# FINAL PROJECT REPORT

---

## Team HAL9000


| | |
|---|---|
| **Project Title** | LA Restaurant Health Inspection and Recommendation System |
| **Date Started** | 01/29/2023 |
| **Date Completed** | 04/24/2023 |
| **Project Sponsor** | Dr. Anna Farzindar *(USC)* & Dr. Alex Liu *(RMDS Lab)* |



**Team Details**

Ziyue Chen, Xiaoyi Gu, Zhenmin Hua, Ying Wang

{ziyueche, xiaoyigu, zhenmin, ywang490}@usc.edu

# Table of Contents

# Chapter 1

## Executive Summary

With the development of human society and economy, the number of restaurants is mounting and public health issue have attracted more and more attention, especially since the onset of the COVID-19 pandemic. Meanwhile, restaurant health inspections become significant for both the local governments and citizens as the number of restaurants has surges in recent years. For the governments, the inspection is a source of cost and efficient resource allocation is needed. For the citizens, they are threatened by direct and indirect contact in restaurants, however, information on restaurant health inspection can be hard to find on social media platforms. Although current smart city projects have already covered many topics of public health, few have researched restaurant health inspection.



Therefore, our project is aimed at filling this gap. Specifically, we will focus on restaurant inspection in Los Angeles and collect data from LA open data and Yelp. This project will propose 3 models to solve the mentioned pain points, including a health risk prediction model, a restaurant segmentation model, and a restaurant recommendation model. We will mainly use machine learning techniques for modeling. The previous two models are designed to improve the efficiency of restaurant inspection, hence optimizing the LA government's resource allocation. The last one will provide LA citizens the insights into restaurant health conditions and offer recommendations based on their preferences. We will also visualize the results and evaluate the performance of the models with predefined metrics.

# Chapter 2

## Project Objectives

- Fetch restaurant features, such as food type, location, price, rating, and review, from Yelp, treat them as potential contributing factors to health inspection scores, and integrate them with the health inspection dataset using common data cleaning and analysis techniques.
- From the combined data, use machine learning strategies to fit and validate various models. Choices include both supervised and unsupervised learning, including regression analysis, neural networks, cluster analysis, etc. We expect to extract valid predictors and their weights on restaurant scores so that they can be used for future predictions on restaurant qualities and health risks. In addition, we will visualize the segmentation of restaurants to provide insights into the adjustment of inspection frequencies.
- Lastly, build a recommendation system using application development tools. The app will allow users to apply different filters based on their needs and to obtain information about their preferred restaurants. The system will recommend similar restaurants based on their preferences.

# Chapter 3

## Lean Six Sigma Report

### 3.1 Define Phase

The first phase of the Lean Six Sigma improvement methodology is Define. During this Phase, our team created a high-level map of the process and began investigating the needs of the process's consumers.

#### 3.1.1 Customer Satisfaction

- Citizen 1 (A person who dines out a lot)
    - He loves the idea of improving the government's inspection work by predicting the health risk of restaurants. He thinks food safety and restaurant health is a really important issue to solve at hand. This app will help make restaurants' sanitary conditions better in the future. He's quite optimistic about it.
    - Also, he is supportive of us combining the health data with existing Yelp data to make restaurant recommendations. However, he points out his concern that restaurants' score on Yelp somehow reflects their sanitary condition in a way. Because if the health condition is really bad in a dining place, its score cannot be high on Yelp.
- Citizen 2 (A person who works at a healthcare delivery company and tries different restaurants regularly)

    - The person was surprised to see this topic as she hasn't seen others analyzing it. She suggested that we try various models, which is exactly our plan so that a more accurate result can be achieved.
    - Secondly, she said the final models can be made for the public instead of for the government only. It sounds interesting to play the model around and to see how a prediction can head to another direction when contributing features are changed.
    - As for the application, she said if the sanitary data is integrated with Yelp or Google Map, then she is willing to download and use it. Simply listing health inspection info. sounds not attractive enough. In addition, some people might not care that much about the health conditions of a restaurant whereas some others take them seriously, so we should take this difference into account while designing the application.
- Citizen 3 (A person who is seeking a restaurant)
    - The person was very glad to hear the idea and expressed her need for such an application as she often ordered outside. She thought it would be very useful if the application could integrate the inspection and the reviews.
    - Meanwhile, she was concerned about the reliability of the data and hoped that the reviews or information should be accurate and not fake.
    - She suggested that the application should offer citizens to leave their own comments. Besides, she advised that it would be better if the application could offer the improvement or change of the inspection result in a certain time period to give users a better insight into the performance of the restaurants. Because some restaurants might spend a lot of energy to fix the problems and these

restaurants, from her perspective, should be trusted. And she also believed that showing the change in the Yelp review score might be helpful.

    - She would prefer a free application instead of a paid one.

- Restaurant (The owner of a restaurant)
    - The interviewee said that he had mixed feelings when he heard about our project since it's both good and bad for him.
    - On the one hand, it means that he will pay more attention to restaurant hygiene and also spend more money and time.
    - However, on the other hand, he said that he had confidence in his restaurant and believed that there would be more customers after our project was released to the public.
    - Besides, he also said that the clustering part would be helpful for him to improve his restaurant by using those with higher rates as a reference.
    - In terms of application, he prefers a specialized version for restaurant owners rather than the universal one for all customers and restaurant owners.

### 3.1.2 Process Map

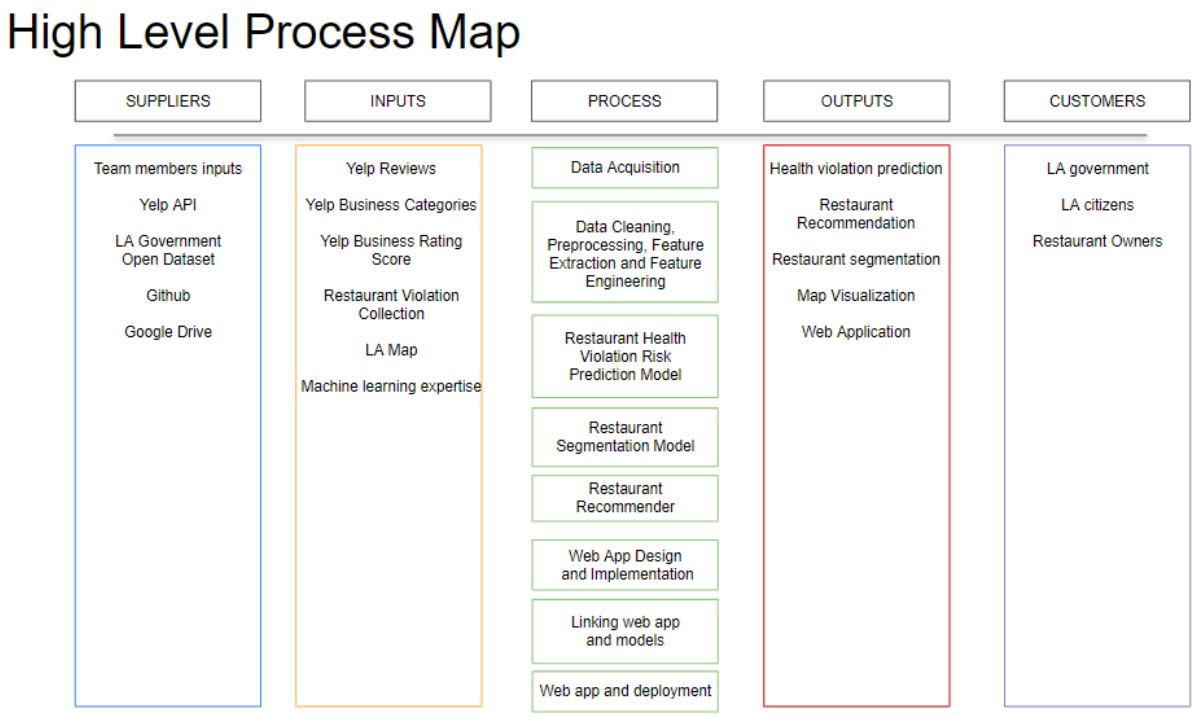



Figure 3.1: SIPOC Diagram

## 3.2 Measure Phase

Based on the project inputs, the Measure phase takes about 2 to 3 weeks. The cooperation of all relevant stakeholders, in particular, is critical in obtaining high-quality data. The measure phase focuses on establishing a baseline for the present process, collecting data, verifying the measurement system, and determining the process capabilities.

### 3.2.1 Process Mapping

Several process maps were created to show the workflow for this project.

