**Final Project Report**

CST8390-Business Intelligence and Data Analytics

**Team Members:**

Om Dankhra (040998226) Section-14

Jeel Asalalia (041002504) Section-14

**Professor:**

Subair Abayomi

**Objective:**

The goal of this final project is to think about how to apply the algorithms we learned in this course to a real dataset.

Table of Contents

[Data Collection: 3](#_Toc110347981)

[Introductory Question: 3](#_Toc110347982)

[Pre-Processing (Data preparation): 4](#_Toc110347983)

[Dataset after Data cleaning. 5](#_Toc110347984)

[Data Analysis 5](#_Toc110347985)

[Complaint Type Analysis 5](#_Toc110347986)

[Status Analysis 6](#_Toc110347987)

[County Analysis 7](#_Toc110347988)

[Data Analysis using KNN Algorithms 7](#_Toc110347989)

[Confusion matrix of KNN “1” 9](#_Toc110347990)

[Results 9](#_Toc110347991)

[Conclusion 9](#_Toc110347992)

# Data Collection:

Reference for the Dataset: <https://opendata.maryland.gov/Energy-and-Environment/Maryland-Department-of-the-Environment-MDE-Air-and/crti-ybyp>

**Maryland Department of the Environment (MDE) - Air and Radiation Administration (ARA) Complaints**

* There are 10 attributes and 625 instances in this dataset. The Maryland Department of the Environment (MDE) provides this data. This dataset gives a comprehensive overview of air quality and other radiological health complaints. We will be able to track the complaints made by people living in different areas of Maryland, as well as the status of the complaints with the received date, complaint type, zip code, complaint date and description, and complaint closed dates, using this information.
* While observing this dataset we can see description of complains and what the complaint is about. This dataset is very straight forward and right to the point so that even a person with minimum knowledge can understand it.
* Observation of different datasets in the past will help us to track this dataset very efficiently and accurately. Initially when we were observing the dataset, we thought that probability air complaints would be more than any other complaints but later on we came to know that odor complaints are highest which is quite interesting fact about this dataset.

## Introductory Question:

This is the question which I would like to discuss with the help of this analysis.

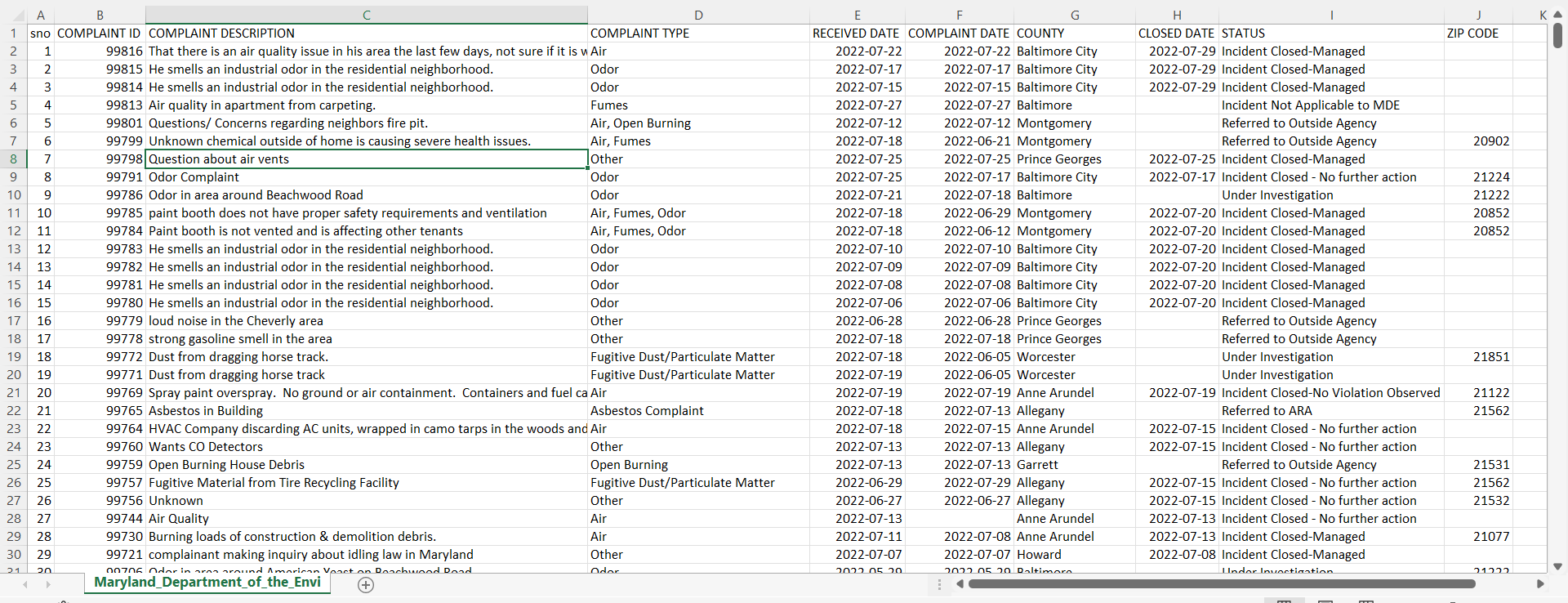
* Out of all the categories, why odor complaints were reported more and in which area of the state is having highest complaints?

