Water Quality Sampling Tracking System

Baluyut, Margherita Kyna Amanda A.
Gonzales, Joaquin Emmanuel J.
Piquero, Geriandre M.
Ruiz, Bonnie Jenniedy J.

Asian Institute of Management
Aboitiz School of Innovation, Technology and Entrepreneurship
Master of Science in Data Science
Capstone Project

Legara, Erika Fille T., PhD
Borja, Benjur Emmanuel L., MSc
Valenzuela, Jesus Felix B., PhD

29 June 2022

Abstract

Our Capstone Project provides solutions to Metropolitan Manila Development Authority (MMDA) - Solid Waste Management Office (SWMO) Division's operational challenges associated with its support functions to the Manila Bay Rehabilitation Project: site inspection and water sample collection. We created four data products to aid in decision-making: 1) Water Quality Prediction Model, which determines the characteristics of potential water sampling sites; 2) Geospatial Recommender System, which analyzes potential hotspots for non-compliant establishments based on a list of previously inspected locations; 3) Fieldwork Tool, which digitizes fieldwork data and improves coordination, and; 4) Dashboard Tool, which digitizes reports, improves coordination, and generates reports efficiently.

The Water Quality Prediction Model, which uses textual descriptions of the previously identified water sampling sites and corresponding wastewater discharge represented by the TF-IDF vectors, achieved a 99% test accuracy and a 94% F1-Score using the Gradient Boosting Method. The SHAP interpretability method applied to the model reveals the most influential features: a site with *inspected* discharge pipes *releasing* wastewater that is visually *turbid* or *murky, greasy,* and has a *foul odor* is potentially non-compliant with water quality standards.

The Geospatial Recommender System, which uses the top features from SHAP and the Hamming distance, produced a 9% higher precision score than the baseline and revealed Singalong Street and Dama de Noche Street in Malate as the potential hotspots of unclean drainage systems. The Geospatial Recommender System, which uses the list of previously identified water sampling sites and both Cosine and Euclidean distance metrics, on the other hand, produced a 2% higher precision score than the baseline and revealed A. Mabini Street and Taft Avenue, which cut through both Malate and Ermita, as potential hotspots of non-compliant wastewater from discharge pipes.

Keywords*: water quality prediction, geospatial recommender system, fieldwork tool, dashboard tool*

Acknowledgement

We would like to acknowledge and give sincere thanks to our capstone mentors: Erika Fille Legara, Jesus Felix Valenzuela, and Benjur Emmanuel Borja for providing valuable counsel and direction throughout the duration of this project. We would also like to thank our examiners, Christian Alis and Damian Dailisan for reviewing our work and assuring its quality. Working with these great minds has inspired us to critically appraise the situation of our country and how data science can be applied to improve not only operations, but also lives.

We would also like to extend our thanks to Director Josias Syquimsiam Jr., Engr. Francis Salazar, Engr. Arlene Parafina, Engr. Desiree Pinca, Engr. Faustina Medina and the entire MMDA-SWMO Division for entrusting us with this project. Their dedication and guidance have truly made the last few months an overall enriching experience.

We would also like to express our gratitude to the MSDS 2022 cohort for the extensive amount of resources and time shared with our capstone team.

Finally, we would like to thank the families of our capstone team members, whose encouragement and patience motivated us to complete this project.

Table of Contents

List of Figures

List of Tables

I. Introduction

Data science has proven to be effective in improving operational standards by providing decision-makers with data-driven insights to solve practical problems, implement changes, and monitor performance, resulting in increased efficiency. (Brooke, 2018). One such organization that can benefit from this is the Solid Waste Management Office (SWMO) of the Metropolitan Manila Development Authority (MMDA). The MMDA-SWMO is guided by Administrative Order No. 16 (PCOO, 2019) and the Clean Water Act (DENR, 2022) to provide support functions to ensure the implementation of critical environmental laws and other relevant laws. This includes expediting the rehabilitation and restoration of Manila Bay's coastal and marine ecosystem, as well as ensuring that establishments adhere to water standards along major waterways that surround it. The MMDA-SWMO, in coordination with the MMDA - Flood Control and Sewerage Management Office (FCSMO), began collecting water samples in 2019 and endorses these for testing to the Department of Environment and Natural Resources (DENR) or Laguna Lake Development Authority (LLDA).

However, the required tasks in collecting samples and performing cleaning operations have been adversely affected by the pandemic due to limited resources, a decrease in sampling points to be studied by the department, dependencies on DENR and LLDA, and difficulties in planning the schedule of their workforce. These operations by the MMDA-SWMO are essential for monitoring and improving the water quality of Manila Bay, a water source on which an estimated 23 million Filipinos rely on for water, food, livelihood and recreation (Senate of the Philippines, 2013).

I.A. MMDA-SWMO Water Sampling Process

The MMDA-SWMO has three field tasks: 1) Ocular Inspection; 2) Site Identification, and; 3) Water Sample Collection. Once the office is informed by the MMDA-FCSMO of the schedule and locations of their cleaning operations, the MMDA-SWMO then coordinates with DENR or LLDA regarding the availability of chemicals for laboratory testing. When given the green light, the field team proceeds to the designated location and joins the MMDA-FCSMO cleaning team.

The Ocular Inspection entails visually checking the site, writing down the exact address of a given drainage system (e.g., In front of No. 1341, Paco Hong Giam Taoist Temple, Perez St., Paco, Manila), and filling out the field checklist describing the drainage type and its content (e.g., checking the columns *'Circular Drainage Manhole (CB)', 'Clogged', 'Many Garbage', 'Sludge', 'Wet Soil', 'Leaves',* and *'Rocks'*).

Site Identification is the process of evaluating whether a site is eligible for water sample collection and laboratory testing. The field team first inspects for the presence of a discharge pipe in the drainage system, then checks for flowing wastewater and other requirements. If it does not discharge wastewater during the inspection, the field team checks the column *'Identified for Water Sampling'*, but writes on the '*Remarks'* column that it is not releasing wastewater and for a revisit. If the discharge pipe does release wastewater during the inspection, the field team visually inspects the wastewater and adds its description to the field checklist (e.g. checks the columns *'Identified for Water Sampling'*, *'Brown Wastewater'* and *'Grease'*, and writes on the '*Remarks'* column *'murky wastewater with human feces and foul odor'*, etc.)

The MMDA-SWMO field team also records the progress of the MMDA-FCSMO's cleaning activities (e.g., writes down 80% on the *'% Clean'* column, and 5 on the *'No. of Sacks Removed'* column). To reiterate, if there are no discharge pipes at the drainage site, the task is solely labeled as Ocular Inspection. If, on the other hand, the drainage site has discharge pipes but no flowing wastewater or clear wastewater is being released, the tasks are Ocular Inspection and Site Identification. Finally, if the drainage site has a discharge pipe that releases a sufficient amount of wastewater that is visually inspected as "dirty", and a connection to an establishment is defined, the three field tasks are carried out in order, and the field team proceeds to Water Sample Collection.

DENR and LLDA require three bottles of water samples per drainage site: one for BOD and fecal coliform testing, one for nitrate testing, and one for phosphate testing. After collecting the water samples, properly labeling the bottles with the address of the establishment, and storing them in the ice box, the field team transports them to the DENR or LLDA laboratory in Quezon City within two hours.

The MMDA-SWMO office awaits copies of the laboratory results for about two to three weeks. Once the results are received, the list of non-compliant sites is forwarded to MMDA's Office of the Assistant General Manager for Operations (OAGMO), which prepares the reports to be submitted to the Manila Bay Task Force. If the water sample fails any of the four water quality parameters being tested, the establishment/s connected to the discharge pipe will be sanctioned by DENR and LLDA, as only they have the mandate to penalize violators of water quality standards.

I.B. Objectives

We recognize the difficulties encountered by the MMDA-SWMO in managing its operations due to limited manpower and unoptimized scheduling of site inspection and water sampling activities. Therefore, we intend to utilize data science techniques to determine the other requirements or characteristics, aside from the presence of discharge pipes, that classify sites as potential water sampling sites. Understanding the features of the water sampling sites identified by the MMDA-SMWO is useful in providing recommendations on potential hotspots of non-compliant establishments with water quality standards. Furthermore, we hope to persuade the MMDA-SWMO to commence digital data collection and storage to enable efficient tracking and monitoring of operations inside the division among involved agencies such as the MMDA-FCSMO, DENR, and LLDA.

This study hopes to do this by achieving the following key goals:

1. Build a data collection tool to digitize operations data, especially from fieldwork activities, with the goal of improving coordination between involved offices and within the MMDA-SWMO.
2. Build a dashboard reporting tool to allow the MMDA-SWMO to seamlessly and timely generate reports with appropriate security access.
3. Create a Risk Map to visualize areas covered by the MMDA-SWMO operations.
4. Develop a machine learning model to predict potential water sampling sites based on textual descriptions, allowing the agency to eliminate some inspection-related tasks, such as checking for and recording all possible wastes present at the sites.

5. Develop a recommender system to identify potential hotspots of non-compliant areas where the MMDA-SWMO can concentrate its efforts.

I.C. Significance of Work

Our project aims to address the MMDA-SWMO's operational challenges related to its functions for the Manila Bay Rehabilitation Project by leveraging data science techniques to develop tools that will help the agency improve its operations.

Our first data product, the Water Quality Prediction model, enables the agency to identify features or criteria that determine whether a site is required for water sampling or not, using textual descriptions of sites included in the field report. The model used this data to determine the characteristics of the sites where the MMDA-SWMO intended to collect water samples, where sampling a site implies that it contains a discharge pipe that is releasing visually polluted wastewater. Using this model to predict whether a site should be water sampled or not will help the agency in determining through text alone what features a site must have in order to be flagged as requiring water sampling and laboratory testing.

Our second data product, the Geospatial Recommender System, addresses the difficulties of the MMDA-SWMO in allocating manpower and other resources by providing recommendations of site locations that are potentially non-compliant with water quality standards. These recommended locations are taken from a list of sites previously inspected by the MMDA-SWMO but not identified for water sampling due to various reasons. We augmented the data with geospatial attributes to identify locations not identified for sampling that have the highest similarity to locations previously identified for sampling. This data product makes use of the output from the Water Quality Prediction Model, where the presence of significant characteristics of

sites identified for sampling is used to evaluate the validity of the recommended locations. Through this, we can identify potential pollution hotspots that were originally overlooked, enabling the MMDA-SWMO to prioritize these locations for future site inspections and focus its operations and available resources on key areas. Furthermore, this tool will generate two sets of recommendations. The first set of recommendations consists of sites with discharge pipes that have not yet been sampled but are potentially producing non-compliant water quality, which can be prioritized for water sampling by the MMDA-SWMO once operations resume. The second set of recommendations includes sites that do not have discharge pipes but are highly likely to contain turbid and murky wastewater as well as other waste, which will assist agencies such as the MMDA-FCSMO in making decisions about potential sources of pollution and cleaning operations. This tool can also be shared with other agencies and divisions involved in the field to improve coordination.

Our third data product, the Fieldwork Tool, improves the data collection and organization processes of the MMDA-SWMO by allowing users to easily log important information collected from their site visits and inspections. This tool is hosted on a no-code development platform called AppSheet and allows users to digitally record and store various characteristics of inspected sites using their mobile devices. This data product also includes a visualization of the locations already visited by the agency using Google Maps as the backend. Through this, the Fieldwork Tool can boost the digitization of processes undertaken by the agency, making operations more efficient and organized.

Our fourth data product, the Dashboard Tool, generates consolidated reports that will aid in data-driven decision-making within the organization. It is hosted on Data Studio, which is an open-source dashboard tool offered by Google LLC, and is

used to consolidate and visualize historical and incoming data from fieldwork operations. It allows for real-time generation of reports, including summary statistics and visualizations such as the Risk Map, that can be effortlessly downloaded by members of the MMDA-SWMO. With this, the agency will be able to effectively monitor their operations and gain insights into the effectiveness of their programs.

Overall, this study expands areas of research in water quality monitoring by designing a methodology that could be applicable to the setting of developing countries. In developing countries where water quality is an ongoing issue, and where there is limited data availability, resources and manpower, this study can serve as a proof of concept to design novel uses of data science techniques for water quality prediction, as well as decision-making tools that can encourage data-driven decision-making and optimization of operations.

I.D. Scope of Work

In this project, we developed data products to streamline identification of sites for water sampling, optimize decision-making in terms of site prioritization, and contribute to the improvement of the data collection system of the MMDA-SWMO. Creating the machine learning classification model maximized the textual data provided by the MMDA-SMWO, and geospatial data was not deemed necessary due to several factors: 1) some previously inspected sites are too close to each other and share similar geospatial properties, but with different drainage and wastewater descriptions, and; 2) the agency did not record any other geospatial information other than the site's address, implying its irrelevance. Instead, we found geospatial data augmentation to be only useful for the Geospatial Recommender System data product. Furthermore, the Geospatial Recommender System only recommends locations from

the list of the MMDA-SWMO's previously inspected sites, ensuring the presence of drainage systems.

The creation and maintenance of a database for all gathered information is also out of scope, as well as the integration of all data products into a single platform. Similarly, incorporating new data points (i.e., information, reports, and laboratory results collected while the project is in progress) into the tools also falls beyond the scope of this project. Finally, this project will only include studying fieldwork results obtained in Metro Manila from 2019 to 2021 as provided by the MMDA-SWMO.

I.E. Review of Related Literature

1) Water Quality Information Systems in Developing Countries

Rapid industrialization has been increasingly contributing to difficulties in monitoring water quality throughout the globe. In developing countries with scarce resources, this problem is even more pronounced. Studies emphasize the importance of a good system in place to organize, analyze and share information on water quality, especially in countries with low resources. In a 2020 study, researchers studied water quality monitoring institutions across six countries in sub-Saharan Africa and evaluated them based on the systems and processes they used to transmit and utilize information on water quality data (Kumpel, 2020). In total, 26 institutions from Ethiopia, Guinea, Kenya, Senegal, Uganda, and Zambia were studied from 2012 to 2016. Data flow diagrams were developed for each institution to map the flow of information between different entities and processes, and a generalized version of all institutions is shown in Figure 1.

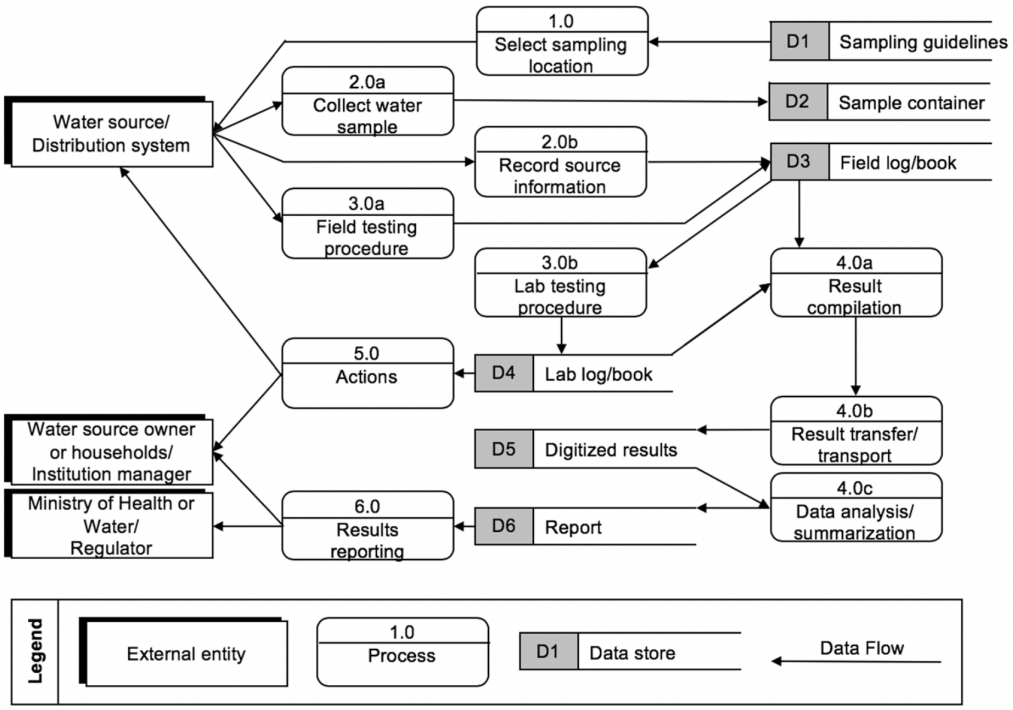


Figure 1. RRL 1: Generalized Data Flow Diagram. The flow of water quality information as depicted between entities and processes is comparable across institutions. (Image from Kumpel, 2020)

The data flow diagram depicted in the figure includes elements such as external entities that institutions must report results to, processes wherein data is transformed or changed in preparation for another stage, data stores wherein data is collected or physically stored, and data flow wherein water quality information is transmitted to another element of the diagram. It was found that comparable structures were employed across institutions. Institutions first selected sampling locations, then collected water samples from these locations, recorded characteristics of the water source, conducted water quality tests on samples collected, compiled and digitized the results, then reported the data to external entities. The study was able to identify key barriers to the proper flow of important information on water quality, many of which are also experienced by the MMDA-SWMO. The major barriers identified were the limited synthesis and analysis of data collected, lack of data literacy skills among

agency staff, poor data sharing procedures between related entities, absence of feedback on the outcomes of the data collected once reports are forwarded to external entities, and insufficient modern infrastructure and technology to support unified platforms and digitization of operations. In this study, researchers emphasized the importance of proper enforcement of testing and reporting by enabling information sharing between related entities, improving the capacity of employees to manage and use data, and integrating water quality data gathering with other associated information systems.

2) Water Quality Monitoring Systems Using IoT devices and Machine Learning

There have been initiatives all over the world to utilize advanced technologies for the optimization of water quality monitoring. A 2018 study conducted in Taiwan used Internet of Things (IoT) technology to create a low-cost, automatic continuous water quality monitoring system, focusing on the Nankan River in Taiwan's largest industrial city, Taoyuan (Chen, 2020). It utilized 100 miniaturized water quality sensors deployed in strategic locations that measured properties of water in real-time over a 24-hour period in a selected area. This system was able to trace upstream river paths towards a source of pollution, provide early warnings for water quality, and detect anomalies in collected data. The study also employed a Recurrent Neural Network (RNN) with Long Short-term Memory (LSTM) as neurons to forecast water quality concentrations from 10 minutes to 3 hours into the future, which was used to identify areas that could have water quality exceeding warning levels. The architecture of the model is shown in Figure 2.