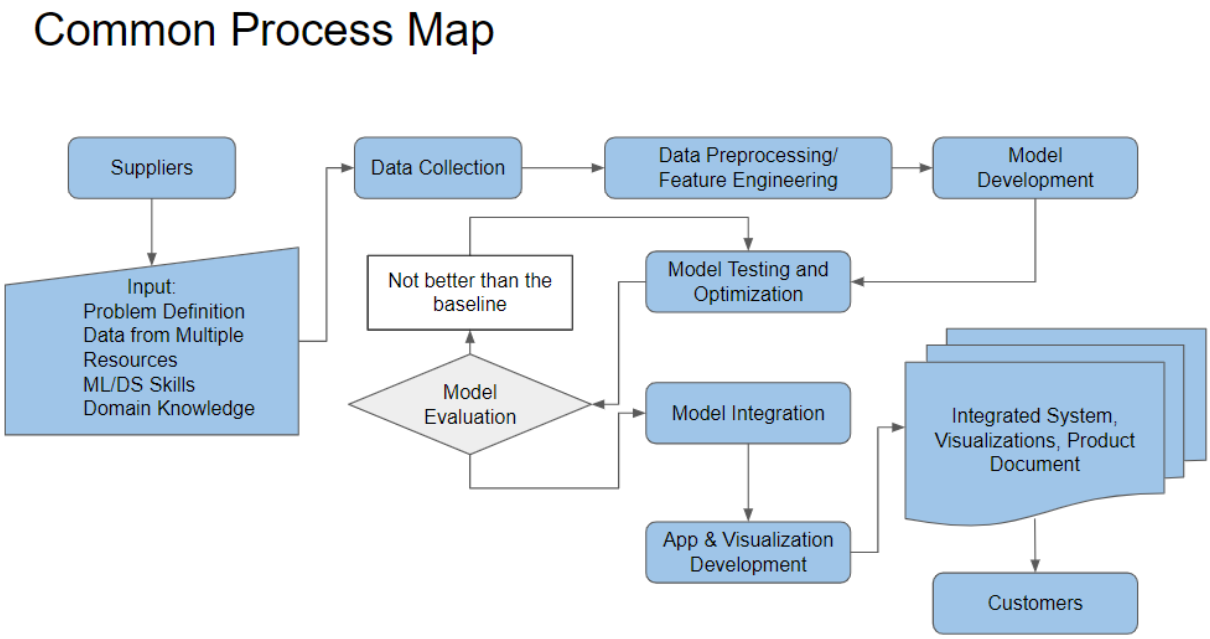



Figure 3.2: Common Process Map

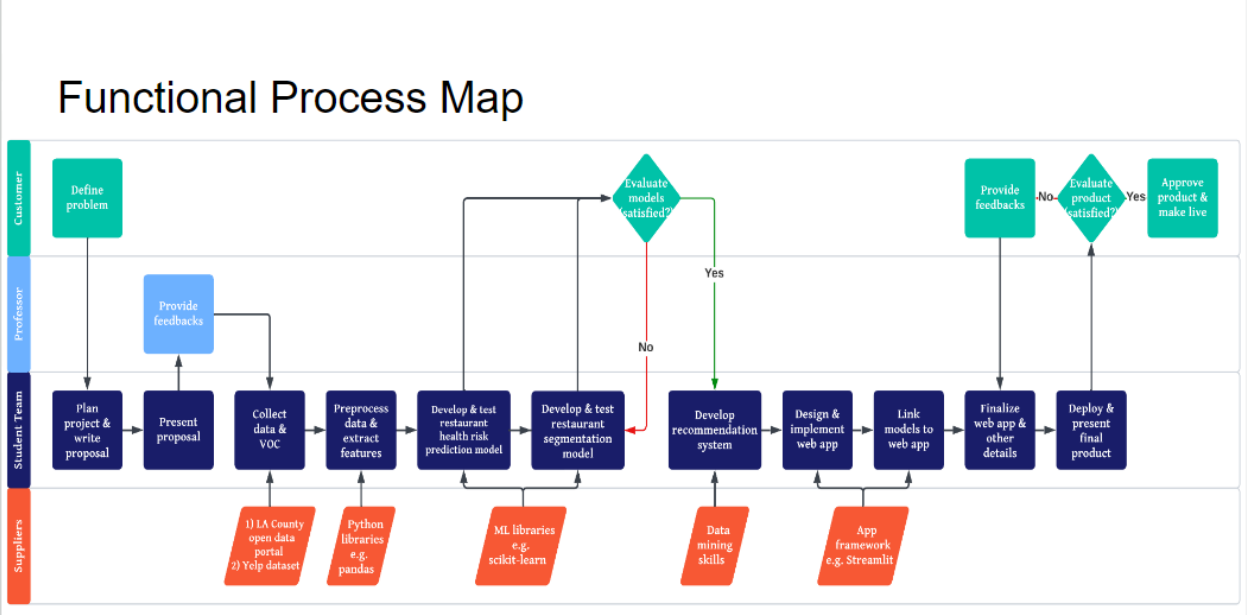



Figure 3.3: Functional Process Map

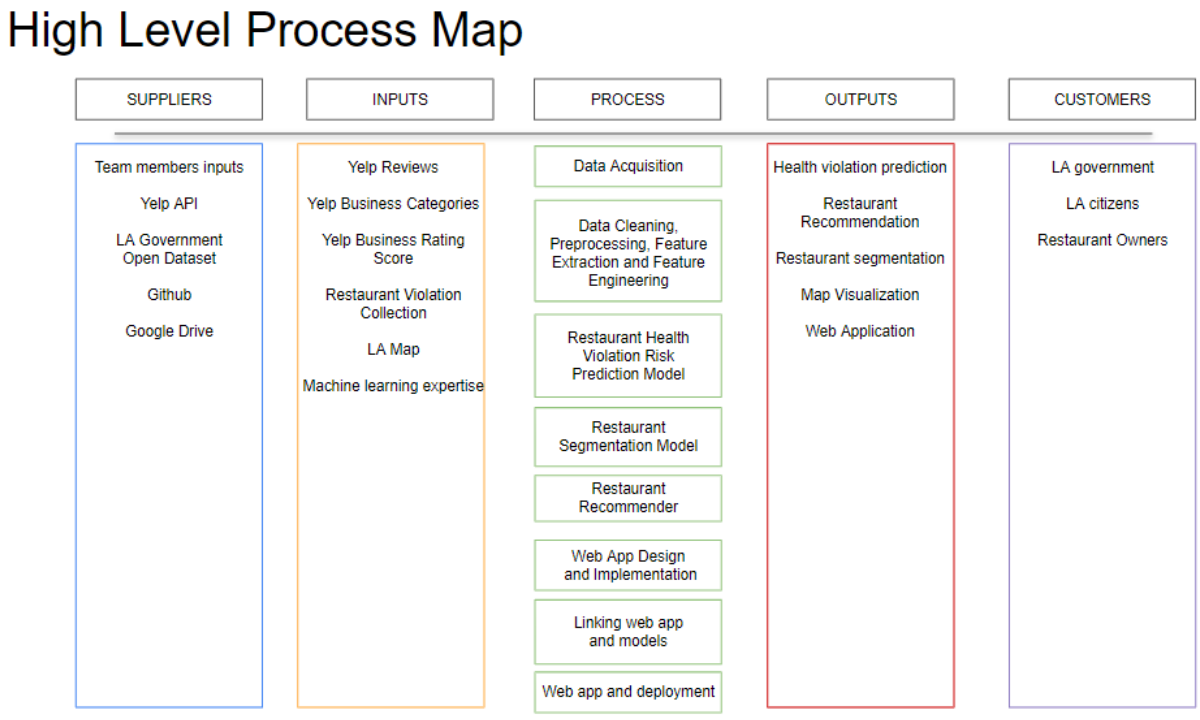



Figure 3.4: High Level Process Map

### 3.2.2 Data Exploration and Preparation

**Data Exploration and Preparation**

- **Data collection 1**: Public Health LOS ANGELES COUNTY RESTAURANT AND MARKET INSPECTIONS (LA Open Data Portal) (https://data.lacounty.gov/datasets/lacounty::public-health-los-angeles-county-restaurant-and-market-inspections-/about)

  Key attributes: activity_date, facility_name, pe_description (violation), facility_address, facility_city, facility_state, facility_zip, score
- **Data collection 2**: Restaurant and Market Health Violations (LA Open Data Portal) https://data.lacity.org/Community-Economic-Development/Restaurant-and-Market-Health-Violations/ckya-qgys

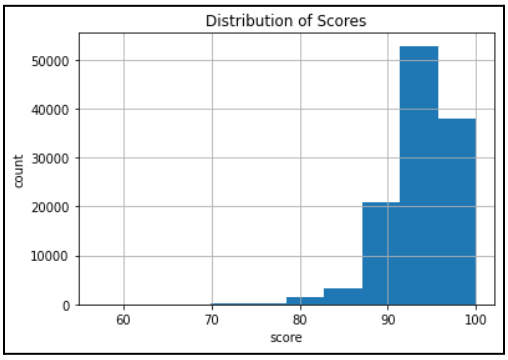
  Key attributes: serial_number, facility_name, grade, violation_code, violation_description
- **Data collection 3**: Restaurant features crawled from social media platforms such as Yelp

  Key attributes: name, address, city, state, zip code, score, and rating
- **Data Preprocessing**
  - Select features (facility info, score, grade, etc.)
  - Merge records (multiple inspections over one restaurant)
  - Join tables (two inspection datasets)
  - Dealing with NULL, duplicates, outliers
  - Add labels for aggregate analysis
    - violation_flag: 0 for no violation, 1 for violation
    - type: restaurant, food market, other (caterer, feeding site, food stand)

- size: 0-30, 31-60, 61-150, 151+
- risk_level: low, moderate, high

- **Exploratory Data Analysis**

a. Distribution of scores (histogram, key statistics e.g. mean, median, standard deviation)

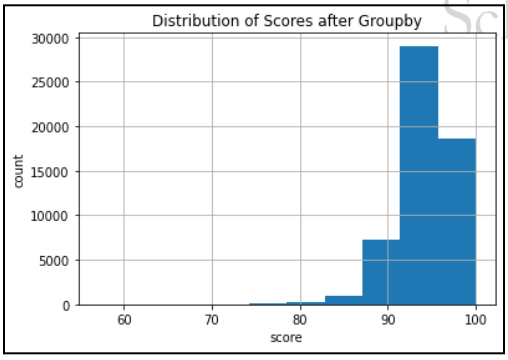



```
count    116705.000000
mean         93.743590
std           3.526726
min          57.000000
25%          92.000000
50%          94.000000
75%          96.000000
max         100.000000
Name: SCORE, dtype: float64
```

Figure 3.5 & 3.6: Distribution of Scores



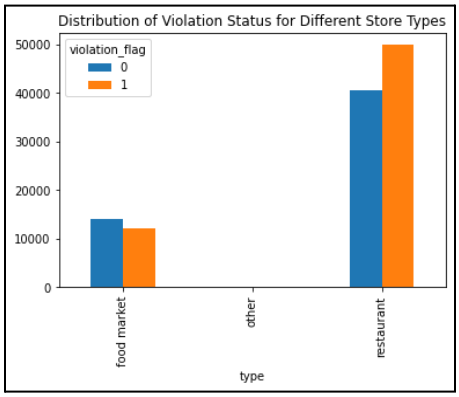
b. Distribution of scores - after averaging scores for the same facility




```
count    56190.000000
mean        94.191471
std          2.989388
min         57.000000
25%         92.500000
50%         94.500000
75%         96.333333
max        100.000000
Name: SCORE, dtype: float64
```

Figure 3.7 & 3.8: Distribution of Scores after Averaging

c. Distribution of violation status for different store type




| type | violation_flag | |
|---|---|---|
| food market | 0 | 14155 |
| | 1 | 12120 |
| other | 0 | 51 |
| | 1 | 10 |
| restaurant | 0 | 40508 |
| | 1 | 49861 |