# Pre-Processing (Data preparation):

**Meaning:** The process of converting raw data into a comprehensible format is known as data preparation. We can't deal with raw data; thus, this is a key stage in data mining. Before using machine learning or data mining methods, make sure the data is of good quality.

**Following are the steps which we have followed for Data cleaning.**

* With the use of Microsoft Excel software, the attributes that were not beneficial in the observation and analysis for the final conclusions were eliminated from the dataset, and some were transformed to nominal attributes with ranges for better comprehension.
* There were 10 attributes at the initial stage of this dataset later after the data cleaning After cleaning the dataset we got 6 attributes only which can be helpful for our data observation.
* This was the data set which we got initially.



* We deleted four attributes: "SNO," "COMPLAINT\_ID," "COMPLAINT\_DATE," and "ZIP-CODE," as well as removing duplicates of the data. We made those adjustments after sorting out our dataset and noticing that a few properties will not be required to identify our main aim.
* Going forward, we renamed “received date” to “complaint received date”, because earlier we deleted the complaint date attribute, which displayed the date the complaint was made on, and received date was original date the complaint was received, so we merged the two.
* We later removed Zip code because it was missing in more than 70% of the data.

# Dataset after Data cleaning.

|  |  |
| --- | --- |
| Attribute Name | Description of Attribute |
| Complaint Description | Contains detailed description on the issue. |
| Complaint Type | Breakdown of complaint types. |
| Complaint Received Date | The date complaint was received. |
| County | Region used for administrative or other purposes |
| Closed Date | Date on which the complaint was closed due to completion (if has any). |
| Status | Status of the complaint raised. |

# Data Analysis

From here, we'd want to begin data analysis for our final dataset; this analysis will aid us in resolving our unanswered issues as well as uncovering some key facts about the dataset.

## Complaint Type Analysis

* First, we will do the complaint type differentiation from all the complaints done.
* We differentiated the complaint types, who were allocated into the following nominal categories: Odor, Air, Other, Fugitive Dust/Particulate Matter, Open Burning, Smoke, Fumes, etc.

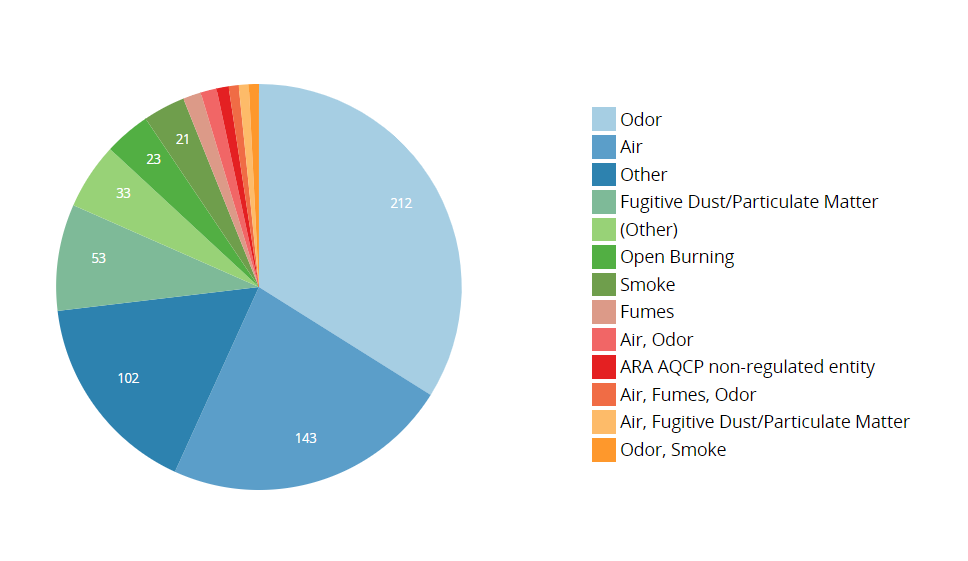


Figure 1 Pie Chart showing Complaint Types

* As you can see, the overall number of complaints was 625 with 34% of complains were about Odor (212), 23% of Air (143), 16% of Other (102) complaints as mentioned by people while reporting the complaint that means complaint type was mentioned while reporting it and all other complaint types like Open burning, Fumes, Fugitive dust, smoke and ARA AQCP non-regulated entity was 27% (168).

