**University of Maryland, Baltimore County**

**Department of Data Science,**

**DATA 602: Introduction to Data Analysis and Machine Learning**

**Project Topic:**

**Analyzing and Forecasting Food Safety Compliance in Chicago: Predicting Violations and Future Trends**

**By**

**Vishal Balaji Kotrike (Campus Id: EE36010)**

**Under the esteemed guidance of**

**Professor Devin Fensterheim**

**December 19th, 2023**

**Problem Statement:** The recent Salmonella outbreak associated with a Chicago taqueria, affecting over 55 individuals, underscores a critical food safety challenge in the city. This incident highlights how important it is to have strict oversight and preventative measures in place to protect health of Chicago residents and safety when they eat out. Our project's objectives are to investigate the outbreak, pinpoint its origins, and evaluate any gaps in food safety procedures that might have contributed to the event. In this project, we are going to achieve the following:

Predictions and Forecasts:

* Develop a model to predict the Inspection outcomes.
* Develop a model to assess the risk levels associated with different facility types.
* Choose the classification model that best fits this dataset by comparing them.
* Displaying the output and conclusions of the model

Analysis:

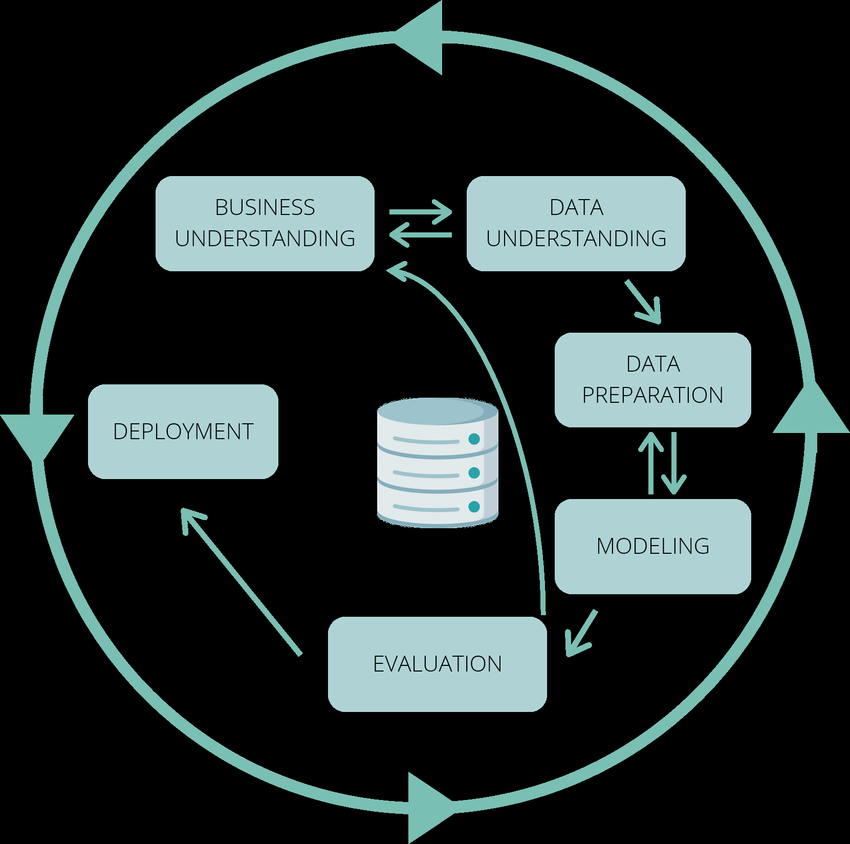
* Identify the most common facility type in the dataset and facility types that are associated with highest and lowest risk factors.
* Create geospatial maps, visualizing the distribution of food establishments and their risk levels across different locations.
* Perform a time series analysis on the dataset, considering features such as Inspection Date.
* Investigate the correlation between different features (e.g., Facility Type, Inspection Type) and the risk level.
* Examine the types of violations reported and their frequency.

**Data Source:**

Our primary data source is the "Food Inspections" dataset provided by the City of Chicago, encompassing information from inspections of local restaurants and food establishments conducted since January 1, 2010. These inspections, conducted by professionals from the Chicago Department of Public Health's Food Protection Program, follow standardized procedures and maintain stringent quality control. Inspection results are methodically entered into a database, subsequently reviewed, and approved by State of Illinois Licensed Environmental Health Practitioners (LEHPs) to ensure data accuracy and regulatory compliance.