Figure 3.9 & 3.10: Distribution of Violation Status

d. Distribution of violation status - restaurant only

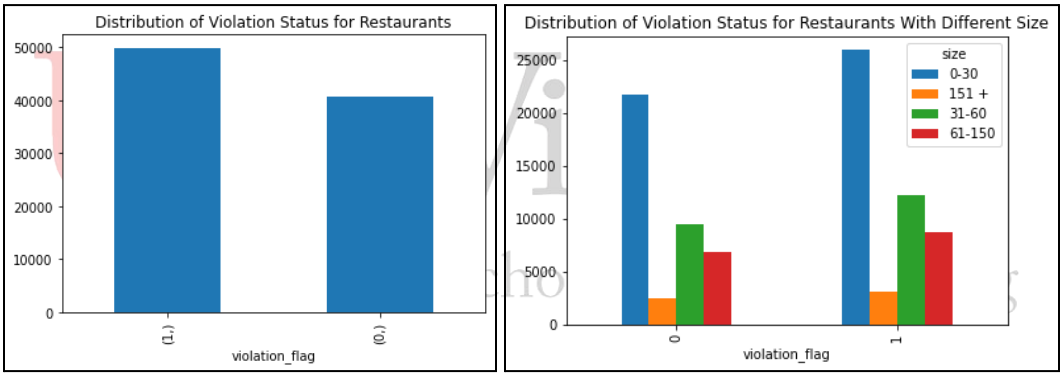



Figure 3.11 & 3.12: Distribution of Violation Status - Restaurant Only

e. Distribution of risk level - restaurant only

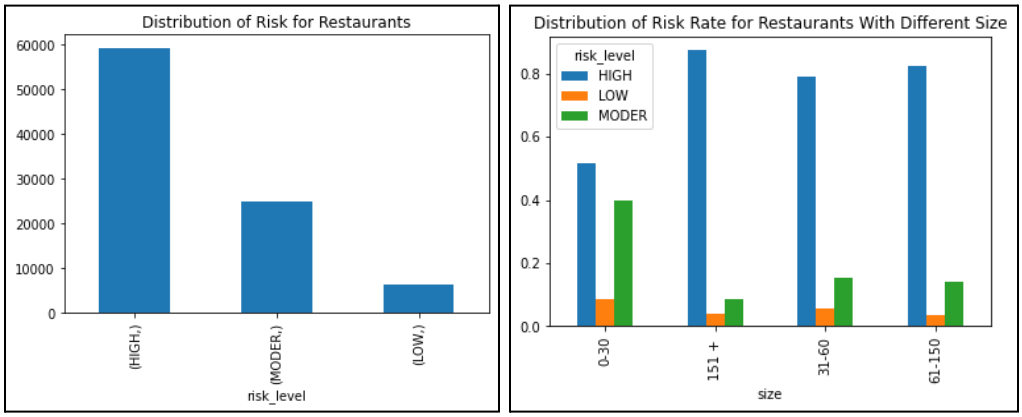



Figure 3.13 & 3.14: Distribution of Risk Level - Restaurant Only

## 3.3 Analysis Phase

The Analysis phase focuses on assessing the process maps, defining potential defects, identifying sources of variation, and determining the critical process parameters for better performance of the models.

### 3.3.1 Selecting charts for Analysis

In order to identify defects in our process, we implemented a set of charts for root cause analysis. The first one is a Pareto chart, which is a graphical tool that introduces the “80/20 Rule” and is able to prioritize quality problems in a process. We paid extra attention to the top problems as they were the most significant ones and could help us formulate corresponding improvements. The second one is a Fishbone diagram which can represent the relationship between a problem and its causes. The third one is 5 Whys, in which we asked at least 5 whys for major defects until we got the root causes. In the end, we collected all identified root causes and kept paying attention to them in the following steps.



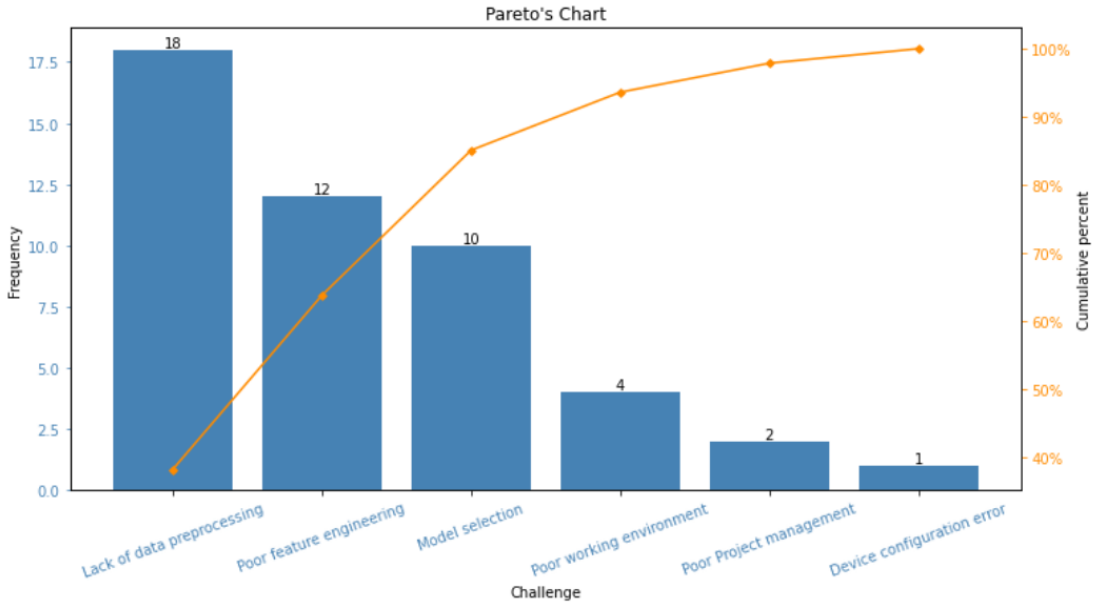



Figure 3.15 : Pareto Chart

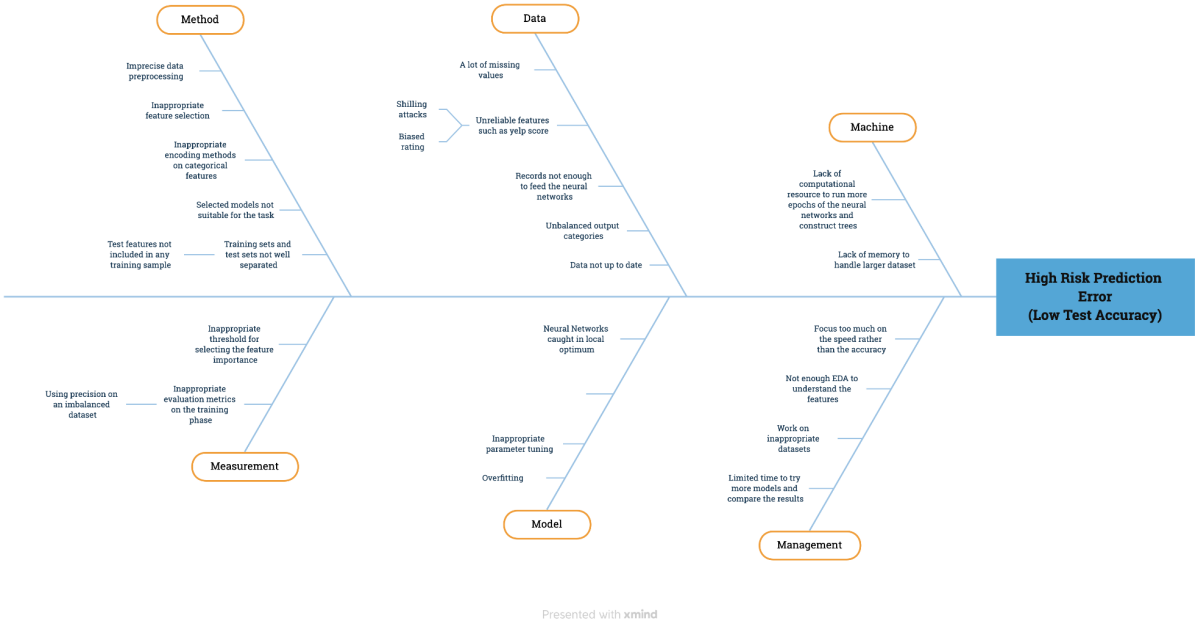



Figure 3.16 : Fishbone Diagram

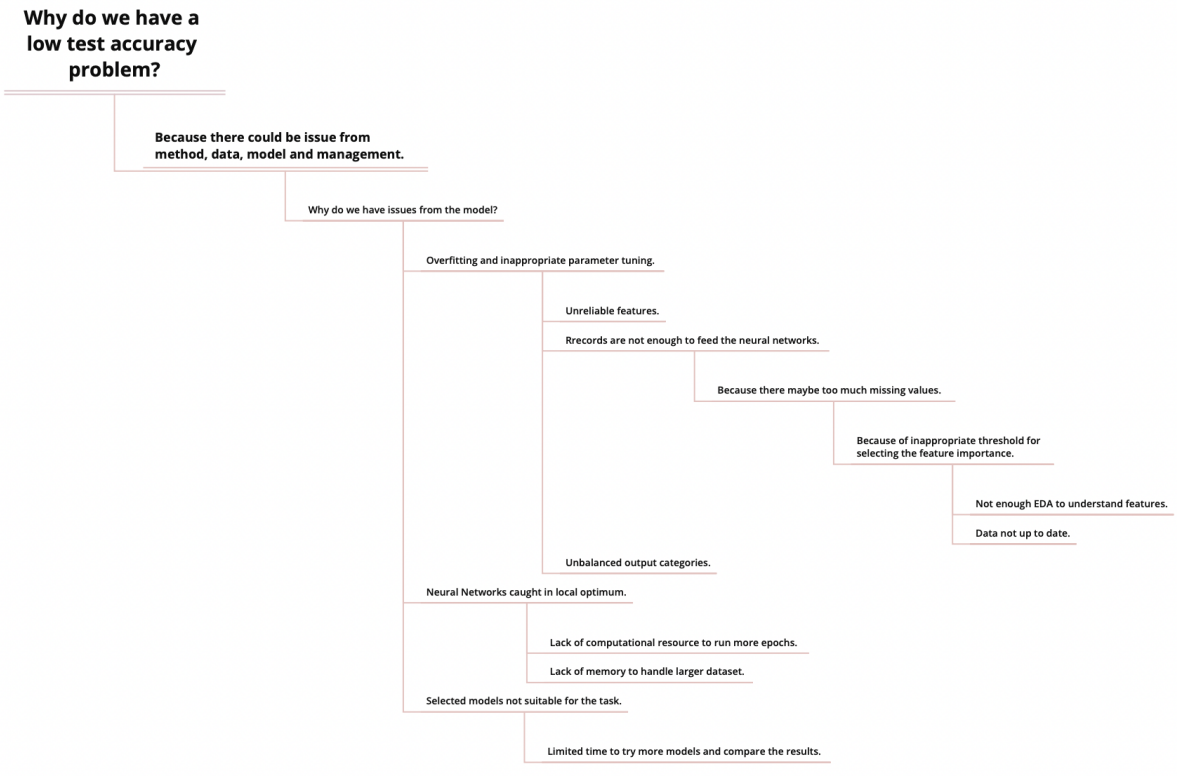



Figure 3.17 : 5 Whys

### 3.3.2 Value Function

In this project, the datasets were obtained from LA Open Portal and Yelp API. We first applied preprocessing techniques and Exploratory Data Analysis to examine the features and potential correlations. In addition, we encoded the categorical variables to facilitate model implementation. After the data was fed to the models, we took time tuning the parameters for the best performances possible. In the end, by integrating the models with a Streamlit web application, our target users, including local government staff and LA citizens, are all able to play with the parameters and obtain the auto-generated results.