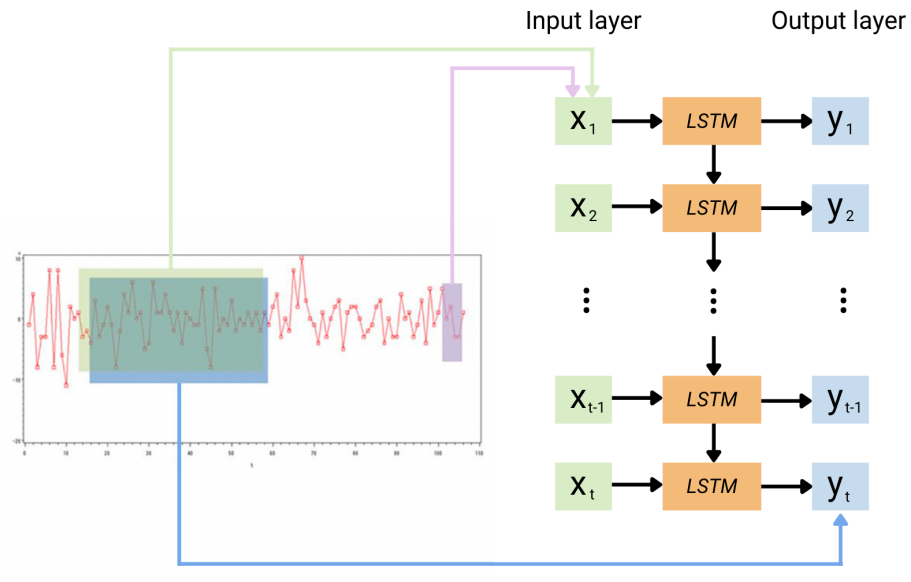


Figure 2. RRL 2: Neural Network Architecture. Forecasting uses historical data of a water quality parameter at defined periods of time apart. (Image adapted from Chen, 2020)

For the input layer of the neural network, the model used historical data of a selected water quality parameter at time t - Δt, where Δt represents the period of time to be forecasted. This is represented by the green block in the figure. The historical data at time t, represented by the blue block in the figure, was then used as the output layer of the neural network. The researchers noted that by feeding real-time monitored data into the prediction model's input layer nodes, represented by the purple block, the physical quantity of the target timing was forecast. The model results were tested on pH data collected from October 2018 to December 2018, where the first 45 days were used to construct the model and the remaining 45 days were used to verify the results. Figure 3 shows these results.

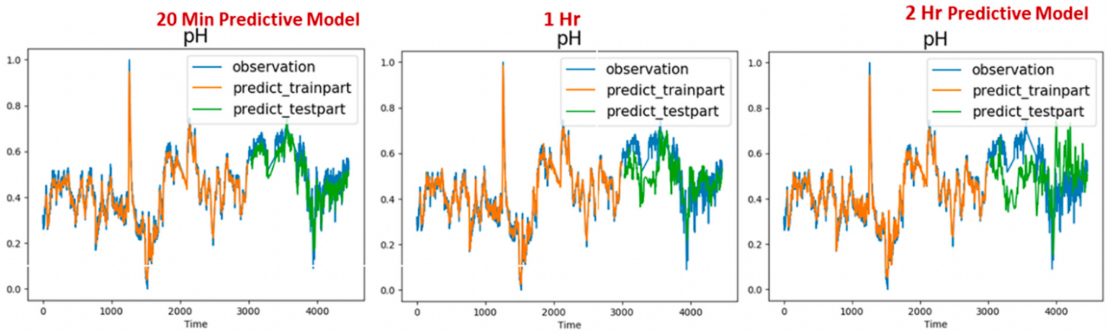


Figure 3. RRL 2: Neural Network Model Results. The 20 minute and 1 hour forecasts showed similar patterns to the ground truth. (Image from Chen, 2020)

The researchers also noted how predictive model forecasts for 20 minutes and 1 hour closely matched the actual data, and the two-hour predictive model matched the overall trend but exhibited errors compared to the actual data. Once the sensors were deployed, the researchers made comparisons of the parameters measured by their devised system with parameter levels determined by traditional laboratory methods. These comparisons were done every three months, and were found to have the following accuracies: 97.7% for pH, 95.7% for temperature, 88.3% for electrical conductivity, 71.3% for chemical oxygen demand, and 67.4% for copper ion content. Along with the data collected by sensors, a database of factories was also used to screen and identify companies within the area of pollution that were likely to have contributed to poor water quality based on waste discharge, manufacturing processes, and even ongoing investigations and public complaints. With this, the system was able to identify spikes in pollutant concentration traced back to electronics factories, and sources of wastewater pollution between monitoring stations. This system effectively reduced the scope of investigation for identifying sources of pollution, improved emergency response time and reduced the requirements of human labor.

3) Water Quality Prediction Using Machine Learning

Significant water quality parameters can also be used as input for machine learning models to predict the water quality index and classification of a given sample. In a 2020 study by Aldhyani et al., deep learning methods such as Nonlinear Autoregressive Neural Network (NARNET) and Long short-term memory (LSTM) were used for water quality index prediction, while machine learning algorithms such as Support Vector Machine (SVM), K-nearest neighbors (KNN) and Naive Bayes were used for water quality classification (Aldhyani, 2020). The water quality index was divided into five brackets and assigned a water quality classification category, as seen in Table 1. The researchers then used historical data from 2005 to 2014 of seven water quality parameters as input to their models. These parameters were dissolved oxygen, pH, conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform.

Table 1. RRL 3: Water Quality Index and Corresponding Water Quality Classification

Water Quality Index Range	Classification
0-25	Excellent
26-50	Good
51-75	Poor
76-100	Very poor
Above 100	Unsuitable for drinking

Table 2. RRL 3: Water Quality Index Prediction Results

Models	Training Data			Testing Data
	MSE	R (%)	MSE	R (%)
NARNET	0.2815	95.97	0.1353	96.17
LSTM	0.1316	93.93	0.1028	94.21

Table 2 shows the complete results of the water quality index prediction. It was found that the NARNET model performed best on the testing data, with an R% of 96.17. The parameters used for the NARNET model were 12 hidden layers, 1:8 number of delays, 100 maximum iterations, 12 maximum epochs, and 1.734 x 10^3 gradients.

Table 3. RRL 3: Water Quality Classification Model Results

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F-score (%)
SVM	97.01	99.23	97.78	94.93	98.54
KNN	83.63	84.73	4.93	87.50	85.84
Naive Bayes	75.20	77.76	91.65	78.08	91.51

Table 3 shows the performance of the water quality classification models based on various evaluation metrics. We can see that the SVM model performed best across all metrics, with a notable F-score of 98.54%. Through this, the study was able to show that using machine learning methods without the use of IoT sensor devices still has the potential of optimizing operations associated with the prediction of water quality.

4) Philippine Research on Water Quality Prediction

In the Philippines, studies have been conducted that explored various methods to predict water quality. One such study even delved into biological methods to assess the quality of water samples. In this 2020 study done in Romblon, macroinvertebrates and coliform presence were used as bioindicators of water pollution for the Bongoy River (Maulion, 2020). Macroinvertebrates present in water samples collected from five sites along the Bongoy River were categorized based on their tolerance to water pollution using the Biological Monitoring Working Party (BMWP) Scoring System. Table 4

shows the resulting score of the Bongoy River, based on adding the individual tolerance scores of each category of macroinvertebrates found among sampling sites.

Table 4. RRL 4: BMWP Score of Bongoy River

Invertebrates	Sampling Sites					BWMP TAXA Score
	1	2	3	4	5	
Odonata: Dragonfly	+	+	-	+	-	8
Plecoptera: Stonefly	+	+	+	-	-	10
Psephinedae: Water Penny	+	-	+	-	+	10
Trichoptera: Caddisfly	+	+	-	-	-	10
Hemiptera: Water Strider	+	+	+	+	+	10
Annelida: Oligochaeta	-	+	-	-	-	1
Crustacea: Shrimps	+	+	+	+	+	6
Mollusca: Pouched Snails	+	-	+	+	-	3
Pelecypoda: Clams	+	+	-	+	-	6
TOTAL						57

Through this method it was concluded that the Bongoy River achieved 57 points for its BMWP score, while the required score of a ‘very clean’ water source is above 150. It was therefore considered to be a ‘Moderately-Polluted’ river.

Another study made in 2020 utilized machine learning algorithms paired with sensor nodes stationed over 23 rural areas in the Southern Luzon Region to develop a data-driven water quality monitoring and classification system (Alipio, 2020). Data collected by the sensors consisted of the features date, time, location, source, pH, turbidity, total dissolved solids, temperature and terrain. Machine learning models were trained on 500 water samples collected from the selected locations over a time period of 90 days, as illustrated in Figure 4.

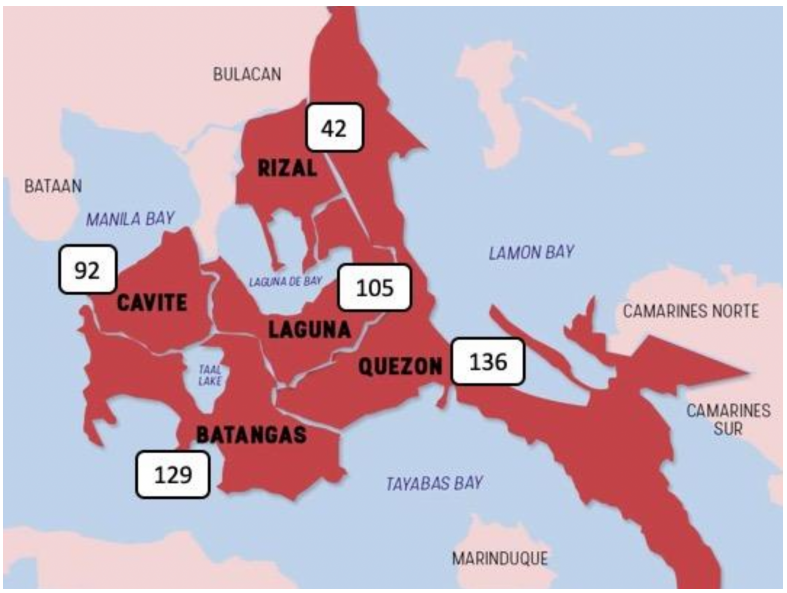


Figure 4. RRL 4: Water Sampling Locations. The 500 water samples were sourced from across the Cavite, Laguna, Batangas, Rizal and Quezon provinces. (Image from Alipio, 2020)

The study employed ensemble learning using a hard-voting method, and used K-Nearest Neighbors (KNN), Naive Bayes, and Classification and Regression Tree (CART) models to make predictions on whether a water sample is potable or non-potable. The model results were then combined into an ensemble voting classifier model using the hard-voting method, which gets the majority vote among all models and uses it as the final prediction. The results were then evaluated according to the following metrics: accuracy, precision, F-measure and kappa statistic. Figures 5a to 5d show the model performance for each metric.

As shown in Figures 5a and 5c, the voting classifier performed best, with an accuracy of 97% and a kappa statistic of 92.48%. The voting classifier also obtained the highest precision at 100% for potable samples and 90% for non-potable samples, which can be seen in Figure 5b. Figure 5d shows that the voting classifier also gave the best performance according to the F-measure score, with 98% for potable samples and 95% for non-potable samples. Additionally, the study used 30 random water

samples to compare their model predictions with traditional laboratory analysis results and found that 90% of samples matched the model's predictions.

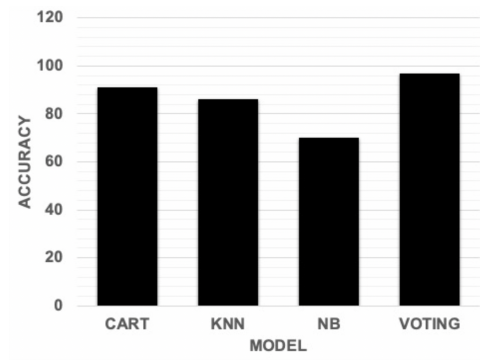


Figure 5a.

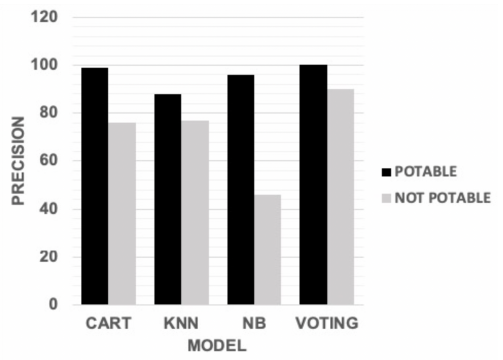


Figure 5b.

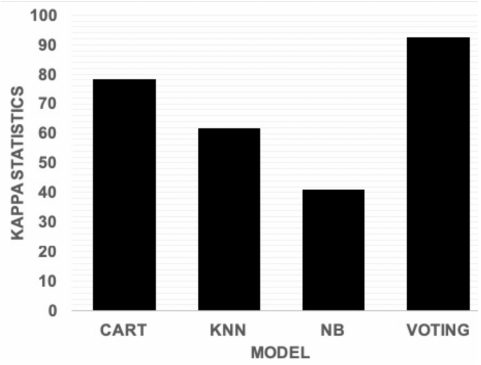


Figure 5c.

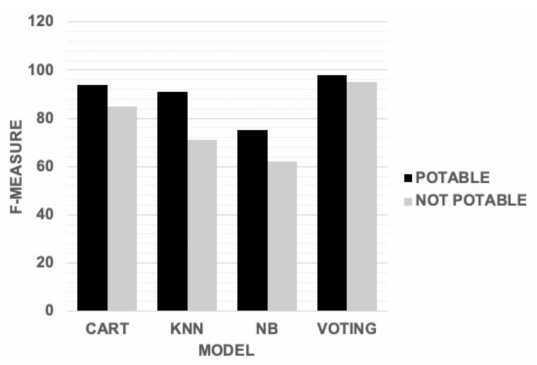


Figure 5d.

Figure 5. RRL 4: Water Quality Prediction Results. The voting classifier performed best across all metrics. (Image from Alipio, 2020)

Indeed, we can see the potential for using data science and data management technologies to aid the ongoing struggle of developing countries when it comes to monitoring water quality and ensuring clean water access for all citizens. Since the first quarter of 2020, the Department of Science and Technology (DOST) has been collaborating with the University of the Philippines (UP) on a project called IM4ManilaBay, which stands for the Integrated Mapping, Monitoring, Modeling and Management System for Manila Bay and Linked Environments (Nazario, 2020). The project involves the University of the Philippines Diliman College of Engineering, the

Institute of Civil Engineering, and the Training Center for Applied Geodesy and Photogrammetry (UP TCAGP). It aims to develop an integrated system to monitor and evaluate solid waste management operations in areas connected to the Manila Bay watershed, use geospatial technologies to map water quality of linked systems to Manila Bay, develop technologies to manage and treat dredged material from selected river systems, and use hydrodynamic analysis to model the transport of materials along the Manila Bay, Pasig River and Laguna Lake water systems.

Our study aims to contribute to this body of research by developing water quality monitoring solutions that use alternative machine learning methodologies and limited available data to predict water quality and recommend potential non-compliant areas. Instead of using sensor devices or water quality parameter data, our methodology utilizes geospatial methods and physical textual descriptions of flagged locations as input to our models.

II. Data and Methodology

II.A. Data Description

The primary data used in this project was derived from the three Microsoft Excel files provided by the MMDA-SWMO as part of the consolidated report for ocular site inspection, site identification, and water sampling activities for the Manila Bay Rehabilitation Project. An example is provided by Figure 6. These files were encoded by the MMDA-SWMO from field checklists filled out by the field team internally called *'Oysters'* to record information of the conducted inspection.

The masterfiles contain several columns with information on each inspected site. The *'Location'* column contains addresses of sites identified by the MMDA-FCSMO that are potentially non-compliant with water quality standards. Additionally, there is no fixed list of sites provided by THE MMDA-FCSMO.

The *'Ocular Inspection'* column shows which sites were visited by the MMDA-SWMO for inspection. The MMDA-SWMO is required to visit all listed sites for ocular inspection, to confirm if they are indeed discharging dirty wastewater.

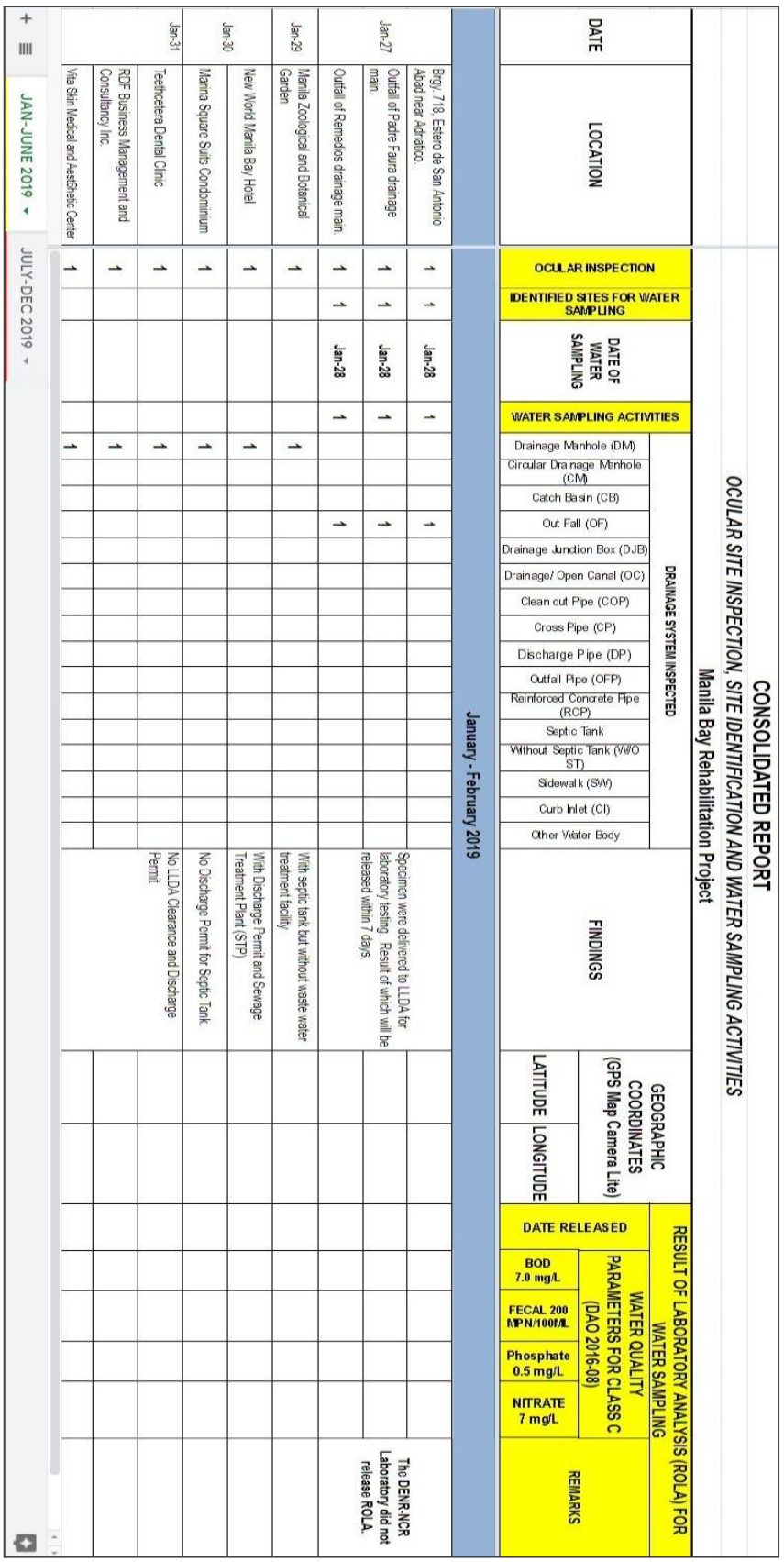
The *'Identified for Water Sampling'* column contains information on whether a site was identified by the agency as in need for water sampling, which depends on whether the site was visually determined to have water quality below standards and whether it contains a discharge pipe releasing dirty wastewater. It is critical to identify sites for water sampling, because these sites will correspond to establishments that are potentially non-compliant and must be sanctioned by the DENR or LLDA. Due to this, the column *'identified for water sampling'* was chosen as our target variable.