**URL:** [Food Inspections | City of Chicago | Data Portal](https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5)

**Project Approach:** For this project, we have adopted the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology.

****

**Business Understanding** defines the business problem, objectives, and requirements for our project.

**Data Understanding** helped us explore and familiarize with the available data to assess its quality, relevance, and potential for analysis.

**Data Preparation** is to clean, transform, and preprocess the data to make it suitable for modeling and analysis.

**Modeling** helps to build and train predictive or descriptive models based on the prepared data.

**Evaluation** assesses the performance and effectiveness of the models in meeting the business objectives.

**Deployment** integrates the selected model into the business environment and makes sure its proper functioning for decision-making.

**Results**:

**By Analysing the past data:**

* More than half of the inspections resulted in Pass across all the facility types.
* The facility types of Risk 1(High) category have the highest Pass percentage.
* Restaurants are the most frequently inspected facility type by the inspection team.
* High Risk is consistently observed in the Restaurants facility type.
* Despite an approximately equal likelihood of receiving either a Pass or Fail result during inspections, it's noteworthy that high-risk levels persist even when inspection results indicate a Pass.
* By doing geospatial analysis we were able to visualise and locate the facility types having high risk.
* By doing Time Series Analysis, We observed that the Frequency of Inspections getting reduced gradually from 2016 and Inspection counts are getting reduced during Winter (Nov, Dec, Jan, Feb)

A graph with a line

Description automatically generated A graph showing a line

Description automatically generated

* Average number of Inspections happening in a month and the week in Restaurants are:

Monthly Average Inspections - 1055

Weekly Average Inspections – 244

* Restaurants has the High Risk factor at the same time, the pass percentage is also more.
* Violation Code 32(Food and Non-Food Contact Surfaces Properly Designed, Constructed and Maintained) has occurred the most across all the inspections.

**Predictions and Forecast:**

**Model 1: Building a model to assess the Inspection Results.**

* Our target feature is Inspection\_Result ('Pass' 'Not Ready' 'Fail' 'Out of Business' 'No Entry' ‘Pass w/ Conditions' 'Business Not Located')
* After cleaning the data, we have applied label encoding. We applied PCA techniques to reduce the columns having multiple unique values.
* We have used various models, and the accuracy scores are listed as below:

|  |  |  |
| --- | --- | --- |
| **Model Used** | **Accuracy score (train)** | **Accuracy score (test)** |
| Random Forest Classifier | 0.69 | 0.69 |
| Decision Tree Classifier | 0.69 | 0.68 |
| XG Boost Classifier | 0.68 | 0.68 |
| K-Nearest Neighbours (KNN) | 0.66 | 0.65 |
| Support Vector Machine (SVM) | 0.69 | 0.68 |
| Voting Ensemble Model | 0.68 | 0.68 |
| Stacking Classifier | 0.68 | 0.67 |

* Among all the models, Random Forest Classifier performed well.

**Model 2: Building a model to assess the risk levels associated with different facility types.**

* Our target feature is Risk (Risk 1, Risk 2, Risk 3)
* After cleaning the data, we have applied label encoding. We applied PCA techniques to reduce the columns having multiple unique values.
* We have used various models and the accuracy scores are listed as below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Used** | **Accuracy score (train)** | **Accuracy score (test)** | **F1 Score (train)** | **F1 Score (test)** |
| Random Forest Classifier | 0.78 | 0.78 | 0.74 | 0.74 |
| Decision Tree Classifier | 0.78 | 0.78 | 0.74 | 0.74 |
| XG Boost Classifier | 0.77 | 0.77 | 0.73 | 0.73 |

* Among all the models, Random Forest Classifier performed well and had less execution time.

**Conclusion:**

The Food Inspection can figure out which places need more attention to keep food safe. Additionally, they will know beforehand what problems might happen in certain areas, making inspections better. Furthermore, they can plan when and where to check places to use their time wisely. Also, they will be able to predict issues in specific spots, so they can help those places do better before any big problems.