### 3.3.3 Sources of Variation

The analysis charts created evidence that the main sources of variation lie in three areas:

1. **Data Collection & Preprocessing**

   Two major problems in this area were the quality of data and the techniques for feature engineering. We realized that the datasets needed to be up-to-date, and the features needed to be relevant to our project goal. By obtaining the data from LA Open Portal and Yelp API, those goals were addressed as the best we could. In addition, feature engineering and EDA were performed in order to pick the best contributing features for our models.

2. **Machine Learning Models**

   For our models, some selected input features were for example restaurant rating, category, risk level, location, and etc, and those features appeared to have no correlation with each other so they were safe to use. We decided to use various models to visualize the relationship between those features of restaurants and their risk level of having health

violations in future inspections, and we wanted to show the results in terms of prediction, segmentation, and recommendation.

a. **Prediction**

   For prediction, we chose to use **Linear Regression** as our baseline model, with alternative models including **Support Vector Machine**, **Random Forest**, **XGBoost**, and **Neural Networks** for better performances in predicting the risk levels.

b. **Segmentation**

   For segmentation, we first performed **Principle Component Analysis**. This step is for dimension reduction as there are many features for a restaurant. Then, **KMeans** was used to learn the representation of the data. In order to visualize the clusters, we used a method called **t-SNE** that could show a 2D visualization of the clusters based on PCA. In the end, since we want to segment the restaurants based on their features, we applied **Topic Modeling** to extract the keywords from comments so that people can see the representative words associated with each cluster.



c. **Recommendation**

   For recommendation, we focused on **content-based recommendation** such that each restaurant will be represented by a set of tags extracted from its features and Jaccard similarities were computed to select top recommendations. Detailed steps are shown below.

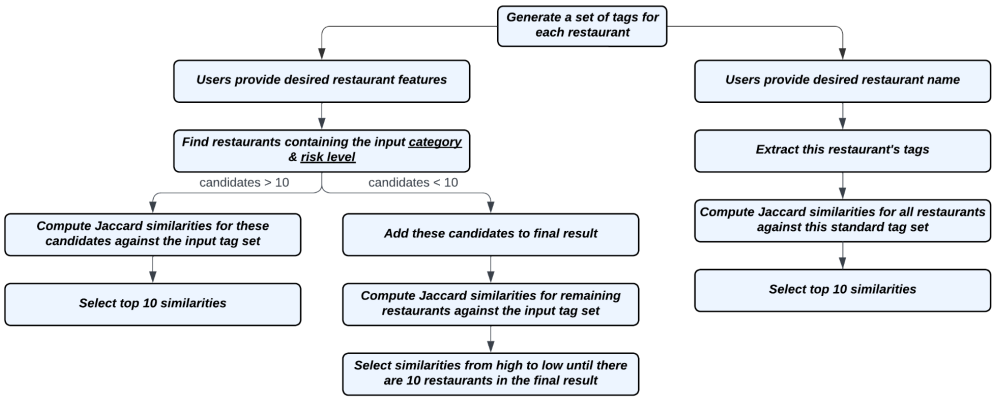



Figure 3.18: Process for Content-based Recommendation

3. **Model Evaluation**

   a. **Prediction**

   **Accuracy** and **ROC_AUC** score were used as the evaluation metrics for prediction models. Our baseline model had 0.835 accuracy and an 0.58 ROC_AUC score. For improvement models, Random Forest had relatively better accuracy and ROC_AUC score. Therefore, Random Forest was chosen as the final model for predicting health risk level for restaurants in the dataset.



| Prediction Model | Accuracy | ROC_AUC |
|---|---|---|
| Logistic Regression | 0.835 | 0.58 |
| SVM | 0.839 | 0.63 |
| Random Forest | 0.842 | 0.67 |
| XGBoost | 0.827 | 0.71 |
| Neural Networks | 0.857 | 0.51 |

Table 3.1: Prediction Model Evaluation

**b. Segmentation**

| Segmentation Model | Evaluation |
|---|---|
| PCA | Explained variance on the first 3 PCs: 0.5 |
| KMeans | Scree Plot: K=5 |
| t-SNE | KL divergence = 0.47 |
| Topic Modeling | We sampled 10 restaurants and checked if the keywords are really included in the descriptions of the restaurants.<br>Included Rate = 70% |

Table 3.2: Segmentation Model Evaluation

**c. Recommendation**



| Recommendation Model | Evaluation |
|---|---|
| Content-based | We ran 10 test recommendations, each time with a random set of input features/restaurant names, to see how many recommended restaurants are actually relevant to the input.<br>Average relevancy = 74% |

Table 3.3: Recommendation Model Evaluation

### 3.3.4 Potential Solutions

Here's a list of potential solutions for each of the major tasks:

1. **Data Collection & Preprocessing**
    - Ensure datasets are reliable, unbiased, up-to-date, and relevant to the central problem.

- Use appropriate preprocessing techniques to remove confusing features, check for repetitive records, find a way to generate a representative value for such records, and tag certain categorical variables to facilitate future analysis.
- Implement Exploratory Data Analysis to visualize data distribution, special trends, outliers, and possible multicollinearity between independent variables.

2. **Machine Learning Models**
    - Handle outliers and abnormal trends in modeling results.
    - Make sure larger input datasets with more numerical and categorical features also allow capturing complex underlying relationships between the features and providing accurate and real-time responses.
    - Incorporate most recent data from the source to continuously tune parameters for better performance.
3. **Model Evaluation**
    - Research on common evaluation metrics used for models selected and understand the meanings of corresponding values.
    - Research on studies with similar central topics and learn from approaches used in those projects.



### 3.3.5 Tools Application

Being able to identify underlying defects that may not be apparent in the earlier stages of a project is crucial, and this is when root cause analysis comes into play. This analysis phase enables us to avoid potential serious problems in later stages. With careful completion of this step, we can smoothly move on to the next step of the project.

## 3.4 Improve Phase

The Improve phase focuses on generating potential solutions, selecting the optimal solution for testing, and developing pilot implementation plans for each team member.

### 3.4.1 Solution Evaluation

We examined the benefits and drawbacks of each of the three essential steps for which alternative solution approaches were established, which are shown below.

1. **Data Collection & Preprocessing**

   The dataset better be .csv or .json files that are more easily to be processed. The data collected should be authentic and not contain any biased views, and it should be up-to-date with features  usable for analysis and modeling. Then, the preprocessing and EDA steps should be clear, meaningful, and reproducible.



| Data Preprocessing & EDA Methods | Pros | Cons |
|---|---|---|
| Remove null values directly | It is simple and can save lots of time. | If there are too many null values/outliers, then removing the whole record can reduce the data size a lot and cause lack of data in the analysis and modeling parts. |
| Replace null values using mean | It is also not complicated and can be done with a few lines of codes. | If a column contains mainly null values, then the average calculated from those cells with values might not be very representative. |
| Merge datasets on certain key(s) | It allows records to be combined and to contain all participating features. It completes the input data | Sometimes the keys refer to the same thing but are in different formats, so it requires extra effort to |

| | | |
|---|---|---|
| | for future steps. | make sure the keys are paired without losing any important records. |
| Encode categorical variables | By converting categorical data to numerical formats, those variables can be used more widely for data analysis, modeling, and data mining. | The process requires extra time and a set of mapping rules for the conversion step. The mapping needs to be explained so that anyone who accesses the dataset can understand. |
| Visualize using histogram | It is easy to generate and can show the trend of the data. | How the distribution looks can easily be affected by outliers and sizes of the bins and thus can mislead the interpretation. |
| Visualize using scatterplot | It can summarize a large numerical dataset, show the trends, and allow comparison between features. | There are too many data points and can cause confusing information. Additional interpretation might be needed. If the features are mainly categorical, then scatter plots cannot work well. |
| Visualize using barchart | It can divide independent variables into subcategories and show the corresponding bars in parallel with different colors. This allows a more direct comparison. | It might require grouping of certain features to identify the independent and dependent variables and the hue, so extra steps and time need to be used. |



Table 3.4: Data Collection & Preprocessing Method Analysis

2. **Machine Learning Models**

   All the models used should be able to predict and recommend fast, to handle the categorical tasks as well as the numerical values, and to define potential relationships between the variables.

| Machine Learning Models | Pros | Cons |
|---|---|---|
| Logistic Regression | LR is simple and straightforward. | Too generalized, might not be able to handle complex relationships. |
| SVM | SVM works well when there is a clear margin of separation between classes and can deal with high dimensional data. | Large datasets and noises can both decrease SVM performance. |
| Random Forest | RF has better generalization performance than an individual decision tree due to randomness. It's also less sensitive to outliers and requires less tuning. It can handle categorical features. | It's fast to train but slow to predict especially when we require a precise prediction with more trees. |
| Gradient Boosting | GB is very accurate and works well with categorical and numerical values. It can also handle missing values. | It's slow to train. GB is expensive to train and not very interpretable. |
| Neural Networks | Neural networks are flexible and good to model with nonlinear data with a large number of inputs. It is reliable in an approach of tasks involving many features. Once trained, the predictions are pretty fast. | NN is data hungry and may lead to overfitting. It is a black box and not very interpretable. It is computationally very expensive. |
| KMeans | KMeans is fast, easy to implement and computationally efficient. | We have to decide k, and the initial center is randomly chosen so it's not very stable. The shapes of clusters can only be circular as it uses distance. |
| Content-based Recommendation | No cold start problem. Any new item can be | We have to select the features. Neighboring |

| | | |
|---|---|---|
| | recommended. It doesn't need users' information and is interpretable. It caters to the users' preference. | information cannot be exploited and we can hardly recommend items outside the users' content profile. |

Table 3.5: Machine Learning Model Analysis

3. **Model Evaluation**

Evaluation metric should be interpretable and better be easily computable. In addition, it should be able to apply to datasets with various distribution patterns.