The columns under *'Drainage System Inspected'* contain indicators of which types of drainage system was found in the site. Within these columns we can find

information such as whether a site contains a discharge pipe or not, specifically, under the column *'Discharge Pipe (DP)'*.

The *'Findings'* column contains textual descriptions of the characteristics of the sites inspected by the MMDA-SWMO, which are summarized by the field team. We decided that descriptions in this column were the most appropriate features we could use to predict our target variable of whether a site would be identified for water sampling or not.

The *'Geographic Coordinates'* column contains the latitude and longitude of selected sites, while the columns under *'Result of Laboratory Analysis'* contain the measured quantities of four water quality parameters: biochemical oxygen demand (BOD), fecal coliform, phosphate, and nitrate. If a location is identified for water sampling and successfully sampled and endorsed for laboratory analysis to DENR, both geographic coordinates and results of laboratory analysis for water sampling will contain values.



CONSOLIDATED REPORT

OCULAR SITE INSPECTION, SITE IDENTIFICATION AND WATER SAMPLING ACTIVITIES

Manila Bay Rehabilitation Project

JAN-JUNE 2019 | JULY-DEC 2019

DATE	LOCATION	OCULAR INSPECTION	IDENTIFIED SITES FOR WATER SAMPLING	DATE OF WATER SAMPLING	WATER SAMPLING ACTIVITIES	DRAINAGE SYSTEM INSPECTED	FINDINGS	GEOGRAPHIC COORDINATES (GPS Map Camera Lite)		RESULT OF LABORATORY ANALYSIS (ROLA) FOR WATER SAMPLING					
								LATITUDE	LONGITUDE	DATE RELEASED	WATER QUALITY PARAMETERS FOR CLASS C (DAO 2016-08) BOD 7.0 mg/L	FECAL 200 MPN/100ML	Phosphate 0.5 mg/L	NITRATE 7 mg/L	REMARKS
							January - February 2019								
Jan-27	Brgy. 718, Estero de San Antonio Abad near Adriatico.	1	1	Jan-28	1	Drainage Manhole (DM)									
	Outfall of Padre Faura drainage main.	1	1	Jan-28	1	Circular Drainage Manhole (CM)									
	Outfall of Remedios drainage main.	1	1	Jan-28	1	Catch Basin (CB)									
Jan-29	Manila Zoological and Botanical Garden	1			1	Out Fall (OF)									
Jan-30	New World Manila Bay Hotel	1			1	Drainage Junction Box (DJB)									
	Marina Square Suits Condominium	1			1	Drainage/ Open Canal (OC)									
Jan-31	Teethcetera Dental Clinic	1			1	Clean out Pipe (COP)									
	RDF Business Management and Consultancy Inc.	1			1	Cross Pipe (CP)									
	Vita Skin Medical and Aesthetic Center	1			1	Discharge Pipe (DP)									
						Outfall Pipe (OFP)									
						Reinforced Concrete Pipe (RCP)									
						Septic Tank									
						Without Septic Tank (WWO ST)									
						Sidewalk (SW)									
						Curb Inlet (CI)									
						Other Water Body									
Jan-27	Outfall of Padre Faura drainage main.						Specimen were delivered to LLDA for laboratory testing. Result of which will be released within 7 days.								The DENR-NCR Laboratory did not release ROLA.
Jan-29	Manila Zoological and Botanical Garden						With septic tank but without waste water treatment facility								
Jan-30	New World Manila Bay Hotel						With Discharge Permit and Sewage Treatment Plant (STP)								
	Marina Square Suits Condominium						No Discharge Permit for Septic Tank.								
Jan-31	Teethcetera Dental Clinic						No LLDA Clearance and Discharge Permit								

Figure 6. MMDA-SWMO Sample of the Consolidated Report. The fields contain information on each inspected site that is collected through ocular inspection.

II.B. Data Extraction and Preprocessing

The full checklists used by the MMDA-SWMO for the years 2019, 2020 and 2021 that serves as the masterfile used for inspection and logging of results was made available. We extracted relevant features from these Excel sheets in order to conduct our analysis, which were the features *'Location'*, *'Identified for Water Sampling'*, and *'Findings'*. Figure 7 below shows the count of locations identified for sampling or not based on the checklist of the MMDA-SWMO for the years 2019 to 2021.

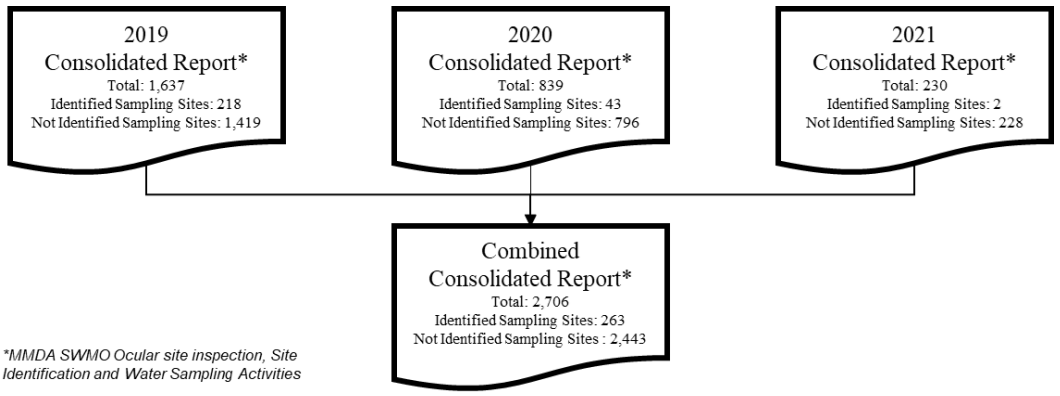


Figure 7. Summary Count of Identified Water Sampling Sites and Non-Water Sampling Sites

To augment the data provided by the MMDA-SWMO, we used Geocode, a Google Sheets extension that uses Google's services to extract the geospatial coordinates of the different sites that were inspected by the agency, as the original dataset only contained longitude and latitude values for those areas that had laboratory results. Once the additional longitude and latitude data was available, we also gathered counts of amenities such as restaurants, fast-food-chains, hospitals, parking spaces, places of worship and other points of interest such as the number of residential areas, commercial establishments, construction activities, among others, which are within 150 meters of each inspected site using Open Street Maps (OSM). By collecting these data points, we were able to determine the possible OSM features which may

contribute to the state of the different drainage systems that are inspected by the MMDA-SWMO.

II.C. Assumptions

This project was undertaken on the assumption that the data products meet the objectives outlined in the project brief: 1) the Fieldwork Tool to improve operational coordination; 2) the Dashboard Tool to seamlessly generate reports with appropriate security access, while also improving operational coordination; 3) the Risk Map, which is integrated into both the Fieldwork and Dashboard Tools, to visualize the areas covered by the MMDA-SMWO operations and to identify potential non-compliant areas, and; 4) the Geospatial Recommender System to identify potential non-compliant areas where the MMDA-SWMO can focus its resources. We also developed a model that determines the characteristics of a site qualified for water sampling, with the goal of assisting the agency in reducing a significant bottleneck in its operations.

Another assumption of this study is the accuracy and completeness of the data provided for the years covered, 2019 to 2021. Similarly, the submitted list of locations is representative of all areas under the MMDA's jurisdiction and adequate for mapping out the agency's operations. The MMDA-SWMO's operations were undisrupted in 2019, and the division inspected 12 of 16 Manila Districts, totaling around 1,600 ocular inspections, as shown in Figure 8, which is a snapshot from a standalone dashboard of their submitted historical data.

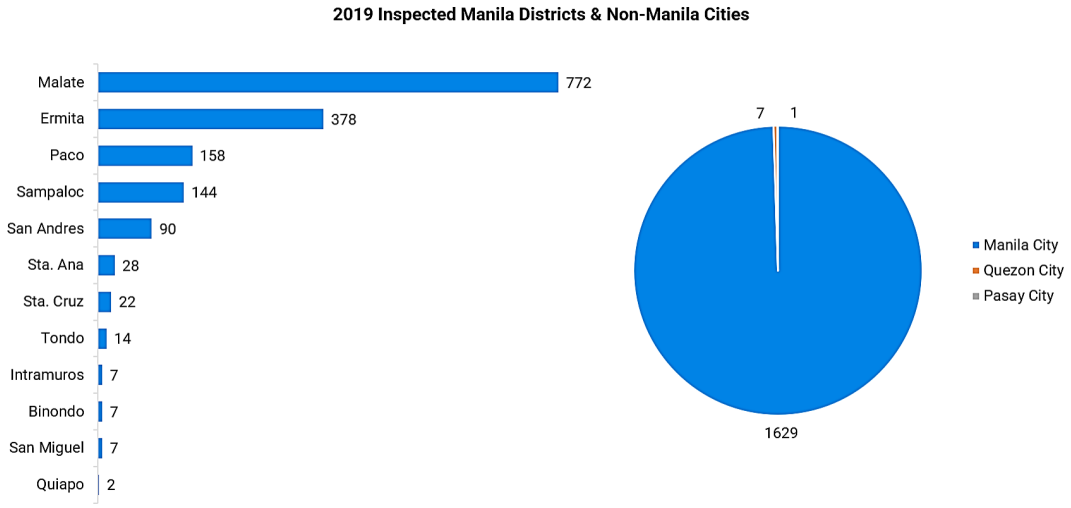


Figure 8. MMDA-SWMO Inspected Manila Districts in 2019. The agency inspected 12 of the 16 Manila Districts, accounting for 99.5% of all locations inspected in 2019.

The Water Quality Prediction Model was built with the assumption that the limited data provided would be sufficient, and that the data imbalance problem would be addressed using oversampling techniques. Furthermore, the discovered discrepancies in the provided data were eliminated when the '*discharge pipe/s*' and *'dp'* were removed from processed text data.

The machine learning model's objective is to determine the characteristics of the previously inspected sites based on the descriptions found in the *'Findings'* column; the team assumed that additional geographic data, such as the amenity features, was not considered necessary because the MMDA-SWMO does not record these details during operations. Furthermore, previously inspected drainage sites may be feet apart yet have different descriptions, thus acquiring and analyzing geographical features per inspected site may be ineffective.

The team also assumed that the insights and recommendations generated in this project would not be used to generalize water sampling sites and operational processes in other areas or agencies outside of the MMDA-SMWO.

II.D. Limitations

This Capstone project is not without limitations. The solutions proposed were based on the limited data provided by the MMDA-SWMO. The breakdown of the 2,706 observations is shown in Table 5.

In order to find the features that characterize a potential water sampling site, the machine learning models utilized the texts from the *'Findings'* column shown in Figure 6, as the laboratory-tested water samples made up only about 3% of the total observations, and all failed the evaluated water quality parameters. Furthermore, the laboratory tests are limited to four parameters only: biochemical oxygen demand (BOD), fecal coliform, phosphate, and nitrate content.

Table 5: Breakdown of the MMDA-SWMO Provided Data

	2019	**2020**	**2021**
Dates Covered	161 field days Q1 to Q4 27 Jan to 18 Dec	63 field days Q1 & Q4 only 02 Jan to 18 Mar 30 Sep to 25 Nov	19 field days Q1 only 21 Jan to 19 Mar
Number of locations inspected	1,637	839	230
Number of observation with written findings	1,625	838	230
Number of sites identified for water sampling	218	43	2
Number of lab tested	83	6	0
Number of released lab results	71	6	0

Table 5: Breakdown of the MMDA-SWMO Provided Data *(continuation)*

	2019	**2020**	**2021**
Number of samples that failed the water quality parameters test	71	6	0
Number of samples that passed all the water quality parameters test	0	0	0

We designed the Geospatial Recommender System with the goal of exploiting the existing list of inspected locations from 2019 to 2021, which totals to nearly 3000 drainage sites. Exploring new locations is not an option because the MMDA-SWMO did not provide a list of new locations of drainage sites, and the Capstone Team has no source of new coordinates that ensures the presence of drainage systems or discharge pipes.

The Manila Bay Rehabilitation Project is focused on areas surrounding the Manila Bay, thus 94% of the inspected locations fell under the jurisdiction of Manila City. Other areas inspected included parts of Pasig City and Quezon City. In line with this, the MMDA-SWMO data was not digitized, posing concerns about information loss and veracity. We lack access and resources to verify the completeness and accuracy of the provided data, particularly the locations and their longitude and latitude coordinates, as well as the written descriptions.

The limited data provided by the MMDA-SWMO was also insufficient to draw conclusions about the applicability of the results during and after the coronavirus pandemic. This public health disruption forced a large portion of the population to

work and study remotely, resulting in increased water consumption and wastewater discharge (Manoiu, 2022).

II.E. Exploratory Data Analysis

The following section shows the results of exploratory data analysis conducted on the data provided by the MMDA-SWMO.

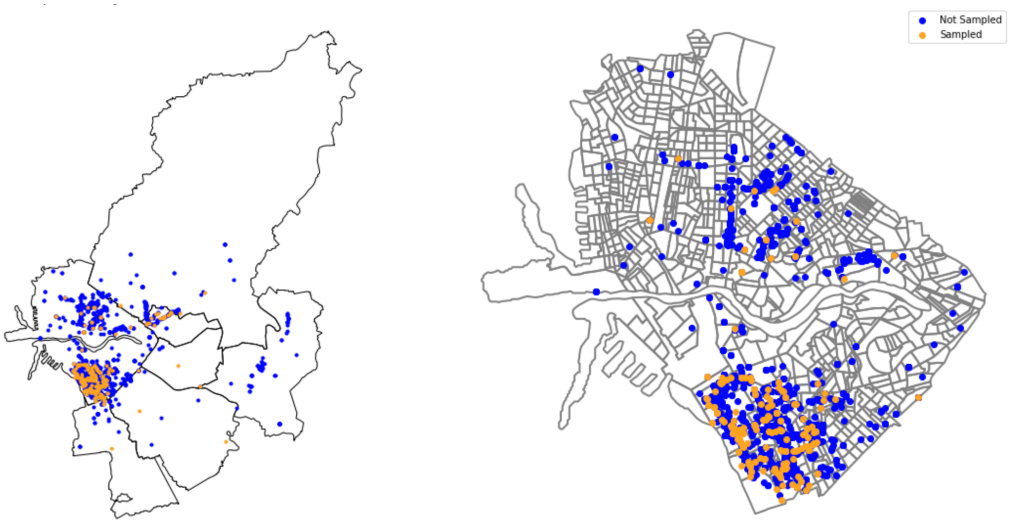


Figure 9. Different Sampling Points from (a) NCR and (b) City of Manila. The MMDA-SWMO's operations spanned across Manila City, Quezon City, San Juan City, Mandaluyong City, Pasay City and Pasig City.

The geospatial representation in Figure 9a depicts a map of the entire National Capital Region (NCR), with blue dots representing the inspected locations for years 2019 to 2021, and the orange dots representing the sites that were identified for sampling. Figure 9b shows a zoom in on Manila City, while the rest of the inspected sites were in Quezon City, San Juan City, Mandaluyong City, Pasay City and Pasig City. Out of the 2,706 sites that were inspected by the MMDA-SWMO in Metro Manila from 2019 up to 2021, only 263 sites were identified for water sampling.

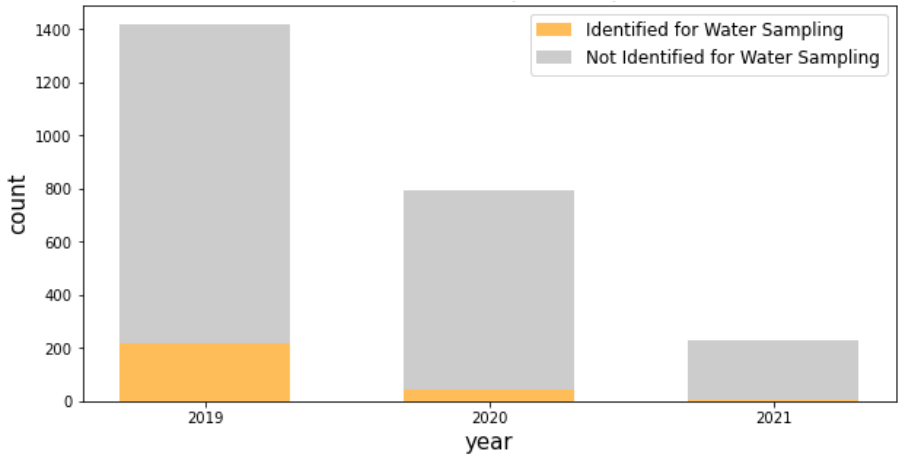


Figure 10. Number of Ocular Inspections per Year. There were more sites not identified for water sampling compared to those that were identified.

When the MMDA-SWMO identifies a site for water sampling, this means they have conducted an ocular inspection at the site location and visually assessed the quality of water in the site to be potentially below water quality standards. They then collect water samples from discharge pipes and endorse these to DENR for laboratory testing to be able to determine if the establishment the wastewater was collected from is deemed to be non-compliant. In Figure 10, we can see that sites where ocular inspection was done consisted of around 1419 locations in 2019, but because operations were affected by the COVID-19 pandemic, this decreased to around 228 in 2021. As of the present, these operations have been stalled due to restrictions brought about by the COVID-19 pandemic. Furthermore, we can see that the number of sites identified for sampling comprises only a small portion of the total number of sites visited by the MMDA-SWMO for ocular inspection, which is a waste of the agency's valuable resources and time.

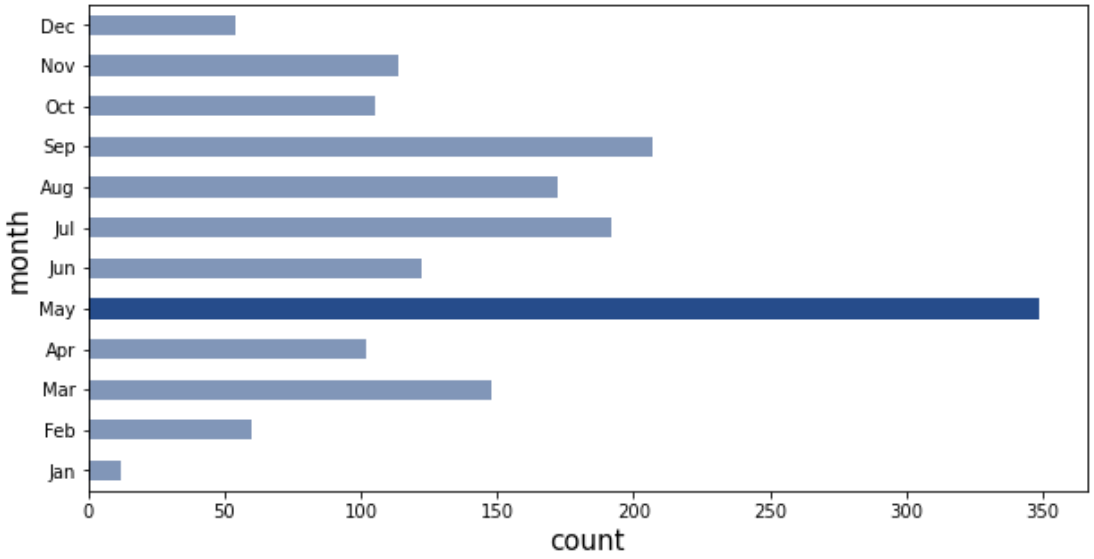


Figure 11. Number of Ocular Inspections per Month in 2019. The month of May showed the most activity in terms of ocular inspections.

