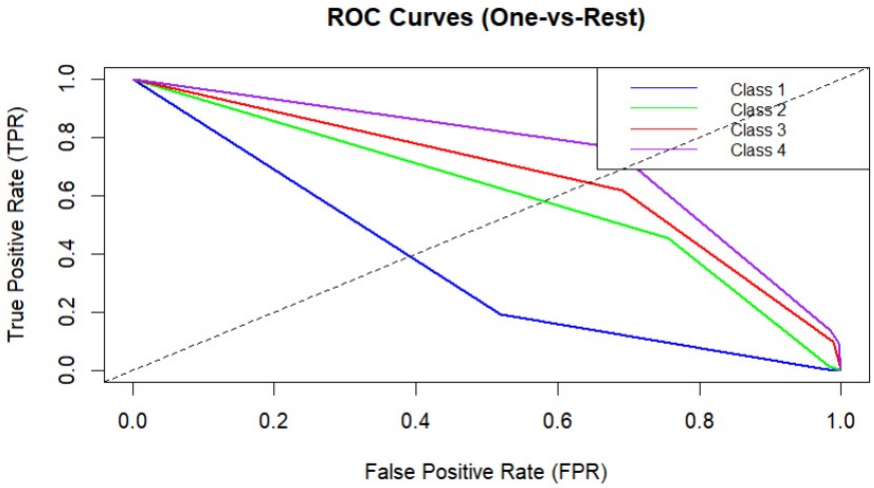


Figure 5.3: ROC[**ROC**]

```
Macro F1 Score for Random Forests: 0.362
```

Figure 5.4: MacroF1 Score[**MacroF1˙RF**]

```
Root Mean Squared Error (RMSE): 0.7075738
```

Figure 5.5: Root Mean Square Error[**RMSE˙RF**]

```
Mean Squared Error (MSE): 0.5006607
```

Figure 5.6: Mean Square Error[**MSE˙RF**]

```
Mean Absolute Error (MAE): 0.425905
```

Figure 5.7: Mean Absolute Error[**MAE˙RF**]

Following is comparision of accuracy for all the models trained on test data.

```
"Random Forests Model accuracy on test set with adjusted class weights: 60.98 %"
```

Figure 5.8: Random Forest Score [**Model˙comparision**]

```
"Neural Network model accuracy on test set: 60.35 %"
```

Figure 5.9: Neural Network Score [**Model˙comparision**]

```
"Naive Bayes model accuracy on test set:  56.57"
```

Figure 5.10: Naive Bayes [**Model˙comparision**]

```
"SVM model accuracy on test set: 60.22"
```

Figure 5.11: Support Vector Classifier Score [**Model˙comparision**]

CHAPTER 6

CONCLUSION

Class Imbalance: The dataset exhibited an imbalance, with underrepresented classes requiring special attention during model training.

Random Forest with Weighted Classes: To address this imbalance, we employed a Random Forest model with adjusted class weights, leading to the best overall performance.

Model Evaluation: We evaluated model performance using several metrics, including:

(i) Accuracy: Random Forest achieved the highest accuracy in predicting inspection grades.

(ii) F1 Score (Macro): These metrics provided a balanced view of model performance across all classes.

(iii) Confusion Matrix: This matrix visualized the distribution of predicted vs. actual grades.

(iv) Other Metrics: We considered additional metrics like precision, recall, RMSE, MSE, and MAE, depending on the model type.

ROC curve : The threshold are given by the roc() function which is taken from the pROC package by using 1 vs rest strategy

SVC and Naïve Bayes: While explored, Support Vector Machines (SVM) and Naïve Bayes models yielded lower accuracy compared to Random Forest.

CHAPTER 7

FUTURE SCOPE

1. Expand Data Sources: Go beyond traditional inspection reports. Include real-time outbreak data from health agencies to identify emerging threats. Analyze consumer reviews on food quality and hygiene to uncover potential issues. Leverage social media sentiment analysis to understand public concerns and identify establishments generating negative buzz around food safety.