| **Evaluation metric** | **Pros** | **Cons** |
|---|---|---|
| Evaluation: F1-Score (classification & recommendation) | F1 score can take imbalance dataset into consideration. | It's not very interpretable. |
| Evaluation: ROC AUC (classification) | ROC curve is very intuitive. It shows a tradeoff between sensitivity and specificity for all possible thresholds rather than just the one that was chosen by the modeling technique. | Not good for an imbalanced dataset. |
| Evaluation: Silhouette Score to select K (clustering) | It considers how close each point in a cluster is to points in the neighboring clusters rather than only within group distance. It can help select k. | The computation is huge. |

Table 3.6: Evaluation Metric Analysis



### 3.4.2 Recommended Solution

The team assignments of major steps are shown as the following:

1. **Data Collection & Preprocessing**

| Action Items | Team Member | Delivery Date |
|---|---|---|
| Data Collection from LA Open Portal | All | 02/15/2023 |
| Data Collection from Yelp API | Xiaoyi Gu | 02/20/2023 |
| Record linkage | Xiaoyi Gu | 02/24/2023 |
| Finalize data & preprocessing | Zhenmin Hua | 02/26/2023 |
| Exploratory Data Analysis | Ziyue Chen, Ying Wang | 02/26/2023 |

Table 3.7: Data Collection & Preprocessing Assignment

2. **Machine Learning Models**

| Action Items | Team Member | Delivery Date |
|---|---|---|
| Prediction - Logistic Regression & SVM | Zhenmin Hua | 04/03/2023 |
| Segmentation - PCA, Kmeans, t-SNE, Topic Modeling | Xiaoyi Gu | 04/03/2023 |
| Recommendation - generate tag sets | Ying Wang | 04/06/2023 |
| Recommendation - compute Jaccard | Ziyue Chen | 04/06/2023 |



Table 3.8: Machine Learning Model Assignment

3. **Model Evaluation**

| Action Items | Team Member | Delivery Date |
|---|---|---|
| Prediction - tune parameters | Zhenmin Hua | 04/06/2023 |
| Segmentation - tune | Xiaoyi Gu | 04/06/2023 |

| | | |
|---|---|---|
| parameters | | |
| Recommendation - run tests, improve algorithm | Ziyue Chen, Ying Wang | 04/09/2023 |

Table 3.9: Model Evaluation Assignment

4. **Web Application**

| Action Items | Team Member | Delivery Date |
|---|---|---|
| Prediction page | Zhenmin Hua | 04/12/2023 |
| Segmentation page | Xiaoyi Gu | 04/12/2023 |
| Recommendation page | Ziyue Chen, Ying Wang | 04/12/2023 |

Table 3.10: Web Application Assignment

### 3.4.3 Pilot Design



In the Improve phase of the project, we designed a pilot implementation which included tasks required for data collection & preprocessing, model selection, evaluation metric analysis, and UI design. All tasks were proposed based on the customer needs so that we could work on the right track.

For data collection & preprocessing, we set some criteria to select the appropriate dataset and features, and EDA was applied to provide a general picture of the data. Then during model selection, we assessed the advantages and disadvantages of various common machine learning models and finalized our selections to the prediction, segmentation, and recommendation models mentioned above. In order to see the performance of those models, we also identified some valid evaluation metrics and applied them to the modeling results. Test recommendations were run as well to indicate the relevancy of recommendation results. Finally, we designed the dashboard for our web application, allowing users to interact with models and see real-time results.

### 3.4.4 Work Breakdown Structure

The tasks involved in improving the project's deliverable can be divided into six main sections:

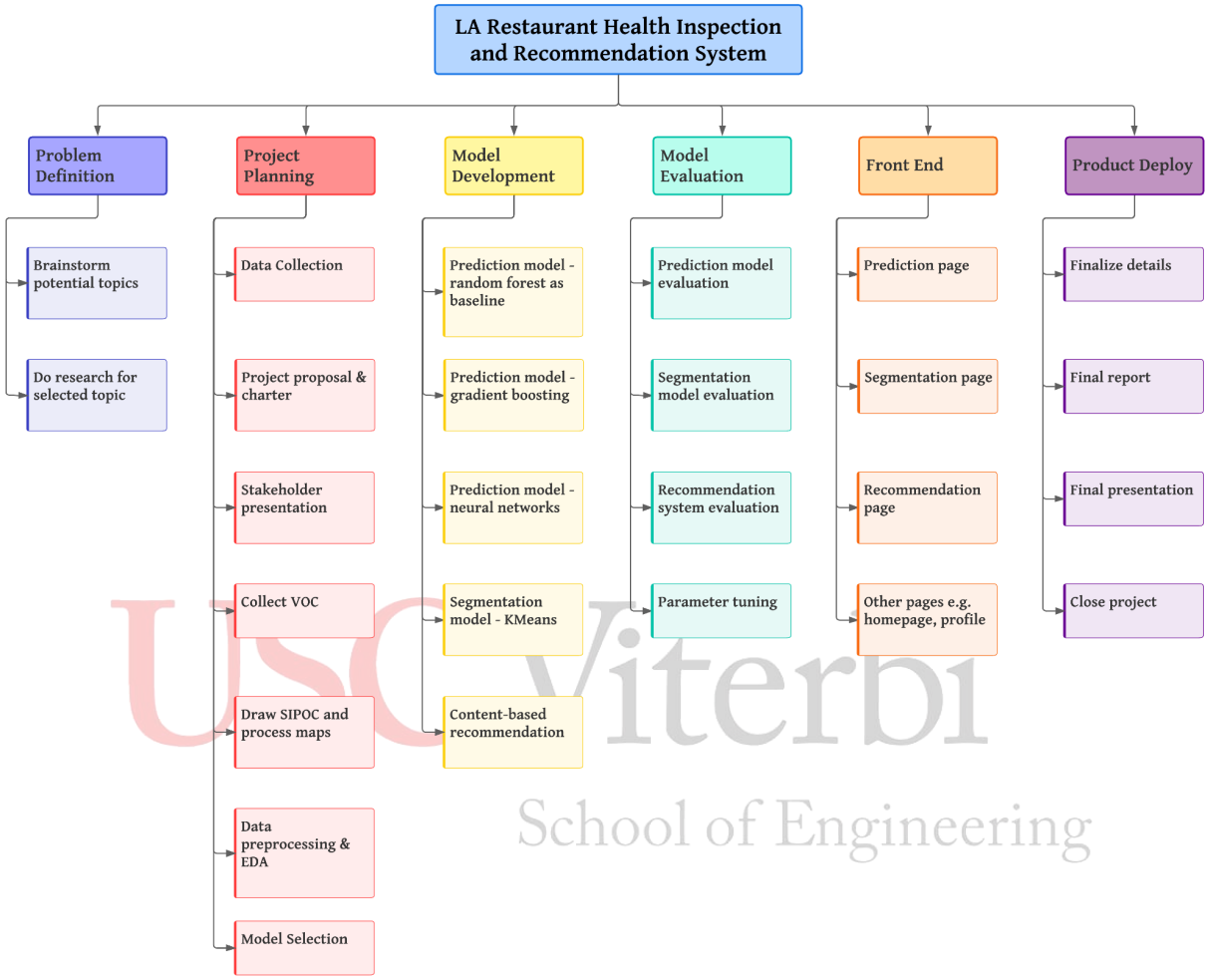



Figure 3.19: Work Breakdown Structure

1. **Problem definition**

   A brainstorm session was held to come up with interesting and innovative ideas first. Then, the single topic was selected, and everyone was assigned to do some research.

2. **Project planning**

   A SIPOC diagram was first created to guide the planning step. Data and the Voice of Customer were collected, and a project proposal was drafted and presented to

stakeholders for feedback. Once the topic was approved, data preprocessing, EDA, and model selection were implemented to better prepare the data for the following modeling development step.

3. **Model development**

   Each team member was assigned a model to work on, and weekly meetings were scheduled to report any progress made and to discuss potential adjustments.

4. **Model evaluation**

   Based on the performance of the models, parameters and algorithms were tuned, and models were re-evaluated until the evaluated metrics provided satisfying values.

5. **Front End**

   Once the models were finalized, several Streamlit pages were created to visualize the results and to provide an interactive user interface.

6. **Product Deploy**

   In the deployment section, details were finalized, and a final presentation was given. The project would close successfully.

## 3.5 Control Phase

The control phase ensures that defects don't recur and the process is being managed and monitored properly, which means the major task is to identify the strategies to keep the performance of the process and maintain the improved state. The control phase helps to make sure that the targets are met. Steps in the control phase include: creating the monitoring plan, implementing the full-scale solution and finalizing the transition. In the control phase, the project's success is constantly monitored and evaluated so that defects will be detected.

### 3.5.1 Control Solutions Considered

In the project, we established the following guardrail methods so that the process would be under control.

1. **Monitoring Plan.** In the beginning, the team developed a monitoring plan to check the performance of the updated process.
2. **Collaborative Platform and Logs.** The team used the collaborative platform to keep track of data and codes. Meanwhile, we had error logging, error reporting, and fallback mechanisms. We recorded challenges and tips in the process so that the previous mistakes could be avoided.
3. **Package Versioning.** For the packages used in the scripts, a version control system was maintained to ensure that the project is robust. When the problem occurs, the version control system can provide a rollback option.
4. **Poka Yoke Control Strategies.** For the user end, we eliminated the possible wrong input such as non-existing restaurant names to prevent any faults from being introduced.
5. **Sync Data System.** We guaranteed the consistency of the data both in the front-end and back-end by using a Streamlit Cloud and GitHub repository.

### 3.5.2 Control Solution Implemented

1. **Monitoring Plan.** We developed a plan defining the frequency of reviewing the process, checking the performance of the models, and reporting the updates. When incurring issues, we looked into the problems.
2. **Collaborative Platform and Logs.** We published the application code on GitHub and collaborated on the data preprocessing and EDA on Google Colab. We saved all the documents and meeting notes in a shared Google Drive folder. So it's convenient for us to record the issues and keep everything up-to-date.
3. **Package Versioning.** As we publish the application code on GitHub, it's easy to control the package versions.
4. **Poka Yoke Control Strategies.** For the clustering part of the application, we allowed the users to select the existing features of the restaurants so that when plotting the images, errors won't occur. And for the recommendation part, we used a select box for the user to choose the restaurants and features in the database, which eliminates the incorrect entries into the platform.
5. **Sync Data System.** We use the Streamlit Cloud to hold the application. The application is connected with GitHub and once the data is updated on GitHub, the application will automatically be updated as well.
6. **Documentation.** We maintained and updated documentation so that all the team members will understand the process. The documentation includes project charters, process maps, client needs, project charts, graphs, and technical reports.