Drilling down per month for the year 2019, which is the only complete year of operations for the MMDA-SWMO, we see the number of ocular inspections conducted varies greatly month to month, with as many as 349 occurring in May and as few as 12 in January.

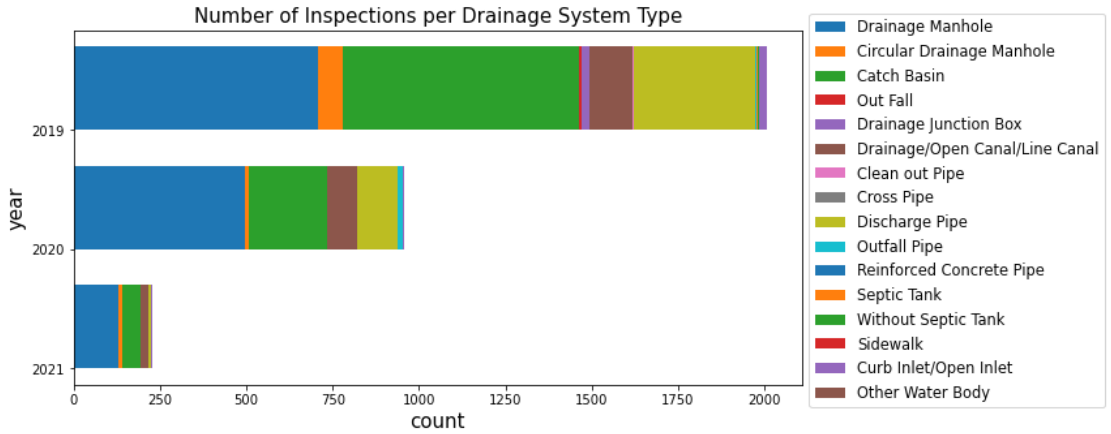


Figure 12. Number of Inspections per Drainage System Type. Majority of the sites visited for ocular inspection possessed drainage systems other than a discharge pipe.

From Figure 12, we can see that among the sites the MMDA-SWMO conducted ocular inspections in, the largest number of them had a drainage manhole,

followed by a catch basin, discharge pipe, and drainage/open canal/line canal. Although the presence of a discharge pipe is not the only requirement needed for water sampling, it is still a prerequisite for sampling, and the chart shows that the majority of sites inspected do not possess one. Consistent with Figure 10, all these counts have been decreasing each year due to limited site visits, and were greatly affected by the COVID-19 pandemic in 2021.

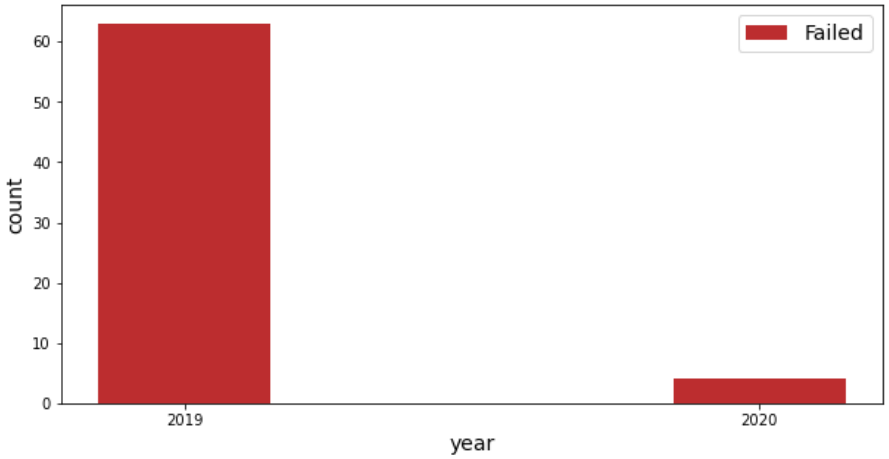


Figure 13. Number of Sites with Laboratory Analysis Results. All available data on water quality parameter levels exhibit failing results.

Upon review of the different laboratory results for all sites, we observed that all sites in which water samples were collected had failed to comply with the water quality parameters mandated by the MMDA-SWMO for the Manila Bay rehabilitation. Due to the fact that there was no available data on water samples that had passed water quality standards, we lacked sufficient data to make predictions based on the water quality parameter results of the DENR or LLDA laboratory analysis.

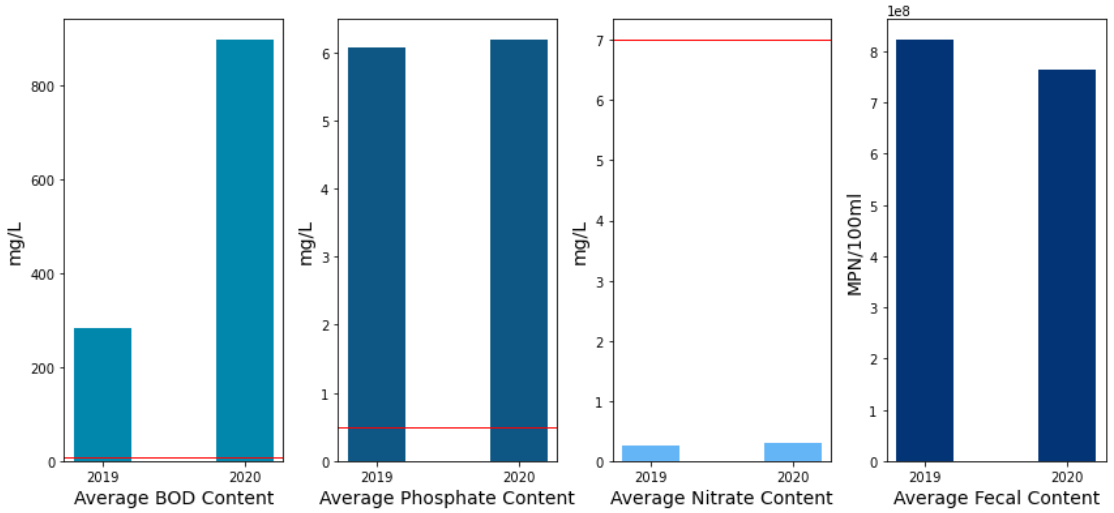


Figure 14. Comparison of Laboratory Analysis Results. Average values of BOD, phosphate, and fecal content are above acceptable limits, represented by the red line.

The available laboratory analysis results of sites endorsed for laboratory testing to DENR is shown in Figure 14. Water samples were tested based on the presence of four parameters: biochemical oxygen demand (BOD), phosphate, nitrate and fecal coliform. Their average value is shown per year and the red line represents the acceptable limits of their values–for fecal coliform, the red line is much lower than the average value which is why it is barely visible. Among these parameters, the average content of BOD, phosphate and fecal coliform exceed the threshold of acceptable limits. For the BOD parameter in particular, we can see the 2020 result was approximately triple the value of that in 2019. Higher values of BOD indicate more oxygen being removed from the water by microorganisms, therefore representing lower water quality.

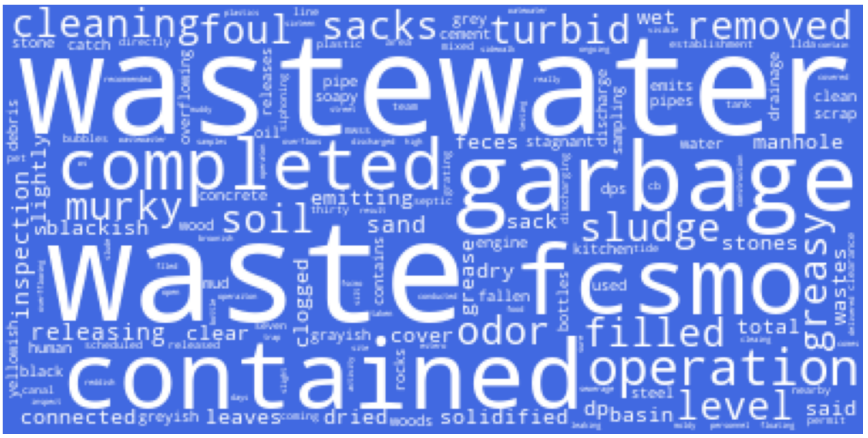


Figure 15. Words Used in *'Findings'* Column. The most common terms used to describe a site include the quality of the wastewater and waste it contains.

Figure 15 focuses on the '*Findings*' column, which contains textual descriptions of the characteristics of the sites inspected by the MMDA-SWMO. The figure shows the most common terms present in this column, and we see the terms 'wastewater', 'waste', 'garbage' and 'contained' are among the most frequently used words to describe the water quality of inspected sites. We therefore decided that descriptions in this column were the most appropriate features that could be used to predict the target variable of whether a site would be identified for water sampling or not.

II.F. Model Building

Utilizing the data provided by the MMDA-SWMO along with the augmented geospatial data from OSM, we used supervised machine learning to identify the characteristics of the sites identified by MMDA-SWMO as water sampling sites (but did not necessarily proceed with actual water sample collection).

For our first data product, the Water Quality Prediction Model, we began by collecting the text descriptions from the *'Findings'* column shown in Figure 6, which

contained observations and findings taken from different sites the agency had inspected. This column contained descriptions for each inspected site based on an internal checklist of the MMDA-SWMO, meaning that it was not free-form. A sample of this checklist is shown in Figure 16. Unfortunately, the actual checklist used is paper-based and no digital version is provided by the MMDA-SWMO.

LOCATION	DRAINAGE MANHOLE	CIRCULAR DRAINAGE MANHOLE	CATCH BASIN	CLOGGED	FILLED	CONTAINED	STAGNANT	WASTEWATER						GREASE			WITH FOUL ODOR	GARBAGE		SOIL		LEAVES	SLUDGE	% CLEAN	NO. OF SACKS REMOVED	NO. OF PERSONNEL	INCLUDING DRIVER	REMARKS
								BLACK	GREY	BROWN	RED	CLEAR	YELLOW	SOLIDIFIED	GREASY	SLIGHTLY		MANY	FEW	WET	DRY							
ALONG: 1161 A. Mabini St. **CORNER:** Malate Mla. **IN FRONT OF:** BF Money Changer/Yue Lai Seafood	✓					✓			✓						✓				✓									Sunny Weather Temp. 35° C
ALONG: **CORNER:** **IN FRONT OF:**																												
ALONG: **CORNER:** **IN FRONT OF:**																												
ALONG: **CORNER:** **IN FRONT OF:**																												

Figure 16. MMDA-SWMO Fieldwork Checklist. Checked columns correspond to descriptions used in the *'Findings'* column of the agency's consolidated report.

Before data preprocessing and applying different machine learning algorithms, we noted that there were 13 records that had missing records in the *'Findings'* column. Out of these 13 records, 8 were endorsed for testing and 5 were not. For modeling purposes, we filtered out these 15 records. We also ensured that the terms *'discharge pipe/s'* and *'dp'* were removed as tokens from our machine learning model, as this is a primary requirement for water sample collection. Next, we constructed a holdout set that consisted of 30% of the data provided by the MMDA-SWMO. This set was randomly selected as it was used to evaluate the capability of the implemented machine learning models to make predictions for new data points that were not used for model training. The remaining 70% of the data was used for model training and

hyperparameter tuning. Figure 17 below illustrates how we divided the data in preparation for machine learning. Given that there were only a few identified water sampling sites, it was necessary for us to address the class imbalance of the target variable when attempting to model it. Hence, we explored resampling techniques on the training set to provide more information to the machine learning models and help capture the inherent patterns found in the data. The text vectorization, resampling, and the model training process was fitted to the training data using scikit-learn's Pipeline function. A validation set was also generated from the training set using a 5-fold Stratified K-Fold cross validation in order to find the optimal hyperparameters of all the machine learning algorithms that were trained. We then used scikit-learn's Grid Search function to loop over all possible combinations of the different hyperparameters that were relevant for each trained model. We then chose the hyperparameters for each model that gave the highest accuracy on the test set.

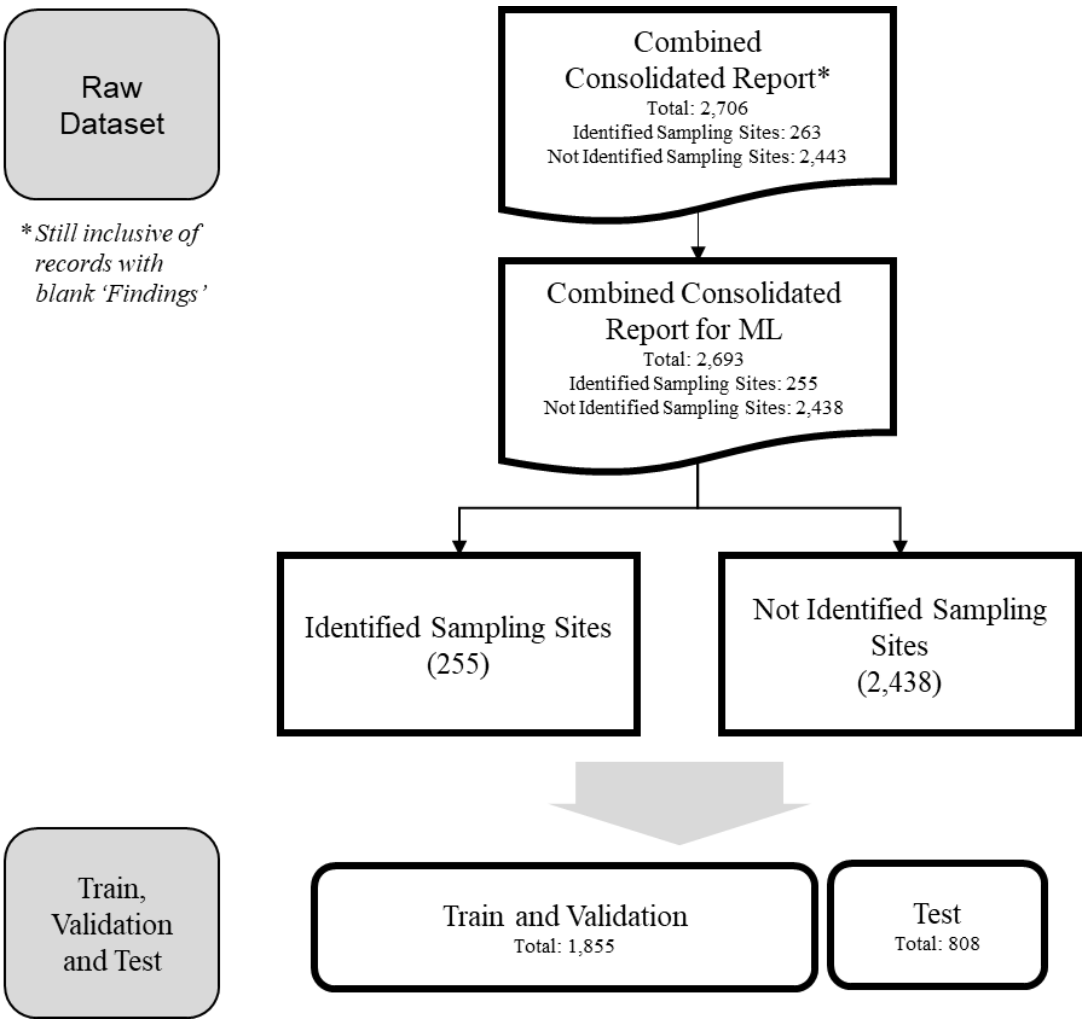


Figure 17. Train, Validation and Test Segmentation of the MMDA-SWMO Data. The data was divided into a test set, and a train and validation set for machine learning.

Using the optimal hyperparameters of each machine learning model, we then evaluated the performance of each model on the holdout set. Aside from using accuracy to evaluate the quality of each model's predictions, we also used the F1-Score to compare the actual target values of the holdout set with the predictions of each model. The F1-Score is an evaluation metric that calculates the harmonic mean between precision and recall. By calculating the F1-Score of the model's predictions on the holdout set, we were able to determine whether the created models could actually recognize all of the sites in the holdout set that were identified as water

sampling sites, rather than the machine learning model simply predicting every inspected site as water sampling sites. We hope that by using our machine learning model, we can assist the MMDA-SWMO in automating the decision-making process of identifying water sampling sites applicable for water sample collection.

In order to determine the keywords from the text data that drive the prediction of whether a site is a water sampling site or not, we employed different global and local model interpretability techniques. We then used the extracted important features to evaluate the performance of the second data product, the Geospatial Recommender System, which is an application of information retrieval that can identify which of the previously inspected sites have similar geospatial features to those identified by the MMDA-SWMO as water sampling sites. The recommended sites by the information retrieval model could serve as hotspots of non-compliant establishments where the MMDA-SWMO can focus its operations and available resources.

The information retrieval problem is a data analytics task often applied to text data which aims to identify the top k most similar documents in the entire database D when compared to a particular query q. For this study, we applied the information retrieval problem to all of the sites that were only inspected but were not considered as water sampling sites, and determined which among these had similar geospatial properties with the identified water sampling sites. Given that the information retrieval problem outputs similar documents per query, the recommendations that the model would generate are only specific to the query of interest. With this, we needed to aggregate the similarity scores generated by the algorithm for all possible queries of interest, which were all of the identified water sampling sites by the MMDA-SWMO. By aggregating the similarity scores, we can make recommendations that are global in nature as all possible queries are fed into the algorithm. The pseudocode for making

global recommendations using the information retrieval problem could be seen in Figure 18 below, where ‘non-identified sites’ refer to not identified water sampling sites and ‘identified sites’ refer to identified water sampling sites.

```
for i in all non-identified sites:

  for j in all identified sites:

    Calculate distance or similarity score of the geospatial
    features between non-identified site i and identified site j

  Getting the mean distance or similarity score for non-identified site i

  Append the mean score to an array

Sort list by increasing order (for distances) or decreasing order (for similarity
scores) to determine the non-identified sites that are most similar to the identified
sites.

Return the top k sites recommended by the model.

Evaluate model recommendations following relevance conditions.
```

Figure 18 .Pseudocode for Global Recommendations. The non-water sampling sites that are most similar to the identified water sampling sites were found using IR.

Once the model was able to generate recommended sites that must be revisited again by the MMDA-SWMO, we evaluated the quality of these recommendations to see whether this was aligned with MMDA-SWMO’s identified water sampling sites. To do this, we used the top keywords that distinguish water sampling sites from those that were not, as generated by the model interpretability methods used to explain the machine learning models created for the first data product, the Water Quality Prediction Model. If a site in the non-identified water sampling group had terms in their site descriptions that were related to the identified water sampling sites, that site was treated as a relevant location among the entire set of locations from which the

MMDA-SWMO failed to collect water samples. For our recommender system to be useful, it must be able to capture most if not all of the relevant locations when finding the top *k* most similar locations to those identified water sampling sites – where all of these sites had the potential for substandard water quality, as described during inspection. In this recommender system, the top *k* most similar locations to the identified water sampling group may differ depending on the distance or similarity metric used to calculate the closeness of the geospatial vectors. This study considers the distance metric as a hyperparameter that can be tuned, as we want to identify the distance metric that yields the most number of relevant locations from the top *k* query. Each recommender system would be evaluated using the model's precision, or the fraction of relevant results that were returned out of the top *k* most similar locations to the identified water sampling. Through this data product, we help the agency practice a more targeted approach in determining which sites to prioritize as soon as field activities resume. A summary of the entire data products building pipeline is found on Figure 19.