## Status Analysis

* We shall now proceed to examine the last known status of the complaint. As we can see, this quality has 10 different types: Incident Closed-Managed, Incident Closed - No further action, Referred to Outside Agency, Incident Closed-No Violation Observed, Incident Not Applicable to MDE, Under Investigation, Referred to LMA, referred to WMA, referred to ARA, no value (Unknown/ not completed).

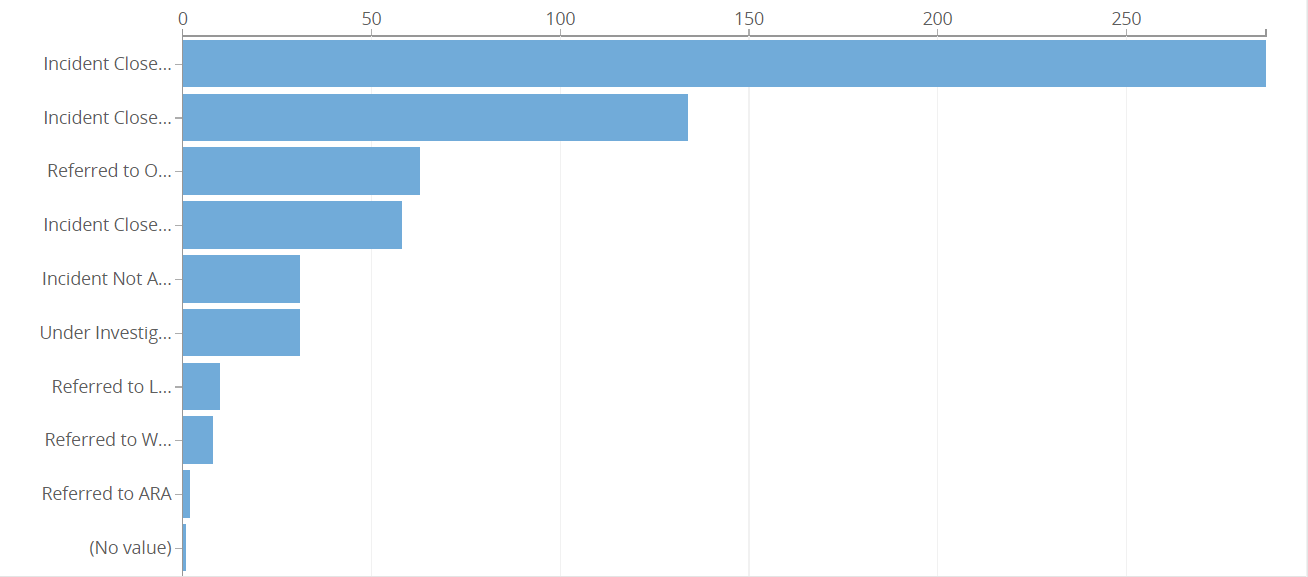
****

Figure 2 Bar Chart showing last known status of the complaint

* We can clearly see that 86% of the complaints were closed, but were further divided into different categories like Incident Closed-Managed, Incident Closed - No further action, and Incident Closed-No Violation Observed.
* Around 5% (31 complaints) were not applicable to MDE. Along with 10% of the complaints were referred to outside agency and 3% complaints were referred to LMA, WMA and ARA.
* 5% of the complaints were under investigation and 1 complain with no status found or unfinished.

## County Analysis

* Here, we can see number of complaints from different county.