## 3.6 Result and System Implementation

### 3.6.1 Machine Learning Approaches

1. **Record Linkage**

We have applied the entity resolution technique we have learned in Knowledge Graph(DSCI558) to combine the LA Open Dataset with the Yelp dataset. As there are 2 datasets and we want to match them, this is a record linkage problem. We firstly generated a unique id for each record in the LA Open Dataset and each restaurant in the Yelp dataset. Altogether, there are over 13303 records and over 9,000 restaurants, which means we may have to check over 13303 * 9000 pairs to see whether they can match. If we use the naive way to compare the pairs, it will be too inefficient. So we apply the Blocking method here. That is, we first check the first 5 digits in the zip code and the first 5 characters of the address of each dataset. We only generate pairs when the first 5 digits/characters of the zip code and address are matched. By blocking, we reduced the pairs to 10981, which is only 0.012% of the original pairs. Meanwhile, we manually coded 49 pairs of ground truth, that's the true matched pairs. We calculated the completeness(=how many true pairs we find / number of total true pairs) of the blocking method, and got the score=1.0. Therefore, all the ground truth pairs are found in blocking.



| FACILITY_ID | FACILITY_NAME | FACILITY_ADDRESS | FACILITY_ZIP | name | address |
|---|---|---|---|---|---|
| FA0242516 | NERANO | 9960 S SANTA MONICA BLVD | 90212 | Nerano | 9960 S Santa Monica Blvd Beverly Hills, CA 90212 |
| FA0177063 | PASTA SISTERS | 3343 W PICO BLVD | 90019-4530 | Pasta Sisters | 3343 W Pico Blvd Los Angeles, CA 90019 |
| FA0008028 | BAMBO GARDEN | 8844 S VERMONT AVE | 90044 | Bamboo Garden Restaurant | 8844 S Vermont Ave Los Angeles, CA 90044 |
| FA0295467 | BARI | 8422 W 3RD ST | 90048 | Bari | 8422 W 3rd St Los Angeles, CA 90048 |
| FA0160554 | TASTE OF TEHRAN | 1915 WESTWOOD BLVD | 90025 | Taste of Tehran | 1915 Westwood Blvd Los Angeles, CA 90025 |

Figure 3.20: Sample Result

As you can see, the facility names of the same restaurant(e.g. In the LA Open Dataset: BAMBO GARDEN & In the Yelp dataset: Bamboo Garden Restaurant) can be slightly different in the two datasets. Hence, we define the entity linking similarity metrics to apply fuzzy matching. We mainly use the jaro-winkler similarity because it considers the order and performs well for some prefixes and thus works well for our task. We set a threshold and any pairs with similarity over the threshold will be considered a match. We apply the similarity metrics to the address, restaurant name and zip code, and filter the candidate pairs with the threshold. Finally, we got 100% precision, 97.95% recall and 98.97% F1 score on the ground truth. We will then use the matched results for our models.

| FACILITY_NAME | FACILITY_ADDRESS | FACILITY_CITY | FACILITY_ZIP | restaurant_id | name |
|---|---|---|---|---|---|
| #1 CAFE | 2080 CENTURY PARK E STE 108 | LOS ANGELES | 90067 | 753 | One Cafe |
| #2 MOON BBQ | 478 N WESTERN AVE | LOS ANGELES | 90004 | 2526 | Moon BBQ 2 |
| 101 ASIAN KITCHEN INC | 7170 BEVERLY BLVD | LOS ANGELES | 90036 | 1080 | 101 Asian Kitchen |



Figure 3.21: Sample Result 2

2. **Prediction Model**

The input features are numerical values such as Yelp rating scores. We also have many categorical features such as location, restaurant category and price(which means not expensive, medium and expensive). Our output will be categorical, the risk level. So we plan to predict low risk, medium risk and high risk. As we predict different levels, it is actually a multiclass classification problem. Logistic Regression is a basic model for multiclass classification. We use it as our baseline. We also use the random forest, SVM, gradient boosting and neural networks. RF and GB both work well for categorical

features. Neural networks can learn complex relationships. We compare these models and select the one with the best results such as the F1 score.

In this task, health scores in the training set are initially given within a range from 0 to 100. By cutting the range into 3 bins, we generate an extra column named “Risk level” to represent the health risk, which has 3 values: high risk, medium risk, and low risk. Next, we select reasonable features to be included in our model, which include zip_code, review_counts, price_level, category, open_duration, and size. Then, we choose multiple machine learning models to classify the restaurants and evaluate the results using accuracy, f1_score and roc_auc.

3. **Segmentation Model**

   We would like to investigate which kind of restaurants tend to violate. This is an unsupervised learning problem and we want to learn the representation of the data. We first run a PCA to reduce the dimension of data. Then we apply KMeans to cluster the features. KMeans can be used to get the cluster centers and based on the groups we can try to discover the key features of each group. To visualize the clusters in a 2-dimension plot, we use t-SNE(t-distributed stochastic neighbor embedding). After that, we apply the LDA topic modeling technique to detect keywords in the restaurant comments in each cluster to get insights.

4. **Recommendation**

   We first select the features to represent the user’s taste and the restaurants. We can calculate the vectors of the restaurants offline based on selected features and each time a user has some preferences, we build the user vector. We calculate the distance between the user and each restaurant and recommend restaurants based on preference.

Content-based recommendation caters to the user's taste. In our dataset, each restaurant is associated with various features, such as location, score, size, risk level, yelp rating, yelp review count, price, and category. We try to use an item content-based recommendation system by building an item profile for all the restaurants, each with a set of tags extracted from the key features mentioned above. Then, whenever a user searches for a restaurant, our recommendation system will search for the top 10 restaurants that have the most similar sets of tags by calculating the cosine distance between restaurants.

### 3.6.2 System Implementation

The LA Restaurant Portal application is designed to provide health risk predictions and restaurant clustering information to the government so that the government can better allocate its inspection resources. Meanwhile, it offers restaurant recommendations to the citizens based on their preferences and helps the citizens to gain a satisfying and healthy dining experience. Therefore, the project contains 3 parts.



1. First of all, it has a prediction model. We selected Random Forest here because it has the best performance. As the user uploads a file containing the restaurant information, the model will automatically predict the health risk level for each restaurant and visualize the result(low risk, medium risk, high risk).
2. Secondly, the system has a clustering algorithm to segment the restaurants into several groups and show the statistics of each cluster such as how many low-performance restaurants are there in cluster 1. Based on the clustering algorithm results, we use t-SNE to plot the distribution of the restaurants. Then the LDA topic model is applied to extract the keywords of the restaurant comments in each cluster. For example, if there are many restaurants with low health scores in cluster 1 and 2, and the government checks

keywords and finds that the word “chicken” is in both cluster 1 and 2 and not other clusters, then the result suggests that the government should pay more attention to the restaurants selling chicken.

3. Thirdly, the system has a recommendation. The recommendation function allows the users to choose the restaurant's features such as categories and price and generates the top 10 restaurants that satisfy the user’s choice. Meanwhile, if the user already has some favorite restaurants, the application will also recommend the 10 most similar restaurants to the user based on the favorite one.
4. All the functions are integrated into the web application so that the users can visit a single website for all the services. We utilized Streamlit and Python to develop the front-end UI, the back-end algorithms, and the managed dataset.



### 3.6.3 Prototype and Demonstration

The system functionalities on every page are detailed below:

1. **Index page.** The user can visit the source code of the application and the raw data from the LA open dataset and review the information of the project.
2. **Prediction page.** The government has a field to upload files or input the features of restaurants such as categories, location, prices and so on. The application will run the prediction model and return the visualized name and predicted risk of the restaurants.

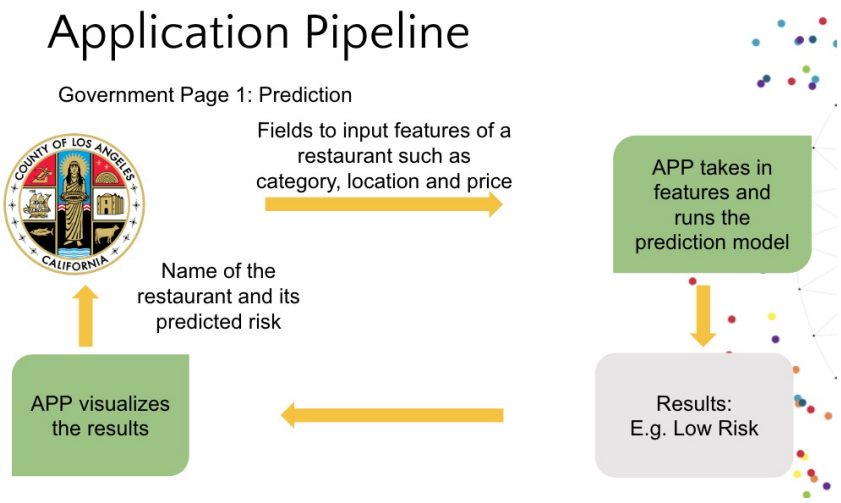



Figure 3.22: Application Pipeline - Prediction Page

3. **Segmentation page.** The user can select different features of the restaurant and check the clustering performance of the features. The user can also select 2-dimensional or 3-dimensional visualization. Meanwhile, the user is able to see the keywords detected in each cluster and the score distribution in each cluster.



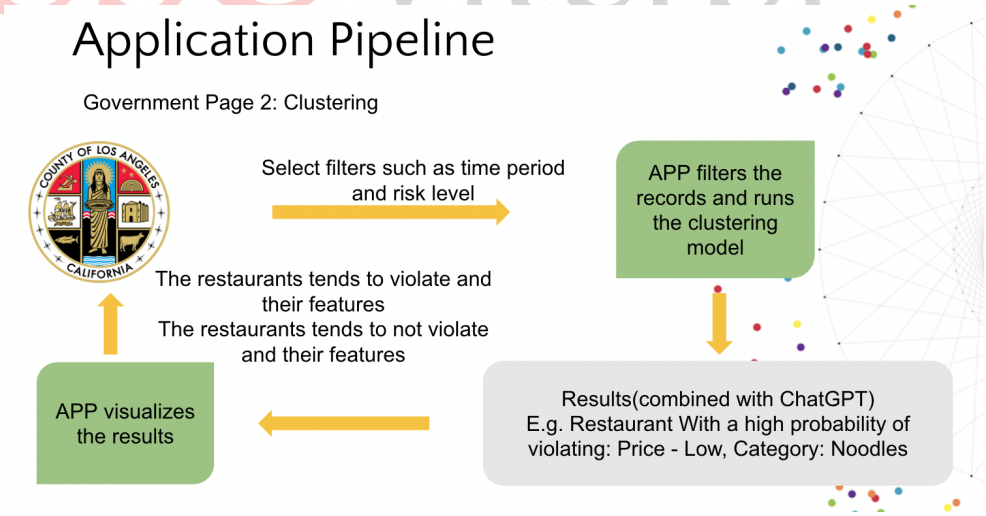



Figure 3.23: Application Pipeline - Segmentation Page

4. **Recommendation page.** The user can either select features of restaurants or input the restaurant they like and get the top 10 recommendations.