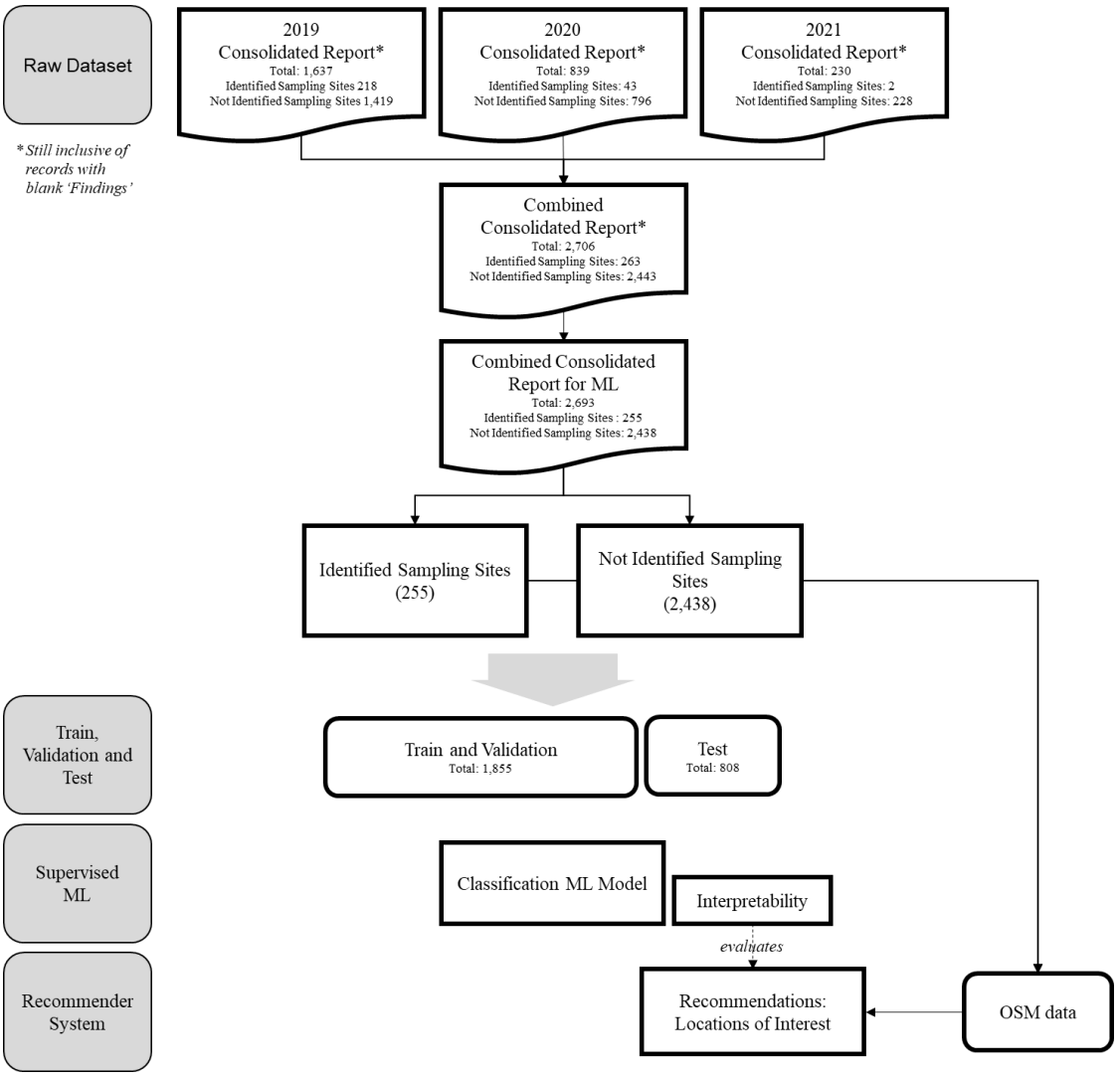


Figure 19. Data Products Build Pipeline.
Illustrates the data used for each data product.

III. Discussion of Models

Our study used different machine learning algorithms from the data exploration up to the data modeling stage. In order to be able to explore insights from the findings of each ocular inspection result along with the nearby geospatial features surrounding each site, we employed four different clustering algorithms to model the distribution of these feature spaces. We used representative-based clustering algorithms such as k-means clustering to partition the feature space into *k* subgroups or clusters which are each defined by a representative point that explains the cluster. For k-means clustering, this representative point is denoted by the mean of all the points within the same cluster. In performing k-means clustering, we aim to maximize the similarities (or conversely, minimize the distances) between each point within a specific cluster relative to the representative point or the centroid of each cluster. We also used Ward's agglomerative clustering method, which is a hierarchical clustering approach. This unsupervised machine learning method creates clusters using a bottom-up approach, where each observation in the dataset initially forms its own respective cluster, and records that are highly similar with each other are then merged together to form bigger clusters. Similarity in Ward's method is ensured by minimizing the increase in distances within points as subclusters are aggregated. Ward's method of linking clusters together was used as this approach tends to form compact and highly balanced clusters unlike other approaches such as the Single-linkage method, which is highly sensitive to outliers as this will merge two clusters even if there is one pair of points among each cluster that is close to one another, irrespective of all other points in the clusters. Probabilistic clustering techniques such as the Gaussian Mixture model was also used to assign each datapoint into soft clusters by assigning a probability that each site falls under a particular cluster. Lastly, Spectral Clustering, a graph theory-based

clustering technique, was also implemented for the data as it does not make inherent assumptions regarding the shape and the distribution of the clusters, unlike representative-based approaches that require clusters to be spherical and convex in form. Connectivity-based approaches like Spectral Clustering first attempt to find points that are immediately next to each other, then map these points into a lower-dimensional space by taking the eigenvectors of the entire dataset before performing the clustering process on the matrix of eigenvectors extracted.

The content of the *'Findings'* column is the basis of the features for the Water Quality Prediction Model. We first performed text normalization, which involves removing articles, determiners, or stop words that are commonly used in the English language and may convey undue signal that such words are important if otherwise retained. We also removed the terms *'discharge pipe/s'* and *'dp'* as tokens, as this is a primary requirement for water sample collection. After performing text normalization, we converted the text data into vectorized form to be able to represent the text in a numeric format. Different vectorization techniques capture a higher quality of information for each machine learning model. One of the basic transformations of text into vectors is the Bag-of-Words (BoW) methodology by counting the frequency of each token for each sample. The frequency of each word for each sample is computed using the CountVectorizer class available in sklearn. We also used the Term Frequency-Inverse Document Frequency (TF-IDF) methodology, which deemphasizes popular words and increases the relevance of tokens that appear less frequently. This is performed using the TfidfVectorizer class available in sklearn. Another vectorization technique used was the Word2Vec methodology, developed by researchers from Google, using the gensim library. The Word2Vec algorithm generates embeddings: representation of the relationships of words which are represented into vectors of

numbers that is derived from the series of texts inputted. Next, we implemented GloVe or Global Vectors for Word Representation, developed by researchers from Stanford, which learns word embeddings by capturing probabilities that a pair of words occur together. GloVe was implemented using the spacy library. Lastly, FastText through the gensim library was implemented to obtain word embeddings as well. An additional feature of FastText, introduced through research from Facebook, focuses on looking at the characters themselves rather than the words only.

Before fitting the vectorized data into a machine learning model, we needed to address the class imbalance problem that was prominent in the data to be modeled. Given that there were only 10% of the sites that were identified for water sampling by the MMDA-SWMO, it was necessary to perform resampling techniques to provide more records of the minority class. By doing this, we minimized the learning bias exhibited by our implemented models towards the majority class. In this work, we decided to use the Adaptive Synthetic Sampling Method (ADASYN) implementation of the imblearn library. It is an oversampling technique that employs a weighted distribution for different minority class examples based on their level of learning difficulty. ADASYN can improve learning with respect to the data distributions by reducing the bias introduced by the class imbalance and shifting the classification decision boundary toward the difficult examples. In this project, the team regarded ADASYN to be superior to Synthetic Minority Oversampling Technique (SMOTE), a similar oversampling technique, because the former not only reduces the bias introduced by the class imbalance, but also adaptively shifts the classification decision boundary toward the difficult examples, which aids in improving our model performance (He, 2008).

For the first data product, the Water Quality Prediction Model, we trained a number of machine learning models to identify the differentiating characteristics of identified water sampling sites from those that were not. The site descriptions in the *'Findings'* column in Figure 6 is the source of features. We looked into different linear models such as the Logistic Regression model and the Stochastic Gradient Descent classifier. Both models were regularized to avoid overfitting on the training set. With regularization, a penalty is added on the loss functions that are optimized by these machine learning algorithms. We applied two different regularization techniques: L1 and L2 regularization. L1 regularization or Lasso regression tends to add a penalty term that consists of the sum of the absolute values of all the weights for each feature. On the other hand, L2 regularization or Ridge regression adds a penalty which consists of the sum of the squared weights for each feature. For the linear models, we also tuned the strength of regularization hyperparameter *C*. Higher values of *C* denote a lower degree of regularization applied when training the models.

Likewise, we also trained linear and non-linear SVM classifiers using the data provided by the MMDA-SWMO. With support vector machines, we aim to find a separating hyperplane that maximizes the size of the boundary between the two classes that we are modeling. For the non-linear SVM, we utilized a sigmoid kernel in defining the shape of the hyperplane used in classification. Similar to the linear classifiers, we also tuned the hyperparameter *C* for these models, which denotes the strength of model regularization.

In addition to this, we trained a Multinomial Naive-Bayes Classifier for the data as this is known to perform well for text classification due to the characteristic of Naive-Bayes models to assume conditional independence of the estimated model coefficients (Kononenko, et.al., 2007). This implies that the Naive-Bayes model

generates a higher bias on its model weights, but at the benefit of having lower variance compared to simple linear classifiers. For cases where there are negative valued features from the data, which was observed when neural-network based embeddings such as Word2Vec, GloVe and FastText were used, we used the Gaussian Naive-Bayes Classifier instead to model the text data.

Aside from linear models, we also explored using tree-based models such as the Decision Tree Classifier, which attempts to infer decision rules that make distinctions on the target variable using the information learned from the features fed into the model. Since tree-based models are generally prone to overfitting on the training set, ensemble models were tested in order to make models that are generalizable to unseen information. Two approaches in ensembling machine learning algorithms would include bagging and boosting. With bagging, we attempt to reduce the variance of our machine learning models by training multiple homogenous weak learners on random samples (generated with replacement) of our original dataset in parallel with each other and then aggregate the model's predictions through averaging. An example of a bagging ensemble model that was used in this study is the Random Forest Classifier, which uses multiple weak decision trees as its base model and trains these decision trees to sub-samples of the data to generate predictions. To compare the performance of the Random Forest Classifier, we also trained a Bagging Classifier that uses the Stochastic Gradient Descent linear classifier as its base learner. In contrast to bagging, boosting algorithms are used to fit models in an iterative manner such that the weights of observations are adjusted on the basis of the quality of the predictions made in the prior iteration. In this study, the boosting algorithm used was the Gradient Boosting Method, which iteratively updates the weights of different decision trees based on the residuals or errors of the model when classifying each observation into

their respective class. For both the tree-based and ensemble algorithms that were implemented, we tuned the maximum depth of the decision tree that was used in model training along with the maximum number of features that would be considered as the machine learning model looked for the best subsample of the data that would help in learning its inherent patterns.

Summarized in Table 6 are the list of machine learning models used to classify the text data, along with the hyperparameters that were tuned for each model experimentation. Given that there were only less than three thousand inspected locations that were recorded by the MMDA-SWMO from 2019-2021, we have decided not to utilize deep learning models as these algorithms require more data in order to capture complex patterns. As 30% of the two thousand locations would be used for model evaluation, the use of deep learning models for text classification may lead to poor performance.

Table 6: Models Implemented and Hyperparameters Tuned

Machine Learning Models	**Hyperparameter Values**
Naive-Bayes Classifier	alpha (smoothing parameter): [0, 0.25, 0.5, 0.75, 1]
Logistic Regression (Lasso) Logistic Regression (Ridge) Stochastic Gradient Descent Classifier (Lasso & Ridge) Linear Support Vector Machines Non-Linear Support Vector Machines	C (regularization strength): [0.0001, 0.01, 1, 10, 100]
Decision Tree Classifier Extra Trees Classifier Adaboost Classifier Bagging Classifier Random Forest Classifier Gradient Boosting Classifier	max_depth: [3, 5, 7, 9, 11, 13, 15, 17, 19] max_features: [0.6, 0.7, 0.8]

Given that we used thirteen different machine learning models which were fed with five different types of vector representations generated from the text data, we were able to implement 65 different machine learning models in total.

The machine learning models are evaluated using the test set accuracy and F1-Score, as discussed in the previous section. The F1-Score was prioritized because it treats false positives and false negatives equally. A higher number of False Positives or misclassifying water sampling sites could lead to additional costs for chemicals used in laboratory testing. Although this expense is outside of the MMDA-SWMO's budget because laboratory testing falls under the mandates of DENR and LLDA, the cost will still be deducted from the Manila Bay Rehabilitation Project budget. Unfortunately, the team has no information regarding the cost or limit of laboratory testing that DENR and LLDA can afford. Similarly, a higher number of False Negatives or misclassifying non-water sampling sites poses an environmental risk that the team cannot quantify. Failure to classify water sampling sites may result in the establishments' continued non-compliance with water quality regulations, further harming Manila Bay. This may delay the Manila Bay Rehabilitation Project and may cause health issues for Metro Manila residents. With this, we utilized the F1-Score to give equal weights to cost, environmental, and health implications.

After performing hyperparameter tuning on the validation set and model performance evaluation on the test set, we explored the use of global and local model interpretability techniques such as the Shapley Additive Explanations (SHAP) and the Local Interpretable Model-Agnostic Explanations (LIME) method. With SHAP, we were able to find the features that either positively or negatively contribute to a prediction of the model. This model interpretability technique is based on coalitional game theory where each feature acts as a player with a respective contribution to the

prediction (Zlaoui, 2021). On the other hand, the LimeTextExplainer class from the LIME library enables us to understand which tokens were relevant to the predicted output for local samples of the model. LIME achieves this by generating synthetic dataset similar to the observation of interest and testing it with surrogate models (Molnar, 2022).

Table 7: Distance Metrics Used to Identify the Top Recommended Sites, where **x** and **y** are N-dimensional vectors

Distance Metric	**Equation**
Euclidean	$\sqrt{\sum_i (x_i - y_i)^2}$
Cosine	$1 - \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$
Cityblock or Manhattan	$\sum_i \lvert x_i - y_i \rvert$
Chebyshev	$\max_i(\lvert x_i - y_i \rvert)$
Hamming	$\sum_i \frac{\delta_i(x,y)}{N}$ where $\delta_i(x,y) = \begin{cases} 1 \text{ if } x_i \neq y_i \\ 0 \text{ if } x_i = y_i \end{cases}$
Jaccard	$\sum_i^{N'} \frac{\delta_i(x,y)}{N'}$ where $\delta_i(x,y) = \begin{cases} 1 \text{ if } x_i \neq y_i \\ 0 \text{ if } x_i = y_i \end{cases}$ and N' is the number of elements where both $x_i \neq 0$ and $y_i \neq 0$

Once we were able to identify the keywords that distinguished between the identified water sampling and those that were not, we used these keywords to determine possible relevant locations from the set of not identified water sampling sites, as these areas tend to have similar text descriptions to those identified sampling

sites. This enabled us to create a recommender system based on information retrieval that recommended sites with geographical properties similar to those identified water sampling sites. To find the recommender system that generated the most number of relevant observations, we iterated over different distance metrics and determined which metric would yield the highest precision. Seen in Table 7 are the different distance metrics that were used in building the information retrieval based recommender system.

IV. Discussion of Results

IV.A. Data Product 1: Findings

For each site inspected, the MMDA-SWMO puts a qualitative description that outlines their observations on a particular area. These are stored in the *'Findings'* column of the checklists provided to us. As we intended to use this information for machine learning modeling, we first needed to explore the different themes observed from the descriptions of each site inspected. To be able to do this, we performed a clustering analysis on the text data. We first tokenized the text using the nltk library's RegexpTokenizer, which splits a string with a regular expression. Following that, we removed stop words or terms that appear frequently in the English language and then lemmatized each word in the corpus using spacy to remove inflections in each term while still returning a word from the English dictionary. We made certain that the terms *'discharge pipe/s'* and *'dp'* were removed as tokens, as this is a primary requirement for water sample collection.

IV.A.1. Clustering Analysis of the Descriptions of the Inspected Sites

After text normalization of the descriptions of the inspected sites, we then vectorized the data into numeric format using the TF-IDF method, which assigns higher scores to terms with lower document frequencies as these are domain-specific words that can contribute additional relevance to the study performed. Doing this prepared the text data to be fed into different clustering algorithms such as k-Means Clustering, Agglomerative Clustering (using Ward's method), Spectral Clustering, and Gaussian Mixture Models.

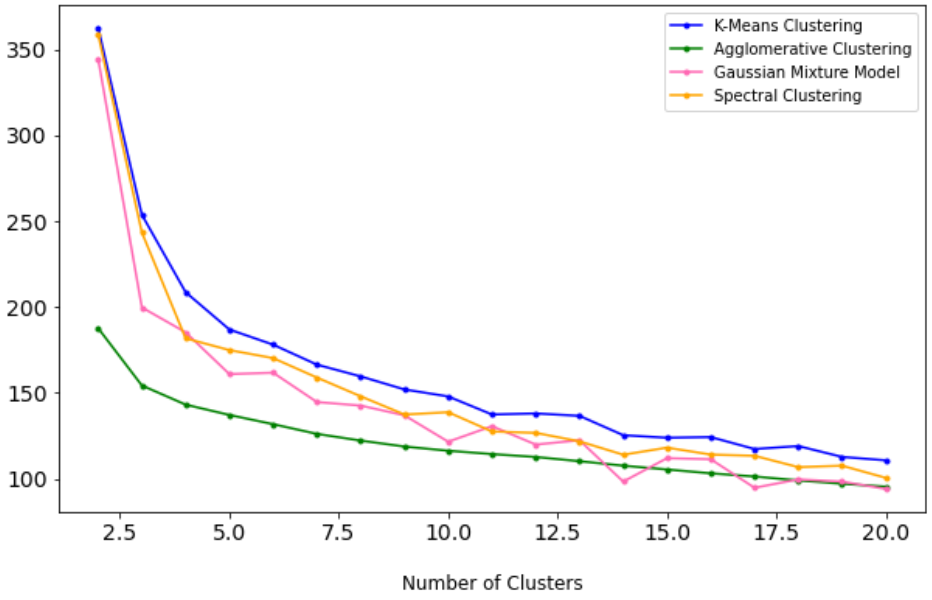


Figure 20. Calinski-Harabasz Scores per Clustering Method. The k-Means Clustering model with two clusters yielded the highest Calinski-Harabasz score.