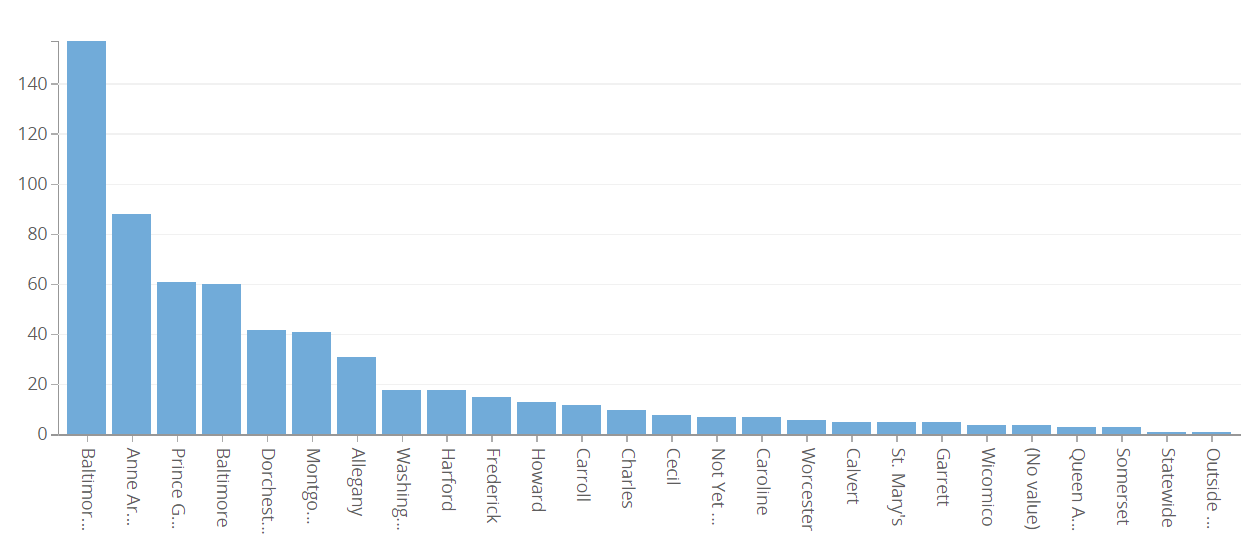


Figure 3 Column chart showing number of complaints from different county.

* The highest number of complaints raised are from Baltimore City 35% (217) that is total of one third of the data. People of Anne Arundel also had issues and raised 88 complaints that’s 14%. So, these two cities had nearly 50% of the air and radiation related issues. By this data we can assume that the amount of air pollution in these two cities of Maryland is highest.
* Carroll and Fredrick have only 2% of total complaints each and would be great in keeping the cities are clean.

## Data Analysis using KNN Algorithms

* KNN algorithms were used on this data set to determine Is the complaints were solved or not and to check the accuracy of our dataset.
* KNN algorithms was run on the different values of K (1, 3, 5, 7, 9, 11, 13, 15).
* As an outcome, the accuracy obtained for the K value of 1 is 61.4498 percent.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Status | TP Rate (KNN 1) | TP Rate (KNN 3) | TP Rate (KNN 5) | TP Rate (KNN 7) |
| Incident Closed-Managed | 0840 | 0.804 | 0.797 | 0.811 |
| Incident Not Applicable to MDE | 0.129 | 0.129 | 0.000 | 0.000 |
| Referred to Outside Agency | 0.290 | 0.306 | 0.306 | 0.306 |
| Incident Closed - No further action | 0.556 | 0.444 | 0.421 | 0.389 |
| Under Investigation | 0.161 | 0.065 | 0.065 | 0.065 |
| Incident Closed-No Violation Observed | 0.632 | 0.544 | 0.596 | 0.596 |
| Referred to ARA | 0.000 | 0.000 | 0.000 | 0.000 |
| Referred to LMA | 0.444 | 0.000 | 0.000 | 0.000 |
| Referred to WMA | 0.000 | 0.000 | 0.000 | 0.000 |
| Correctly Classified Instances | **61.4498 %** | 55.6837 % | 54.6952 % | 54.6952 % |

## Confusion matrix of KNN “1”

From the matrix, it is clearly determined that most of the Incident Closed-Managed

**Text

Description automatically generated with medium confidence**

# Results

* Now, I would like to discuss results from our data analysis.
* While performing K nearest neighbour (KNN) we concluded that 86% of complaints were closed by managing it.

# Conclusion

* For the conclusion I would like to address our question for which we did this analysis.
* I'd want to break down my questions into three parts: 1) “Out of all kinds of complaints, which type of complaint was reported the most?" As we previously noted in data analysis, "Odor" was the most reported complaint. 2) “Where were the more complaints reported?" According to our findings, we got almost 1/3 complaints from “Baltimore city". 3) “Were all complaints were solved and how?" according to our KNN analysis, we can clearly state that 86% of the complaints were closed/solved by managing them.

**References:**

**1)** <https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/>