2. Advanced Analytics: Harness the power of AI. Develop models that analyze inspection data, consumer reviews, and outbreak information to predict establishments with a high risk of food safety violations. Automate report analysis to identify recurring issues and prioritize inspections. Use network analysis to understand connections between suppliers, distributors, and establishments, pinpointing potential weak links in the food chain.

3. Public Engagement: Empower consumers! Create a user-friendly platform displaying inspection results, hygiene ratings, and consumer reviews. Partner with food delivery apps to showcase this information alongside menus, allowing users to make informed choices.

4. Policy and Regulation: Turn data into action. Analyze data to identify gaps in regulations or areas needing stricter enforcement. Recommend data-driven policy improvements to address emerging food safety challenges. Advocate for standardized data collection across food safety agencies to enable comprehensive analysis.

5. Scalability and Replication: Share the knowledge! Develop a framework for replicating this project in other regions. Create a global database of food safety information, fostering collaboration between countries and building a data-driven ecosystem for ensuring safe food everywhere.

CHAPTER 8

BIBLIOGRAPHY

[1] Kosola, M. et al. (2024) 'Association between food control inspection grades and regional incidence of infectious foodborne diseases in Finland', International Journal of Environmental Health Research, 34(2), pp. 885–897. doi: 10.1080/09603123.2023.2183942.

[2] Oldroyd, Rachel A., Michelle A. Morris, and Mark Birkin. 2021. "Predicting Food Safety Compliance for Informed Food Outlet Inspections: A Machine Learning Approach" International Journal of Environmental Research and Public Health 18, no. 23: 12635. https://doi.org/10.3390/ijerph182312635

[3] Lee P, Hedberg CW. Understanding the Relationships Between Inspection Results and Risk of Foodborne Illness in Restaurants. Foodborne Pathog Dis. 2016 Oct;13(10):582-586. doi: 10.1089/fpd.2016.2137. Epub 2016 Sep 28. PMID: 27680283.

[4] Leinonen, E., Kaskela, J., Keto-Timonen, R., Lundén, J. (2023). Results of routine inspections in restaurants and institutional catering establishments associated with foodborne outbreaks in Finland. International Journal of Environmental Health Research, 33(6), 588–599. https://doi.org/10.1080/09603123.2022.2041563.

[5] Yeager VA, Menachemi N, Braden B, Taylor DM, Manzella B, Ouimet C. Relationship between food safety and critical violations on restaurant inspections: an empirical investigation of bacterial pathogen content. J Environ Health. 2013;75(6):68-73.

[6] Huang A, de la Mora Velasco E, Farhangi A, Bilgihan A, Jahromi MF. Leveraging data analytics to understand the relationship between restaurants' safety violations and COVID-19 transmission [published correction appears in Int J Hosp Manag. 2022 Oct;107:103328]. Int J Hosp Manag. 2022;104:103241.

[7] Kimberly J. Harris, Kevin S. Murphy, Robin B. DiPietro, Gretchen L. Rivera, Food safety inspections results: A comparison of ethnic-operated restaurants to non-ethnic-operated restaurants, International Journal of Hospitality Management, Volume 46, 2015, Pages 190-199, ISSN 0278-4319, https://doi.org/10.1016/j.ijhm.2015.02.004.

[8] Charlotte Yapp, Robyn Fairman, Factors affecting food safety compliance within small and medium-sized enterprises: implications for regulatory and enforcement strategies, Food Control, Volume 17, Issue 1, 2006, Pages 42-51, ISSN 0956-7135, https://doi.org/10.1016/j.foodcont.2004.08.007.

[9] Irwin K, Ballard J, Grendon J, Kobayashi J. 1989. Results of routine restau-

rant inspections can predict outbreaks of foodborne illness: the Seattle-King County experience. Am J Public Health. 79(5):586–590. doi:10.2105/AJPH.79.5.586.

[10] Jones TF, Pavlin BI, LaFleur BJ, Ingram A, Schaffner W. 2004. Restaurant inspection scores and foodborne disease. Emerg Infect Dis. 10(4):688–692. doi:10.3201/eid1004.030343.