The number of clusters generated by each algorithm was a hyperparameter that was tuned, and the best model is the one with the highest Calinski-Harabasz (CH) score. The CH metric computes the ratio of inter-cluster (between cluster) dispersion to intra-cluster (within cluster) dispersion. The higher the CH score, the better the performance as this implies that the generated clusters are dense and well separated. Figure 20 confirms this and demonstrates that the optimal clustering of the text descriptions was observed when the number of clusters was set to two for all of the models used. Across all of the clustering algorithms evaluated, the k-Means Clustering model produced the highest Calinski-Harabasz score. This indicated that the clusters created by this algorithm were highly separable from one another, but the points within the same cluster were quite close together.

Cluster 1 *Cluster 2*

Figure 21. Word Cloud of Site Descriptions: Cluster 1 pertains to the physical state of inspected drainage sites, while Cluster 2 pertains to the MMDA-FCSMO's cleaning operations.

From the visualized word clouds of the clusters, displayed in Figure 21, we identified inspection-related descriptors as part of the first cluster, which contained terms associated with the actual physical state of the drainage sites inspected by the MMDA-SWMO. Among the descriptors in this cluster, the words with the high weights are *'garbage', 'wastewater', 'turbid', 'contain', 'release',* and *'sludge'*, which allude to the characteristics of an inspected drainage system that *contains garbage* and *sludge*, with the presence of a discharge pipe that *releases turbid wastewater.*

On the other hand, the descriptions observed in the second cluster were more related towards site interventions undertaken by the MMDA-FCSMO. The descriptors that were part of this cluster were related to the operations done by the agency to fulfill their drainage cleaning mandate. The terms with highest weights were *'fcsmo', 'remove', 'sack', 'cleaning',* and *'complete*'. The MMDA-SWMO field team also records the completion progress of the cleaning process and the number of sacks removed from the drainage by the MMDA-FCSMO.

We further segmented the clustering results performed above by dissecting them based on whether the samples were identified sampling sites or not. Figure 22 below shows how the identified sampling sites were clustered on the basis of their text descriptions.

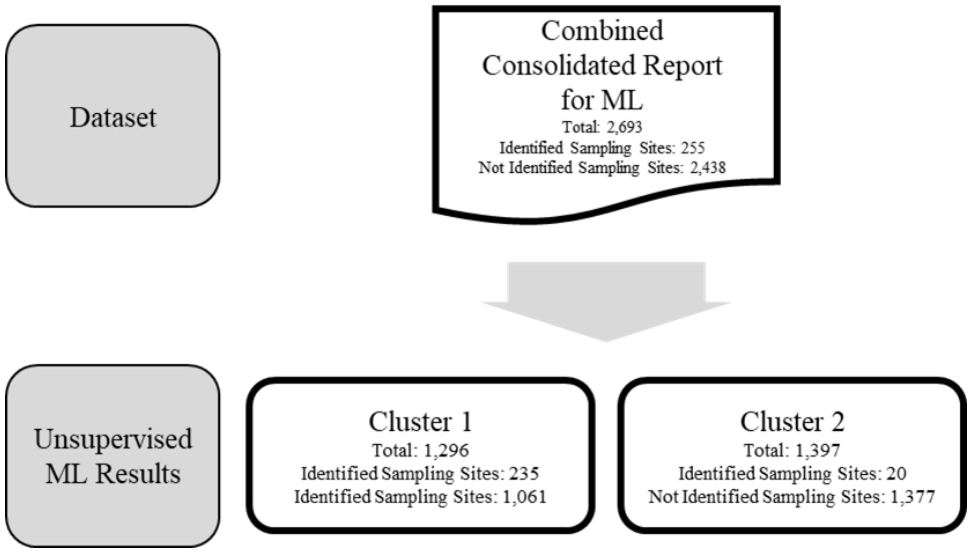


Figure 22. Clustering Results with Count of Sampling Site Identification. The majority of the identified sampling sites are present in Cluster 1.

We can deduce from the clustering results that most of the identified water sampling sites (235 out of 255) were present in the cluster that contained the observed physical state of the inspected drainage sites. On the other hand, the second cluster only contained a few identified water sampling sites (20 out of 255). It is important to note that the clustering performed above was only based on the *'Findings'* column and whether the samples were identified water sampling sites or not; on whether the sites actually underwent water sample collection was not considered.

The next step entails training a supervised machine learning model with a target variable to predict whether or not a site is a water sampling site, using the text descriptions in the *'Findings'* column as features.

IV.A.2. Machine Learning Models to Identify Water Sampling Sites

We were able to develop thirteen different machine learning models that distinguished between identified water sampling sites and those that were not. To do this, we split the data into training, validation, and test sets. Each machine learning

model was fed with the training data, and the hyperparameters of each model were optimized using the validation set. To find the best performing model, the accuracy is optimized on the test set using various hyperparameters.

To reiterate, the machine learning models were also evaluated using the F1-Score, as discussed in the previous section. We prioritized the F1-Score because it treats false positives and false negatives equally. A higher number of False Positives or misclassifying water sampling sites could lead to additional costs for chemicals used in laboratory testing. Although this expense is outside of MMDA-SWMO's budget because laboratory testing falls under the mandates of DENR and LLDA, the cost will still be deducted from the Manila Bay Rehabilitation Project budget. Unfortunately, the team has no information regarding the cost or limit of laboratory testing that DENR and LLDA can afford. Similarly, a higher number of False Negatives or misclassifying non-water sampling sites poses an environmental risk that the team cannot quantify. Failure to classify water sampling sites may result in the establishments' continued non-compliance with water quality regulations, further harming Manila Bay. This may delay the Manila Bay Rehabilitation Project and may cause health issues for Metro Manila residents. With this, the team utilized the F1-Score to give equal weights to cost, environmental, and health implications.

The performance of the thirteen models on the test set using Bag-of-Words vector representation, TF-IDF vector representation, Word2Vec embeddings, GloVe embeddings, and FastText embeddings are found in Tables 8 through 17.

Table 8. Model Results Using Bag-of-Words Vector Representation

Model	Test Accuracy	Test F1-Score	Optimal Hyperparameter	Training Time (s)
Multinomial Naive-Bayes	77.50%	41.84%	alpha = 0.25	0.13
Linear SVC	85.82%	49.60%	C = 10	1.13
Non-linear SVC	72.74%	30.75%	C = 10	0.51
SGD Classifier - L1	72.82%	30.94%	alpha = 0.0001	0.15
SGD Classifier - L2	72.93%	31.16%	alpha = 10	0.13
Logistic Regression - L1	72.52%	30.58%	C = 100	8.20
Logistic Regression - L2	72.70%	30.59%	C = 10	0.33
Decision Tree Classifier	92.35%	70.49%	max_depth = 17 max_features = 0.6	0.15
Extra Trees Classifier	91.60%	68.70%	max_depth = 19 max_features = 0.6	3.17
Adaboost Classifier	78.83%	41.96%	learning_rate = 1	0.19
Bagging Classifier	90.27%	65.34%	max_features = 0.6	0.53
Random Forest Classifier	92.28%	70.37%	max_depth = 19 max_features = 0.7	4.39
Gradient Boosting Method	94.80%	77.92%	max_depth = 11 max_features = 0.6	2.39

Using the Bag-of-Words vector representation, the Gradient Boosting Method performed the best on the test set in terms of accuracy and F1-Score, with 94.80 percent and 77.92 percent, respectively.

Table 9. Model Results Using TF-IDF Vector Representation

Model	Test Accuracy	Test F1-Score	Optimal Hyperparameter	Training Time (s)
Multinomial Naive-Bayes	81.77%	48.04%	alpha = 0.25	0.13
Linear SVC	86.78%	56.05%	C = 10	0.22
Non-linear SVC	75.57%	38.50%	C = 1	0.80
SGD Classifier - L1	74.86%	37.83%	alpha = 0.01	0.14
SGD Classifier - L2	73.90%	37.06%	alpha = 0.0001	0.13
Logistic Regression - L1	75.42%	37.55%	C = 100	20.03
Logistic Regression - L2	75.04%	37.89%	C = 100	0.43
Decision Tree Classifier	94.88%	75.70%	max_depth = 13 max_features = 0.7	0.22
Extra Trees Classifier	96.14%	81.43%	max_depth = 17 max_features = 0.6	7.07
Adaboost Classifier	83.22%	50.22%	learning_rate = 1	0.89
Bagging Classifier	98.89%	94.02%	max_features = 0.3	1.25
Random Forest Classifier	96.81%	84.48%	max_depth = 13 max_features = 0.6	18.96
Gradient Boosting Method	98.96%	94.47%	max_depth = 17 max_features = 0.6	9.31

When the models were trained using the TF-IDF vector representation, the Gradient Boosting Method also outperformed the others on the test set in terms of accuracy and F1-Score, achieving 98.96% and 94.47% respectively. These results were higher than those obtained using the Bag-of-Words vector representation, indicating that utilizing a more robust vector representation improved model performance.

By looking at the confusion matrix of the Gradient Boosting Method on the test set, we observed that the TF-IDF vectorization tends to reduce the number of

misclassified sites from 37 to 9. In particular, the use of TF-IDF for text vectorization was able to increase the number of correctly identified sites for water sampling from 65 to 72.

Table 10. Confusion Matrices of Bag-of-Words and TF-IDF

Bag-of-Words			
		Predicted	
		Not Sampling Site	Sampling Site
Actual	Not Sampling Site	707	25
	Sampling Site	12	65

TF-IDF			
		Predicted	
		Not Sampling Site	Sampling Site
Actual	Not Sampling Site	728	4
	Sampling Site	5	72

Table 11. Model Results Using Word2Vec Embeddings

Model	**Test Accuracy**	**Test F1-Score**	**Optimal Hyperparameter**	**Training Time (s)**
Gaussian Naive-Bayes	99.38%	96.86%	var_smoothing = 1e-11	0.24
Linear SVC	88.00%	59.41%	C = 100	13.90
Non-linear SVC	85.89%	47.71%	C = 0.01	1.49
SGD Classifier - L1	86.01%	47.93%	alpha = 0.01	0.17
SGD Classifier - L2	86.01%	48.87%	alpha = 0.01	0.15
Logistic Regression - L1	85.64%	46.30%	C = 0.01	0.36
Logistic Regression - L2	85.89%	48.18%	C = 10	0.60
Decision Tree Classifier	98.51%	92.59%	max_depth = 19 max_features = 0.6	0.57
Extra Trees Classifier	98.89%	94.41%	max_depth = 19 max_features = 0.8	6.53
Adaboost Classifier	82.55%	47.58%	learning_rate = 1e-6	2.45
Bagging Classifier	99.38%	96.82%	max_features = 0.4	4.56

Table 11. Model Results Using Word2Vec Embeddings (*continuation*)

Model	Test Accuracy	Test F1-Score	Optimal Hyperparameter	Training Time (s)
Random Forest Classifier	98.51%	92.50%	max_depth = 19 max_features = 0.8	63.05
Gradient Boosting Method	99.51%	97.40%	max_depth = 19 max_features = 0.6	38.65

Using Word2Vec embeddings to train the models resulted in a further increase in performance. The Gradient Boosting Method was again the best performing model, with 99.51% test accuracy and 97.40% F1-Score. Despite comparable test accuracies, the GBM model using Wor2Vec embeddings produces a considerable boost in F1-Score when compared to the TF-IDF's 94.47% F1-Score. It is worth mentioning that the latter has a 30 second faster runtime.

Table 12. Model Results Using GloVe Embeddings

Model	Test Accuracy	Test F1-Score	Optimal Hyperparameter	Training Time (s)
Gaussian Naive-Bayes	99.13%	95.42%	var_smoothing = 1e-07	0.15
Linear SVC	81.68%	44.78%	C = 10	11.41
Non-linear SVC	83.91%	45.38%	C = 1	1.66
SGD Classifier - L1	83.66%	45.00%	alpha = 0.0001	0.59
SGD Classifier - L2	83.79%	44.73%	alpha = 0.01	0.24
Logistic Regression - L1	83.79%	45.19%	C = 100	97.84
Logistic Regression - L2	82.92%	43.90%	C = 10	0.77
Decision Tree Classifier	97.65%	88.34%	max_depth = 19 max_features = 0.8	1.37
Extra Trees Classifier	98.02%	90.12%	max_depth = 19 max_features = 0.6	11.68
Adaboost Classifier	89.73%	53.63%	learning_rate = 1	0.58

Table 12. Model Results Using GloVe Embeddings (*continuation*)

Model	Test Accuracy	Test F1-Score	Optimal Hyperparameter	Training Time (s)
Bagging Classifier	99.01%	94.59%	max_features = 0.4	8.53
Random Forest Classifier	97.15%	86.39%	max_depth = 19 max_features = 0.6	115.72
Gradient Boosting Method	99.13%	95.42%	max_depth = 11 max_features = 0.8	66.05

Two models achieved the highest test set accuracy by using GloVe embeddings for training. Again, the Gradient Boosting Method is among the best models, achieving the same test accuracy of 99.13% and F1-Score of 95.42% as the Gaussian Naive-Bayes. Interestingly, employing GloVe embeddings did not improve model performance when compared to utilizing Word2Vec embeddings, which resulted in a significantly higher test accuracy and F1-Score for its best model - Gradient Boosting Method.

Table 13. Model Results Using FastText Embeddings

Model	Test Accuracy	Test F1-Score	Optimal Hyperparameter	Training Time (s)
Gaussian Naive-Bayes	99.38%	96.82%	var_smoothing = 1e-11	0.21
Linear SVC	87.62%	45.65%	C = 0.0001	0.13
Non-linear SVC	89.85%	12.77%	C = 0.01	2.15
SGD Classifier - L1	90.01%	13.04%	alpha = 0.0001	0.25
SGD Classifier - L2	90.01%	13.04%	alpha = 0.01	0.15
Logistic Regression - L1	90.01%	13.04%	C = 1	7.72
Logistic Regression - L2	90.01%	13.04%	C = 1	0.22
Decision Tree Classifier	98.02%	90.12%	max_depth = 17 max_features = 0.7	0.60

Table 13. Model Results Using FastText Embeddings *(continuation)*

Model	Test Accuracy	Test F1-Score	Optimal Hyperparameter	Training Time (s)
Extra Trees Classifier	98.89%	94.41%	max_depth = 17 max_features = 0.7	8.32
Adaboost Classifier	84.53%	46.35%	learning_rate = 1e-06	4.26
Bagging Classifier	99.38%	96.86%	max_features = 0.6	7.70
Random Forest Classifier	98.64%	93.33%	max_depth = 17 max_features = 0.8	93.88
Gradient Boosting Method	99.38%	96.82%	max_depth = 13 max_features = 0.7	41.41

Lastly, using FastText embeddings to train the models showed that three methods (Gradient Boosting Method, Gaussian Naive-Bayes, and Bagging Classifier) achieved the highest accuracy score of 99.38%. In terms of F1-Score, the Bagging Classifier performed best, attaining 96.86%, with the former two models close behind at 96.82%.

Table 14. Confusion Matrices of Word2Vec, FastText and GloVe

Word2Vec			
		Predicted	
		Not Sampling Site	Sampling Site
Actual	Not Sampling Site	731	2
	Sampling Site	2	75

FastText			
		Predicted	
		Not Sampling Site	Sampling Site
Actual	Not Sampling Site	729	4
	Sampling Site	1	76

Table 14. Confusion Matrices of Word2Vec, FastText and GloVe *(continuation)*

GloVe			
		Predicted	
		Not Sampling Site	Sampling Site
Actual	Not Sampling Site	730	3
	Sampling Site	4	73

Although the best models using FastText embeddings outperformed those using GloVe embeddings and TF-IDF features in terms of test accuracy and F1-Score, they did not outperform the Gradient Boosting Method model using Word2Vec Embeddings, which achieved the highest test accuracy of 99.51% and F1-Score of 97.40%.

Comparing the confusion matrices of the Gradient Boosting Method on the test set using the features generated by the neural-network based embeddings, we can see that the 97.4% F1-Score generated by Word2Vec translates to only four misclassified sites relative to five and seven sites when the FastText and GloVe embeddings were used respectively. Although the FastText embedding tends to classify correctly most of the sites that are intended for water quality sampling, it tends to be less precise in identifying sites that are not for sampling activities. Hence, using the FastText embedding relative to Word2Vec could lead to additional costs in water quality sampling even though it is not actually needed for the inspection site. In general, the neural network generated word embeddings tend to outperform simple vectorization techniques such as Bag-of-Words and TF-IDF by netting higher performance metrics on the test set. However, utilizing these embeddings for feature generation requires a

significant amount of training time and identifying the keywords that contribute to the predictions of the model is difficult since each feature generated through word embeddings are not attributable to a single token only.

With this, the team has selected the Gradient Boosting Method trained on features generated through TF-IDF vectorization as the most appropriate predictive model for the MMDA-SWMO's use case given all the models that were implemented in this work primarily due to its excellent performance, interpretability, and shortest training time. The model achieved an F1-Score of 94.47% and a test accuracy of 98.96%, both of which is higher than the proportional chance criterion of 82% for the dataset. Using TF-IDF vectorization also makes the model readily interpretable through methods like SHAP, but neural network-based embedding techniques do not. In addition, when utilizing TF-IDF vectorization, the Gradient Boosting Method has the fastest run time and can be trained on the data in 9.31 seconds, as opposed to 38.65 seconds for Word2Vec.

IV.A.3. Model Interpretability

The Gradient Boosting Method using the features from the TF-IDF vectorization is regarded as the best model due to its interpretability and excellent performance. The SHAP method was applied to this model to generate charts that aid us in understanding how the model arrived at its predictions. Figure 23 shows the absolute value of the Shapley values for each token, and it can be seen that the tokens with the highest values are *'level', 'release', 'turbid', 'murky', 'complete',* and *'inspection'*.

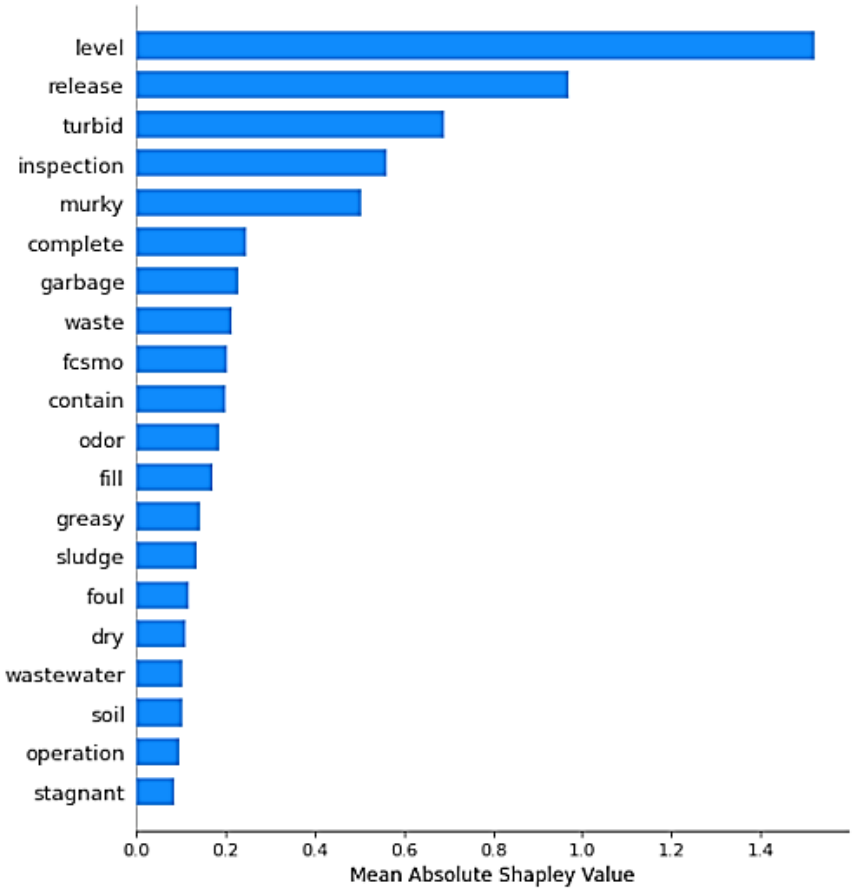


Figure 23: Top 20 Tokens with Highest Shapley Values (in absolute value). The tokens with top absolute Shapley Values include *'level', 'release', 'turbid', 'inspection',* and *'murky'*.