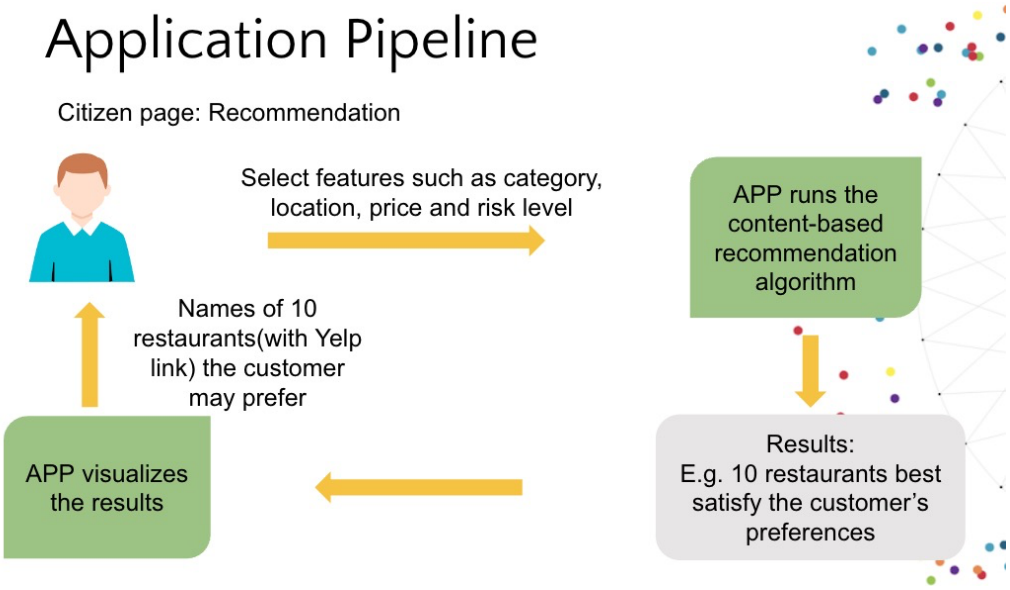



Figure 3.24: Application Pipeline - Recommendation Page

# References


1. N. H. M. Shamsuddin, N. A. Ali and R. Alwee, "An overview on crime prediction methods," 2017 6th ICT International Student Project Conference (ICT-ISPC), Johor, Malaysia, 2017, pp. 1-5, doi: 10.1109/ICT-ISPC.2017.8075335.
2. M. Feng et al., "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," in IEEE Access, vol. 7, pp. 106111-106123, 2019, doi: 10.1109/ACCESS.2019.2930410.
3. Iqbal, Rizwan, et al. "An experimental study of classification algorithms for crime prediction." Indian Journal of Science and Technology 6.3 (2013): 4219-4225.
4. Raschka, Sebastian. "Model evaluation, model selection, and algorithm selection in machine learning." arXiv preprint arXiv:1811.12808 (2018).
5. Llorente, Fernando, et al. "Marginal likelihood computation for model selection and hypothesis testing: an extensive review." SIAM Review 65.1 (2023): 3-58.
6. Barron, Andrew R. "Predicted squared error: a criterion for automatic model selection." Self-organizing methods in modeling. CRC Press, 2020. 87-103.
7. Choi, Jinkyung, Douglas Nelson, and Barbara Almanza. "Food safety risk for restaurant management: Use of restaurant health inspection report to predict consumers' behavioral intention." Journal of Risk Research 22.11 (2019): 1443-1457.
8. Choi, Jinkyung, Douglas Nelson, and Barbara Almanza. "Food safety risk for restaurant management: Use of restaurant health inspection report to predict consumers' behavioral intention." Journal of Risk Research 22.11 (2019): 1443-1457.
9. Siering, Michael. "Leveraging online review platforms to support public policy: Predicting restaurant health violations based on online reviews." Decision Support

Systems 143 (2021): 113474.

10. Luna, Julio César, et al. "Food safety assessment and risk for toxoplasmosis in school restaurants in Armenia, Colombia." Parasitology research 118 (2019): 3449-3457.