In addition, a Beeswarm plot can also be generated through SHAP for a more in-depth analysis on how each feature positively or negatively contributed to the model's prediction. We can see in Figure 24 that high TF-IDF values for the token *'release'* contribute to the model predicting that a given site should be identified for water sampling. This is consistent with how the MMDA-SWMO conducts its water sampling operation, in which sampling a site is reliant on the presence of flowing wastewater from the discharge pipe, among other things. Other notable tokens whose presence in the text positively contributes towards a prediction of an inspected site being qualified for water sampling are *'inspection', 'turbid', 'murky', 'foul'*, and *'greasy'*. The *'odor'* token with a comparable contribution as the *'foul'* token was

removed in subsequent implementations, i.e., Geospatial Recommender System, to avoid redundancy as the former always exists in the context of *'foul'*. In contrast, the presence of the word *'level'* and *'stagnant'* indicate that the site is not required to be water sampled. Upon further inspection, the word *'level'* forms part of the phrase *'level of cleaning'* and is succeeded by the percent completion progress of cleaning by the MMDA-FCSMO as of inspection date, while *'stagnant'* refers to the wastewater that is trapped within the drainage. We can deduce from the inspected Beeswarm plot that a site with an inspected discharge pipe releasing wastewater that is described as turbid or murky, greasy, and has a foul odor in the *'Findings'* column may potentially have water quality below standards, thus, should be sampled and forwarded to laboratory testing.

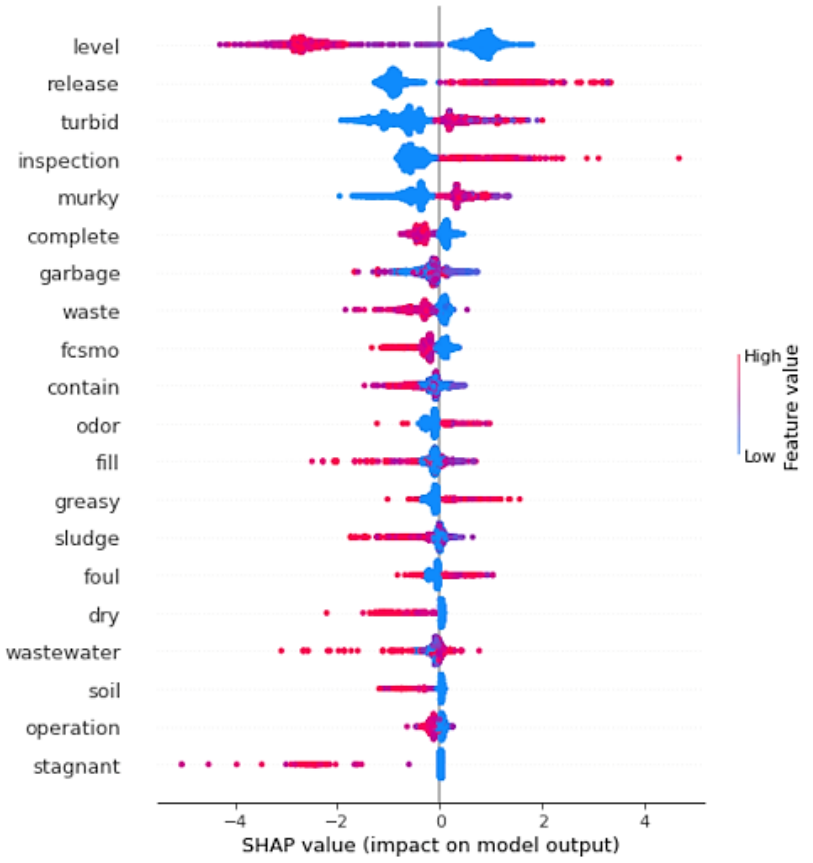


Figure 24: The SHAP Beeswarm Plot of the Top 20 Tokens reveals the top features that positively (i.e., *'release', 'turbid', 'inspection'*, etc.) and negatively contribute (i.e., *'level', 'stagnant'*) to the model's prediction.

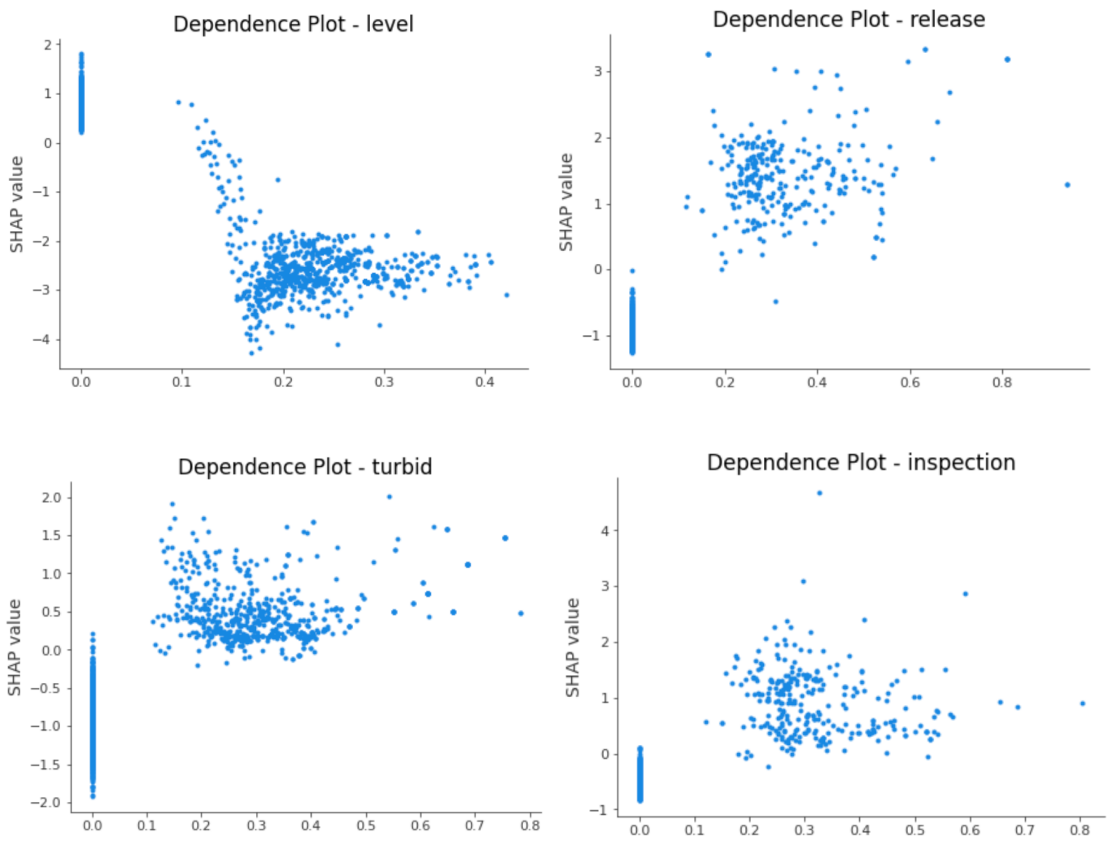


Figure 25. SHAP Dependence Plots of the Top Four Keywords. There is a positive correlation between the TF-IDF values and SHAP values of the words *'release'*, *'turbid'*, and *'inspection'*, while there is a negative correlation for the word *'level'*.

IV.A.4. Correlation of the Top Features to the 'dp' and 'discharge pipe' Tokens

To further interpret the results of our machine learning model, we examined the correlation of the top features identified through SHAP's Beeswarm Plot to the primary requirement of a water sampling site: presence of *'dp'* or *'discharge pipe'*, which were removed from the tokenization process of the text to avoid putting a variable or feature in our model that could perfectly predict the target. As discussed in the previous section, the tokens *'level'* and *'stagnant'* negatively contributed to the model's prediction, whereas the tokens *'inspection'*, *'turbid'*, *'murky'*, *'foul'*, *'odor'*, and *'greasy'* positively contributed to the model's prediction.

Table 15. Correlation of the Top Features from SHAP's Beeswarm Plot to the *'dp'* and *'discharge pipe'* Tokens

	'dp'	**'discharge pipe'**
level	-0.23	-0.03
stagnant	-0.02	-0.02
release	0.84	0.17
inspection	0.79	0.15
murky	-0.04	-0.04
turbid	-0.05	0.00
greasy	0.01	0.00
foul	-0.05	0.03
odor	-0.04	0.03

Table 15 shows that among the top tokens, only *'release'* and *'inspection'* are correlated to the *'dp'* token, while no correlation was observed for the token *'discharge pipe'*. The correlation of *'release'* and *'inspection'* to *'dp'* can be attributed to the wastewater release, or lack thereof, from the discharge pipe being inspected by the field team during the Site Identification task. The majority of the top features are not correlated to either of the tokens pertaining to discharge pipe, implying that the identified characteristics for water sampling site eligibility are not known indicators based on MMDA-SWMO's existing procedures. With the top keywords generated by the machine learning model, we provided the field team a new set of general guidelines in determining whether a site should be water sampled and endorsed for laboratory testing or not, with no disruptions to their current practice of visual inspection and recording textual descriptions.

IV.A.5. LIME Results: MMDA-SWMO Identified Water Sampling Site

Local interpretability can also be performed through LIME. Figure 26 shows a single sample record of a water sampling site previously evaluated by the MMDA-SWMO. The example shown is a drainage manhole with a discharge pipe located in front of Jollibee N. Domingo corner A. Luna Street, Brgy. Balong Bato, San Juan City. Through LIME, we can interpret that the GBM model identifies the tokens *'release'*, *'emit'*, *'connect', 'greasy', 'wastewater', and 'odor'* as indicators that the site is qualified for water sampling. In contrast, the word *'inspection'* does not contribute to the site being classified as such.

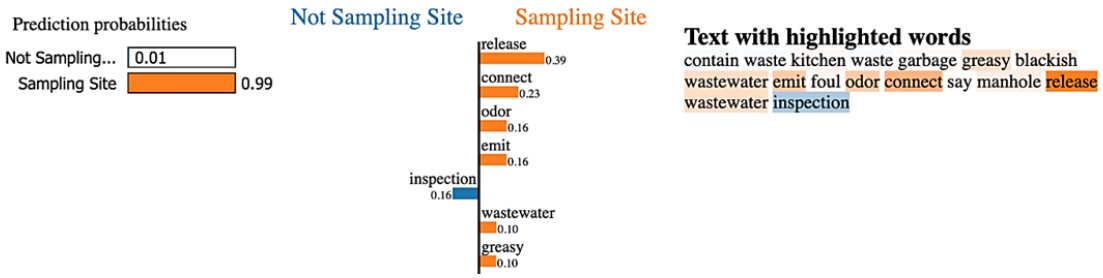


Figure 26. LIME Text Explainer Results for a Sample Record of Identified Water Sampling Site. The tokens *'release'*, *'connect', 'odor'* are some indicators that the site is qualified for water sampling.

IV.A.6. LIME Results: MMDA-SWMO Not Identified Water Sampling Site

For the other label, Figure 27 shows a sample record of a site that the MMDA-SWMO did not consider for water sample collection. The example shown is a catch basin with no discharge pipe located on the left side of No. 1026, DJM Building, Belen Street, Paco, Manila. LIME identifies the tokens *'level', 'scrap', 'wood', 'odor',* and *'foul'* as indicators that the site is not required for water sampling. The token *'level'* is consistent with the results of SHAP, whereas the tokens *'greasy', 'turbid',* and *'leak'* normally contribute to a water sampling site but are not sufficient in this case.

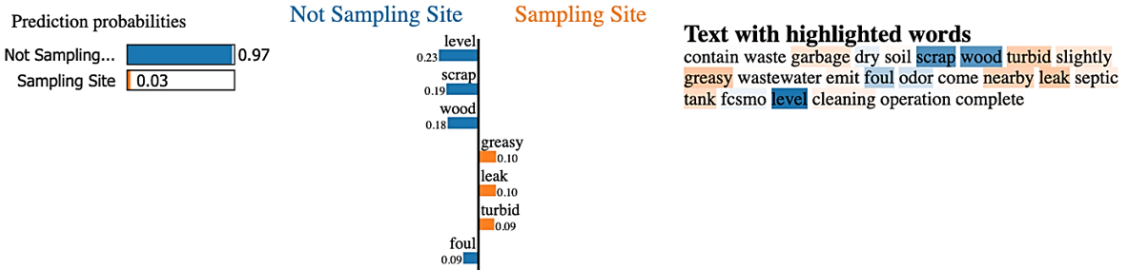


Figure 27. LIME Text Explainer Results for a Sample Record of Not Identified Water Sampling Site. The tokens *'level'*, *'scrap'*, *'wood'* are some indicators that the site is not required for water sampling.

IV.B. Data Product 2: Findings

IV.B.1. Recommendations for Site Reinspection

To be able to predict areas which may potentially be non-compliant to water quality standards mandated by the DENR, we developed an information-retrieval based recommender system. In building this data product, we aim to identify areas that were not identified as water sampling sites, but have similar geospatial properties with those identified water sampling sites. In order to perform this, we extracted the number of amenities for all locations with longitude and latitude data, either provided by MMDA-SWMO or extracted using Geocode. For inspected sites with geo-coordinates, we extracted the number of amenities and land use classification within 150 meters using OSM. These geospatial features and their respective frequencies are then treated as feature vectors for each location. We computed the distances of the feature vectors of each non-water sampling site against each water sampling site, using various distance metrics identified in Table 7 and computed the average. We next sort by the smallest average distance for each distance metric. The locations with the smallest distance would be the most similar to the locations of water sampling sites based on the geospatial features available from OSM.

For each distance metric, we chose to retrieve the top 100 locations with smallest average distances as the list of recommendations. The reasonableness of these recommended locations are then evaluated using the precision scoring metric. We measure precision for this information retrieval task if the keywords arising from the SHAP results in Figure 24 are present in the *'Findings'* column that contains the written descriptions of the inspected site. Specifically, a non-water sampling site is identified to be relevant if the *'Findings'* contain any of the terms *'release', 'turbid', 'murky', 'inspection', 'foul',* and *'greasy'*. These words are the top six tokens identified through SHAP that contribute to the prediction of a site as a water sampling site. Having a relevance condition allows us to evaluate the usefulness of the recommendations towards the agency's objectives.

Table 16. Precision Scores of the Recommender System for Site Reinspection (k=100)

Distance Metric	**Precision** (no normalization)	**Precision** (with normalization)
Euclidean	73%	70%
Cosine	71%	71%
Cityblock	75%	68%
Chebyshev	67%	69%
Hamming	82%	77%
Jaccard	73%	73%

Table 16 shows that using the Hamming distance to generate recommendations yielded the most number of relevant observations. Out of the top 100 sites recommended by the Hamming distance, 81 of them contain at least one of the keywords mentioned above in their site descriptions. The recommender system also performs better relative to revisiting each site in a random order, as there are 1,702 out

of 2,335 sites or 72.6% from the set of sites that were not identified as water sampling sites, but contain the SHAP's most important keywords to be identified as relevant.

Figure 28. Recommended Sites for Reinspection per Distance Metric. The charts highlight that most of the recommended sites are located in Ermita, Malate, and Paco.

As a further step for validation, the vector features are normalized by transforming the frequencies of the surrounding amenities and land use classification into a unit vector to avoid bias for amenities with a significantly high count. The results show that the recommendations using the Hamming distance still provides precision better than the baseline of 72.6%.

Seen in Figure 28 are the exact locations in Manila of all the recommended sites for each distance metric. Regardless of distance metric used, most of the top 100 recommended sites produced by the model are located in Ermita, Malate, and Paco, Manila. The main difference between the recommendations generated by the Hamming distance relative to the other models implemented is that its top 100 recommendations are relatively close to each other. Provided in Appendix 1 is the complete list of the top 100 recommended sites produced by the Hamming distance metric. Aside from Manila City, some of the sites that were recommended for reinspection are located in Pasig City and Quezon City.

IV.B.2. Priority Locations for Site Reinspection

Table 17 displays the top ten critical sites identified using the Hamming distance, which was noted as the best metric to use for the second set of recommended locations. These areas lack discharge pipes but are highly likely to contain dirty wastewater and other wastes. Appendix 1 contains the full list of the top 100 sites from the list of previously inspected locations that we propose prioritizing for reinspection and cleaning once the MMDA-SWMO resumes field operations with the MMDA-FCSMO.

Table 17. Top 10 Recommended Sites for Reinspection

Date of Inspection	Location	Street	District	City	Site Type	Drainage System	Findings	Rank
04 Jul 2019	In front of a vacant lot	Pedro Gil St.	Paco	Manila	Non-Residential	Catch Basin (CB)	Contained few garbage that emits foul odor, rocks, sand, wet soil	9
09 Jul 2019	In front of Betamia General Merchandise	Angel Linao St.	Malate	Manila	Non-Residential	Drainage Manhole (DM)	Contained murky wastewater, few garbage and sludge	5
06 Aug 2019	In front of No. 2219	Singalong St.	Malate	Manila	Non-Residential	Catch Basin (CB)	Contained murky wastewater, few garbage and sludge, dried leaves	1
06 Aug 2019	In front of No. 2181, PSA Bike Shop	Singalong St.	Malate	Manila	Non-Residential	Catch Basin (CB)	Clogged with stagnant turbid wastewater with few garbage and sludge	3
06 Aug 2019	At the side of No. 2181, PSA Bike Shop	Singalong St.	Malate	Manila	Non-Residential	Catch Basin (CB)	Clogged with stagnant turbid wastewater with few garbage and sludge	4
15 Aug 2019	In front of No. 2008	Kasoy St.	San Andres	Manila	Residential	Drainage-Open Canal (OC)	Contained murky wastewater, some garbage and sludge, sand, scrap wood	2
17 Sep 2019	In front of No. 2328-B	Dama de Noche St.	Malate	Manila	Residential	Drainage-Open Canal (OC)	Contained turbid wastewater, some garbage	7
17 Sep 2019	In front of No. 2334	Dama de Noche St.	Malate	Manila	Residential	Drainage-Open Canal (OC)	Filled with murky wastewater, some garbage, sand, stones	8
17 Sep 2019	At the side of No. 2323	Dama de Noche St.	Malate	Manila	Residential	Drainage-Open Canal (OC)	Contained murky wastewater, sand, and stones	10
17 Sep 2019	In front of No. 2329	Dama de Noche St.	Malate	Manila	Residential	Drainage-Open Canal (OC)	Contained turbid wastewater, some garbage and sludge	6

These critical sites for reinspection and re-cleaning are concentrated in Malate, specifically along Dama de Noche and Singalong streets. The list includes an equal number of residential and non-residential locations. The previously inspected drainage systems were split into catch basins and open drainage systems that contained murky and turbid wastewater, as well as various waste such as garbage, sludge, grease, soil, sand, rock, wood, and leaves. It is also worth noting that all ten sites were last inspected nearly three years ago.