# Appendix

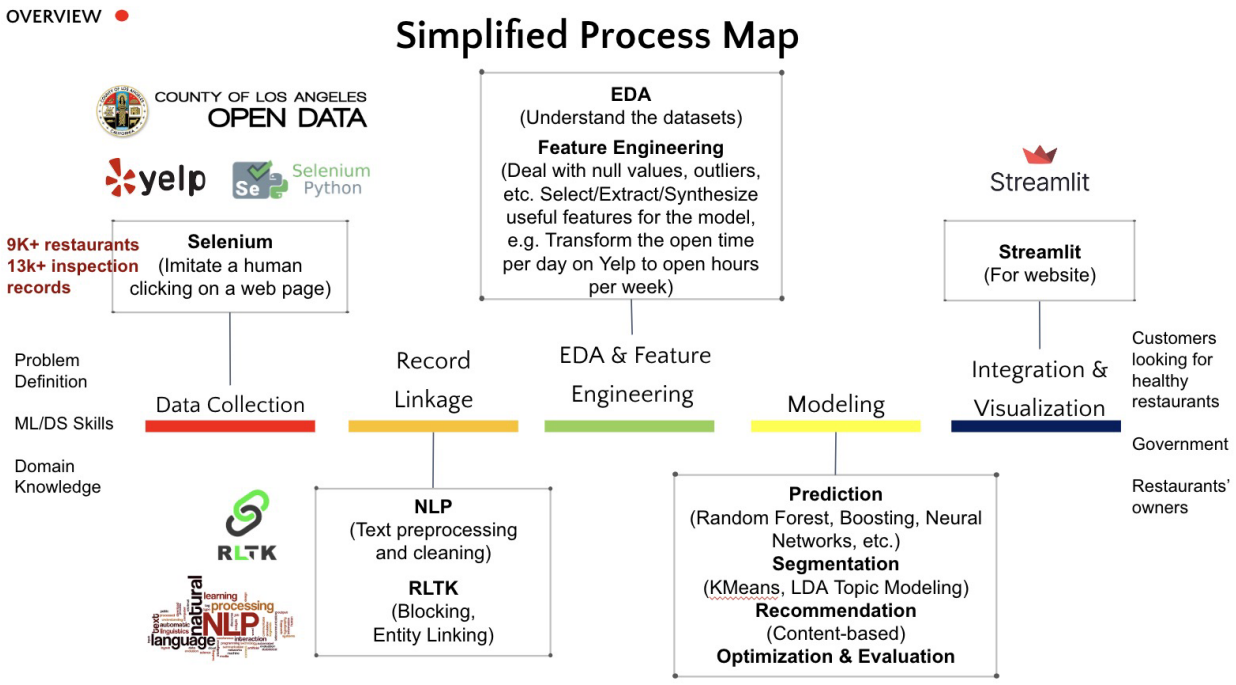



Figure A.1: Project Pipeline



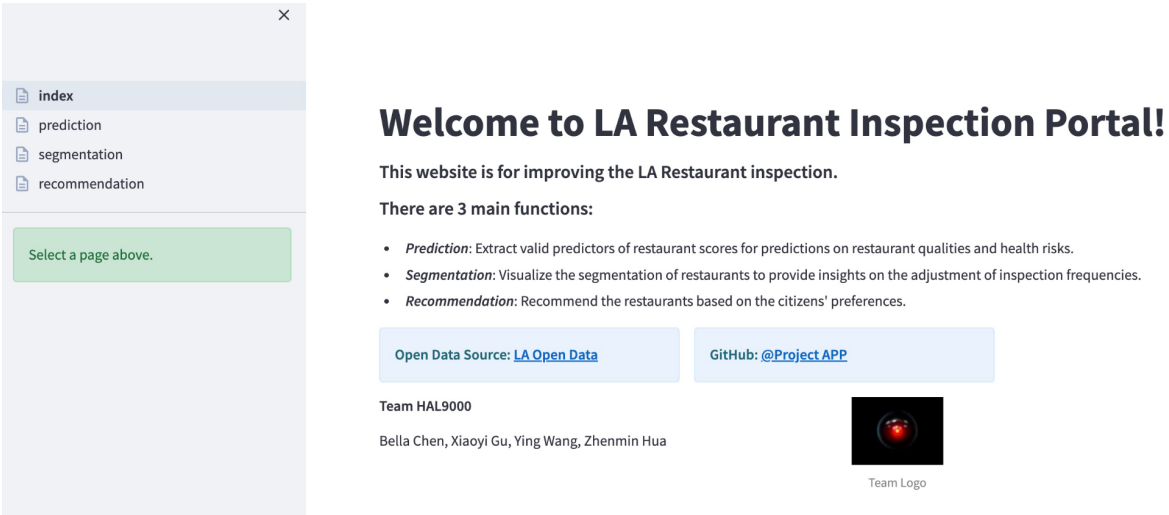



Figure A.2: Index Page

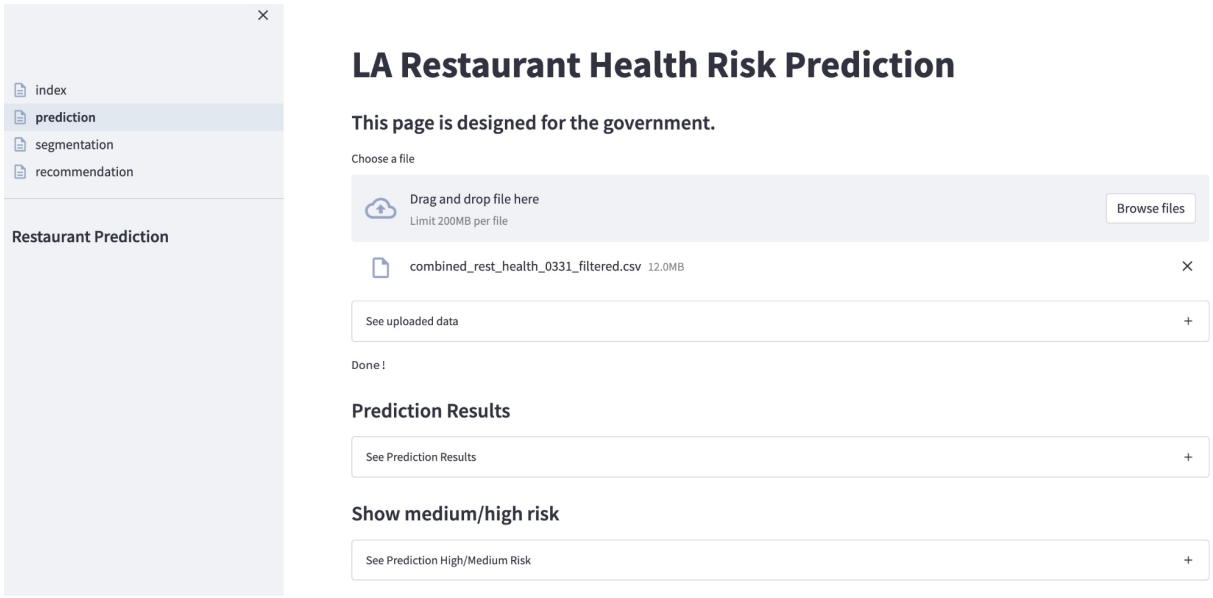



Figure A.3: Prediction Page



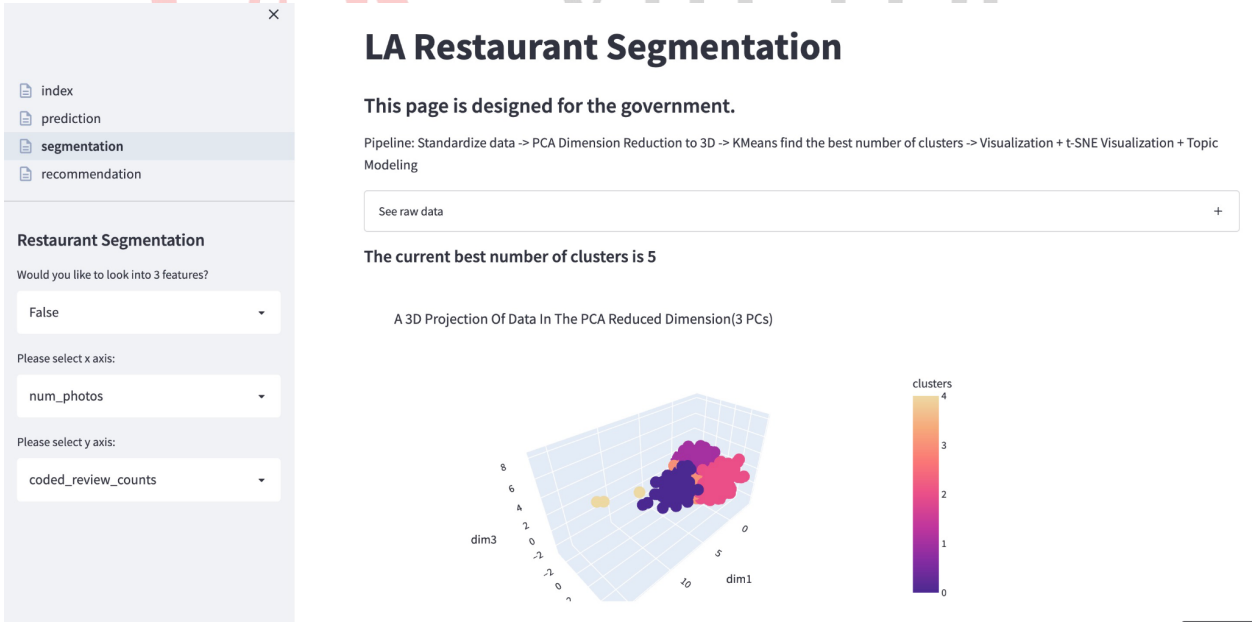



Figure A.4: Segmentation Page-1


Figure A.5: Segmentation Page-2



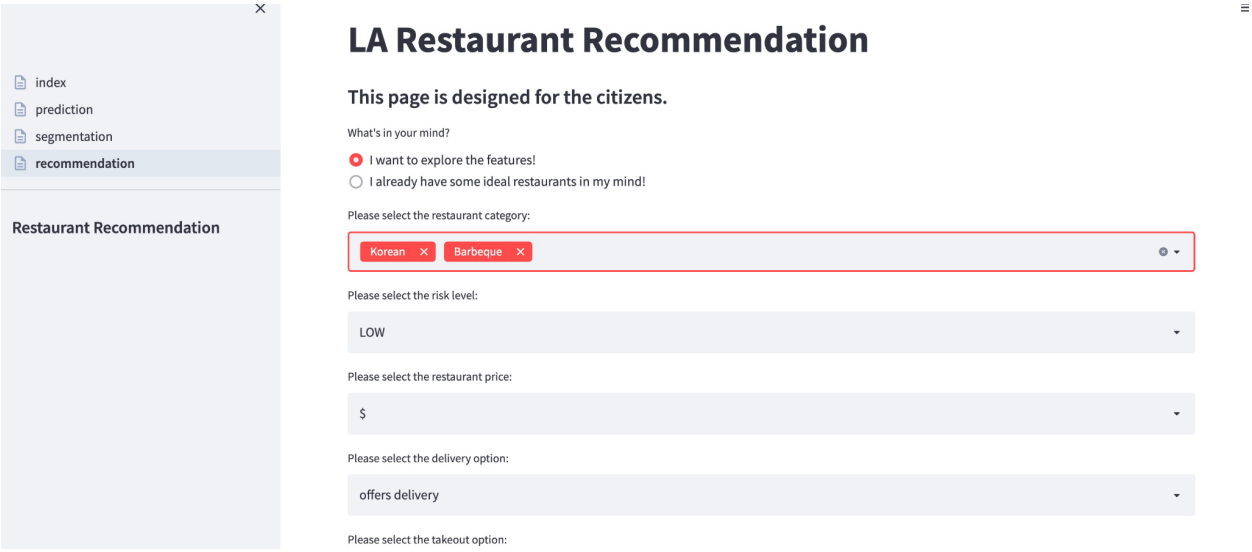



Figure A.6: Recommendation Page-1

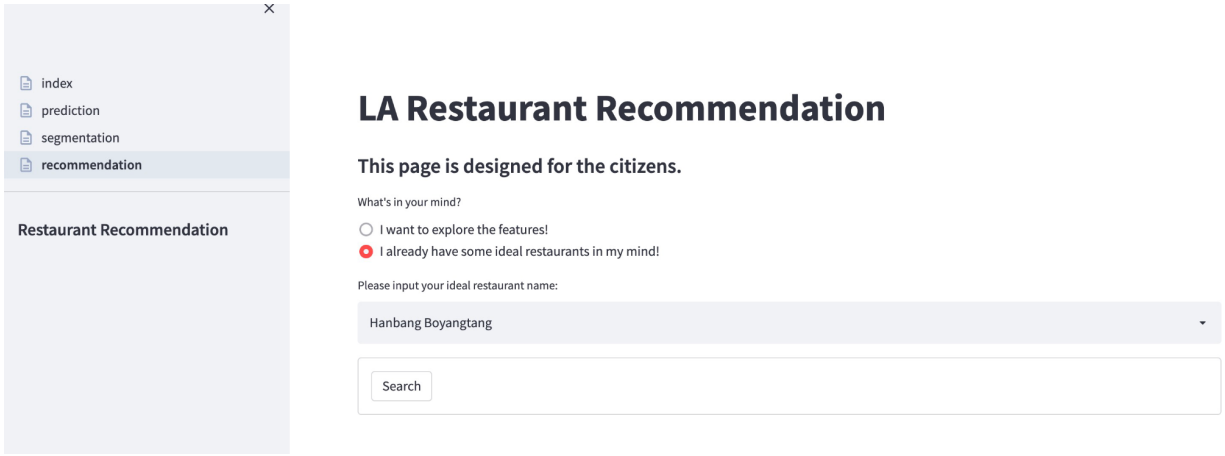



Figure A.7: Recommendation Page-2

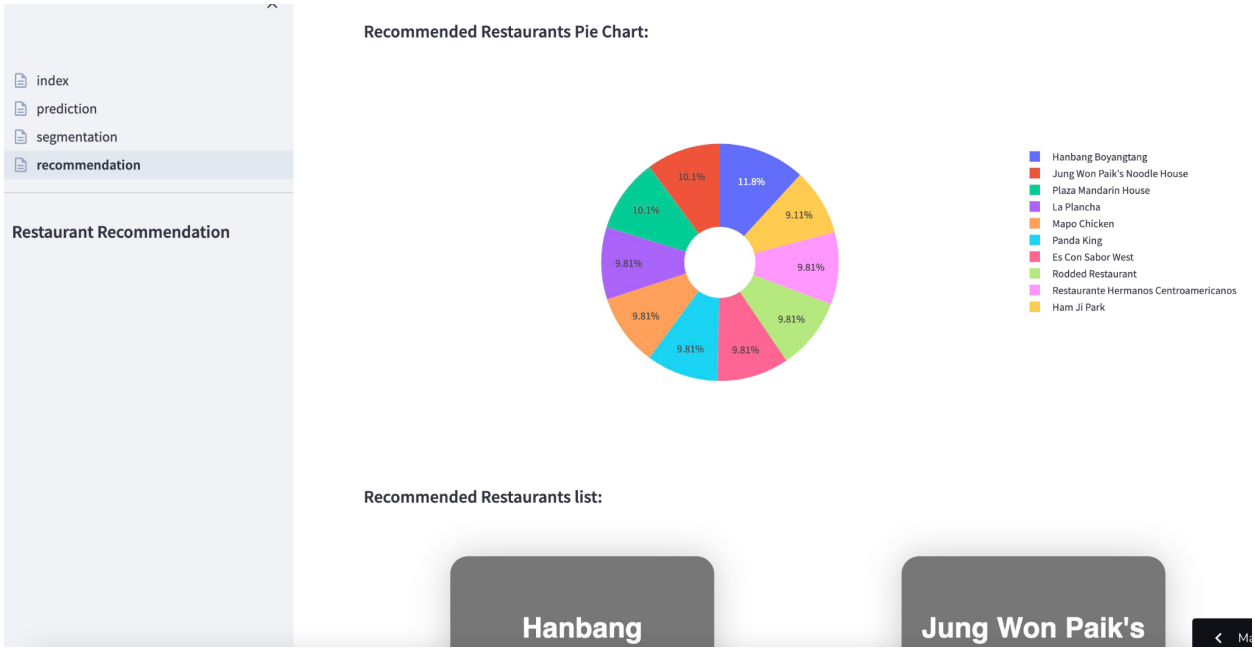



Figure A.8: Recommendation Page-Results