IV.B.3. Recommendations for Water Sample Collection

Given that the MMDA-SWMO can only collect water samples from drainage sites with discharge pipes releasing wastewater, not all of the relevant sites based on the SHAP keywords can be endorsed to MMDA-SWMO for water sample collection and to DENR or LLDA for laboratory testing. Hence, we needed to modify the relevance condition in order to make recommendations for sites where water sample collection is feasible. With this, we modified the relevance condition to only include those sites with the term *'dp'* on their site descriptions. With this filtering, there are only 5.7% or 134 relevant locations out of the 2,335 that were identified by

MMDA-SWMO as non-water sampling sites. Likewise, we implemented our information-retrieval based recommender system by iterating over different distance metrics, and then evaluated the quality of the recommendations generated using the precision metric, but taking into account the new relevance condition.

Table 18. Precision Scores of the Recommender System for Water Sample Collection (k=100)

Distance Metric	**Precision** (no normalization)	**Precision** (with normalization)
Euclidean	8%	7%
Cosine	8%	8%
Cityblock	5%	9%
Chebyshev	5%	3%
Hamming	0%	1%
Jaccard	3%	6%

From the table above, we observed that the Euclidean distance metric tends to yield the most number of relevant recommendations as nine out of its top 100 recommendations tend to have discharge pipes within the drainage sites. The recommender system also performs better relative to reinspecting all locations randomly, as there are only 5.7% relevant observations from the set of non-water sampling sites. Aside from the Euclidean distance metric, the recommender system which uses the Cosine distance metric to generate recommendations tends to have a precision that goes beyond the performance of the random recommendation.

Figure 29. Recommended Sites for Water Sample Collection per Distance Metric. Sites are mostly located in barangays located in Malate and Ermita, Manila.

Although the Hamming distance achieved the highest precision when making recommendations for sites that should be reinspected and re-cleaned, as indicated in Table 16, this metric did not yield any relevant recommendations for sites that are intended for water sample collection. Similarly, the precision is computed for the various differences when the feature vectors are normalized, and the results are also shown in Table 18. Using the Euclidean and the Cosine distance still provide stronger performance against the baseline of 5.7%. It is also worth noting that the recommendations using the Cityblock distance improved to 9% when using normalized feature vectors. However, when compared to the baseline, the precision using raw values in the feature vectors and the Cityblock distance result in lower precision. As such, we decided to use the recommendations based on Cosine and Euclidean distance metrics as they display stability in precision computed for both raw and normalized feature vectors.

Shown in Figure 29 are the locations of the relevant sites where water samples could be collected and endorsed to DENR for laboratory testing and are part of the Top 100 recommendations for each distance metric. Most of the recommended sites for water quality sampling are located within specific barangays located in Malate and Ermita, Manila. The models also provided specific locations in Pasig City and Quezon City where MMDA-SWMO can collect water samples from discharge pipes.

IV.B.4. Priority Locations for Water Sample Collection

Based on the MMDA-SWMO's data, there are about 300 previously inspected sites with discharge pipes that were endorsed for sampling but no water sample was collected due to various reasons concerning the wastewater discharge, as shown in

Figure 30. The absence of flowing wastewater during the inspection was the primary reason, and the sites were marked for revisit.

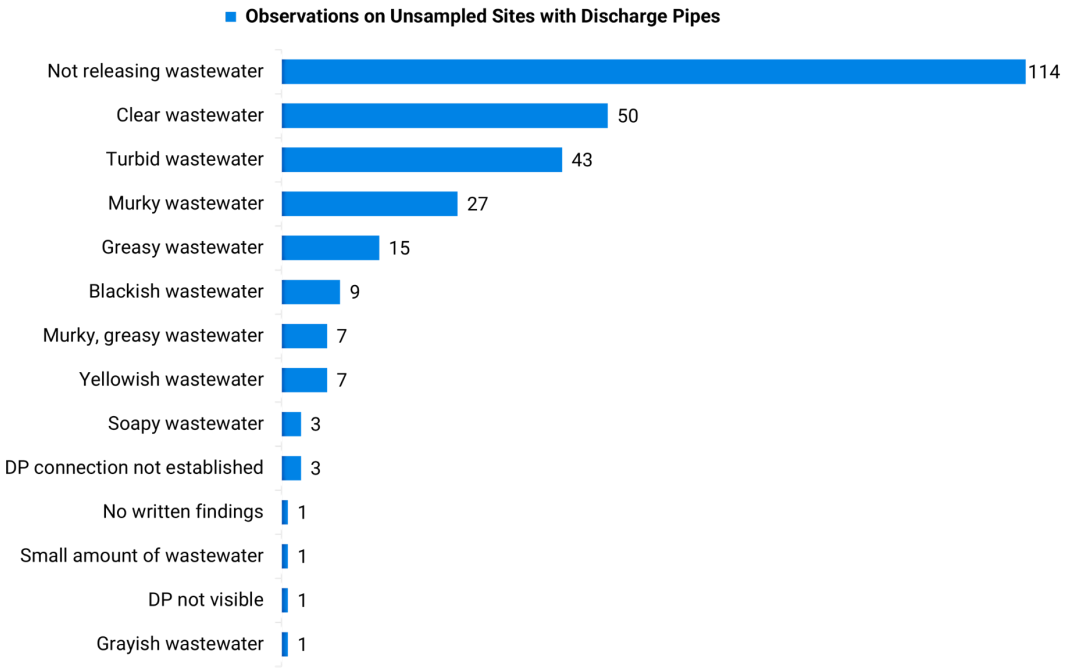


Figure 30. Observations on Unsampled Sites with Discharge Pipes. The lack of flowing wastewater at the time of inspection was the primary reason for not collecting water samples.

As both Euclidean and Cosine Distance metrics produced the highest precision score for previously identified water sampling sites with the presence of discharge pipes, the team deep dived into the top 10 sites from both measures. Table 19 displays the top 10 critical sites determined by the Cosine Distance, whereas Table 20 shows the top 10 determined by the Euclidean Distance. Both top 10 sets have no similar sites, however the top 100 from both metrics contain 69% duplicate water sampling sites, as shown in Appendices 2 and 3. Combining both lists, the team recommends the top 123 sites to be prioritized for water sample collection when the MMDA-SWMO's field operation resumes.

Zooming in on Table 19, The majority of the top 10 sites determined by the Cosine Distance were located along Taft Avenue in Malate. Ermita has the same number of sites in the top 10, although they are distributed among three different thoroughfares. Furthermore, the majority of the sites are non-residential, with drainage systems divided between manholes and catch basins. It is also worth noting that sites were previously not sampled due to the lack of wastewater discharge.

Table 19. Top 10 Recommended Locations for Water Sample Collection (Cosine Distance)

Date of Inspection	Location	Street	District	City	Site Type	Drainage System	Findings	Rank
19 Feb 2019	No. 2172-A, Tuazon Compound, District 5, Zone 78, Brgy. 709 (in front)	Taft Ave.	Malate	Manila City	Residential	Catch Basin (CB)	Murky wastewater; with few garbage	2
19 Mar 2019	No. 2172-I, Tuazon Compound, District 5, Zone 78, Brgy. 709 (in front)	Taft Ave.	Malate	Manila City	Residential	Catch Basin (CB)	Slightly turbid wastewater; with garbage, sludge, feces, solidified grease	3
22 Mar 2019	No. 1661, Sari-Sari Store (in front)	Sagrada Familia St.	San Andres	Manila City	Non-Residential	Catch Basin (CB)	Not releasing wastewater; with some garbage	8
09 May 2019	Gram Care Diagnostics Center (in front)	A. Mabini St.	Ermita	Manila City	Non-Residential	Drainage Manhole (DM)	Not releasing wastewater; filled with soil	9
10 Jul 2019	Little Ceasar's Pizza (at the side)	Alhambra St.	Ermita	Manila City	Non-Residential	Drainage Manhole (DM)	Not releasing wastewater; with few garbage	7
29 Jul 2019	No. 2172-J, Tuazon Compound, District 5, Zone 78, Brgy. 709 (in front)	Taft Ave.	Malate	Manila City	Residential	Catch Basin (CB)	Clear wastewater; with few garbage and sludge	4
16 Oct 2020	No. 17-D (in front)	Araneta Ave.	Sta. Mesa	Manila City	Residential	Drainage Manhole (DM)	Clear wastewater	1
30 Oct 2020	Cowboy Grill (in front)	A. Mabini St.	Ermita	Manila City	Non-Residential	Drainage Manhole (DM)	Murky wastewater; clogged with garbage and sludge. Replacement with a larger size drainage pipe that does not obstruct flow was recommended.	6
10 Nov 2020	No. 2140 Apostolic Nunciature (in front, center gate)	Taft Ave.	Malate	Manila City	Non-Residential	Drainage Manhole (DM)	Not releasing wastewater; with sludge	5
19 Feb 2021	Starbucks at Bayview Park Hotel (at the side)	United Nations Ave.	Ermita	Manila City	Non-Residential	Drainage Manhole (DM)	Not releasing wastewater; filled with solidified grease and has a foul odor	10

The top 10 sites determined by the Euclidean distance, on the other hand, identified Malate solely as the municipal hotspot of previously endorsed water sampling sites, with four roads cutting through it; however, three sites are located in A. Mabini Street. Similar to the top 10 sites determined by the Cosine distance, the majority are non-residential with connected discharge pipes not releasing wastewater during inspection. In contrast to the former set, 8 of the 10 previously inspected drainage systems were manholes.

Table 20. Top 10 Recommended Locations

for Water Sample Collection (Euclidean Distance)

Date of Inspection	Location	Street	District	City	Site Type	Drainage System	Findings	Rank
19 Feb 2019	First Mission Center of Skills Development, Inc. (in front)	A. Mabini St.	Malate	Manila City	Non-Residential	Drainage Manhole (DM)	Greasy wastewater with human feces	2
19 Mar 2019	Hotel 2016 Manila (in front)	A. Mabini St.	Malate	Manila City	Non-Residential	Drainage Manhole (DM)	Not releasing wastewater; with garbage and sludge	3
22 Mar 2019	Bayad Center (in front)	A. Mabini St.	Ermita	Manila City	Non-Residential	Drainage Manhole (DM)	Not releasing wastewater; with few garbage	8
09 May 2019	Cluster 1, Zone 78, Brgy. 707	Quirino Ave.	Malate	Manila City	Non-Residential	Drainage Manhole (DM)	Clear wastewater; contained few garbage	9
10 Jul 2019	Unit No. 1 Agoncillo Townhomes (in front)	Agoncillo St.	Malate	Manila City	Non-Residential	Catch Basin (CB)	Not releasing wastewater; with few garbage	7
29 Jul 2019	KB24 Restaurant, Ronis Bldg. (in front)	San Marcelino St.	Malate	Manila City	Non-Residential	Catch Basin (CB)	Not releasing wastewater; with dried leaves	4
16 Oct 2020	Décor Modern Furniture and Lighting (in front)	Araneta Ave.	Brgy. Doña Imelda	Quezon City	Non-Residential	Drainage Manhole (DM)	Clear wastewater; contained some garbage	1
30 Oct 2020	Kasara Resort Residences (in front)	Eagle St.	E. Rodriguez	Pasig City	Residential	Drainage Manhole (DM)	Yellowish wastewater with human feces	6
10 Nov 2020	No. 299 (in front)	N. Domingo St.	Brgy. Ermitanyo	Quezon City	Non-Residential	Drainage Manhole (DM)	Grayish wastewater	5
19 Feb 2021	US Dontics Dental Supply (in front)	Ramon Magsaysay Blvd.	Sta. Mesa	Manila City	Non-Residential	Drainage Manhole (DM)	Not releasing wastewater; with few garbage	10

By consolidating the insights from the two sets of top 10 recommended sites for water sample collection, Malate and Ermita are identified as the municipal hotspots of potential non-compliant establishments, with Taft Avenue and A. Mabini Street as the hotspot thoroughfares. The majority of the top sites are non-residential with connected discharge pipes in either manholes or catch basins, but were not previously sampled due to the lack of wastewater flow. Additionally, none of the top sites have been revisited to collect water samples in two to three years now due to the suspension of field operations brought about by the coronavirus restrictions.

V. Platform or App Description

V.A. Python Scripts and Insights

The 1) insights obtained from the machine learning models, and the 2) priority locations to revisit obtained from the recommender system will be delivered and presented to the MMDA-SWMO in PowerPoint and PDF formats. The codes will be compiled and submitted to AIM as a Jupyter notebook.

V.B. Data Product 3: Fieldwork on AppSheet

The Fieldwork Tool, with the Risk Map feature, is hosted on AppSheet, an application that provides a no-code development platform for various applications owned by Google LLC. Specifically, the Fieldwork Tool will help assist the MMDA-SWMO field team and engineers by having a digital version of the checklist, shown in Figure 16, on their mobile, laptop or desktop devices (a sample is available on https://bit.ly/fieldworktool). Using Google Sheets as the backend source of data, an application can be built to aid in the data gathering aspect of the site inspection task of the MMDA-SWMO, which is also useful for having real-time data available to multiple stakeholders.

The map of the City of Manila and points of interest can be easily visualized through a mobile device or through a computer screen. This capability is made possible by Google Maps. An example is provided in Figure 31. Previously visited locations that have available longitude and latitude data from the masterfile checklists can be migrated into the source Google Sheets. Likewise, new locations can be added through the application.

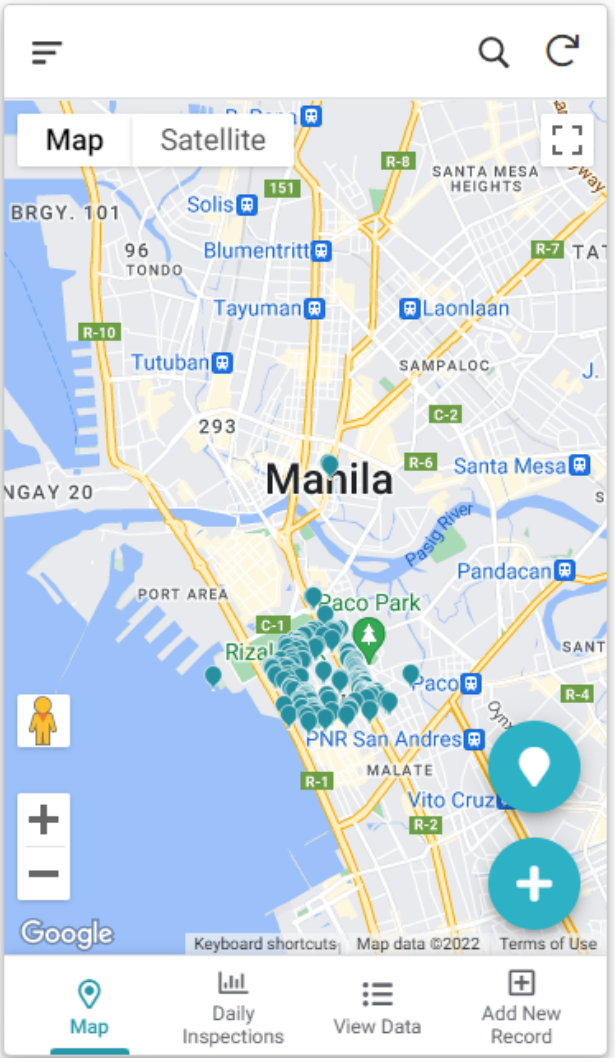


Figure 31. AppSheet Interface with Map and Locations using a Mobile Device. Previously inspected sites with available coordinates are shown as pins.

When conducting site inspections, a fieldwork personnel can access the AppSheet through an internet connection, then add or update records as needed. The features of the fieldwork tool are the same as the columns of the field checklist, with the addition of image uploading. The coordinates can be obtained through the Global Positioning System (GPS) of the field personnel's mobile phone, or from a third-party open-source mobile camera application. Sections of the sample interface are shown in Figures 32 and 33.

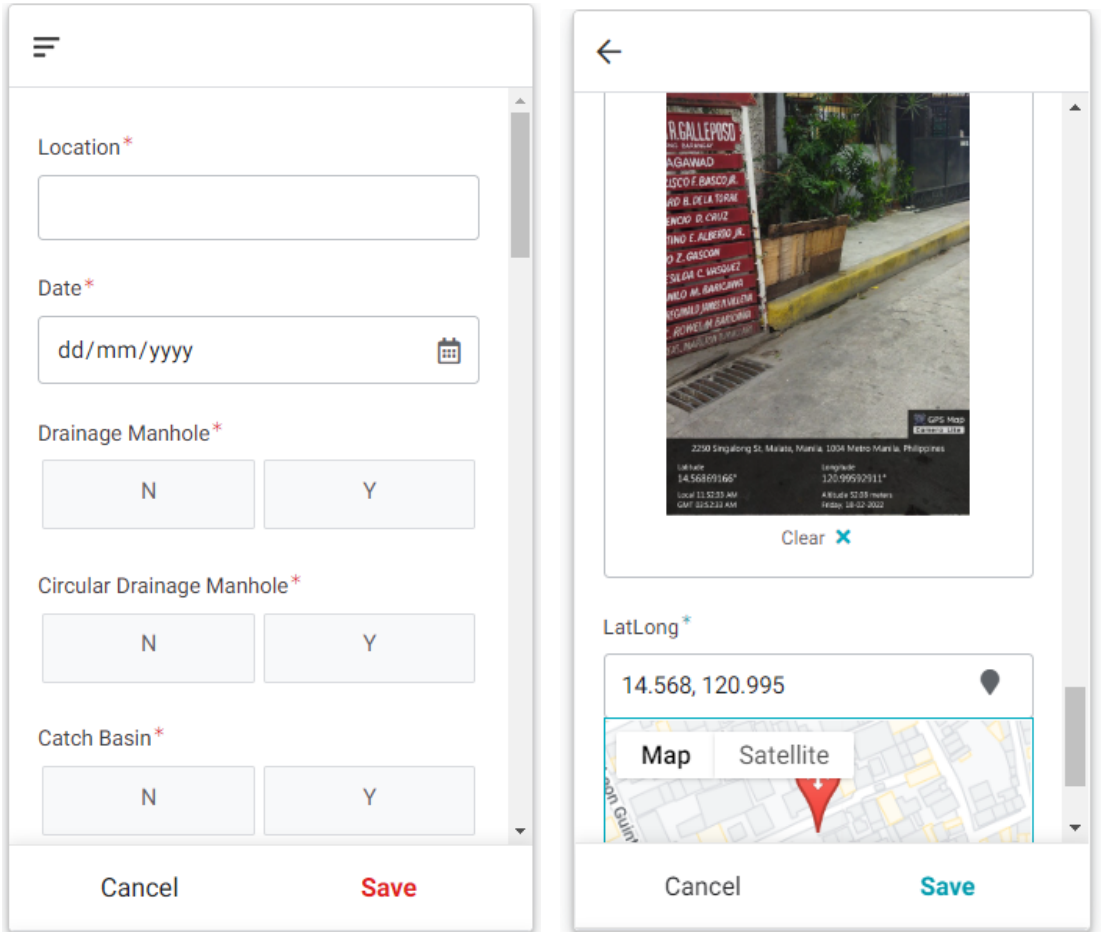
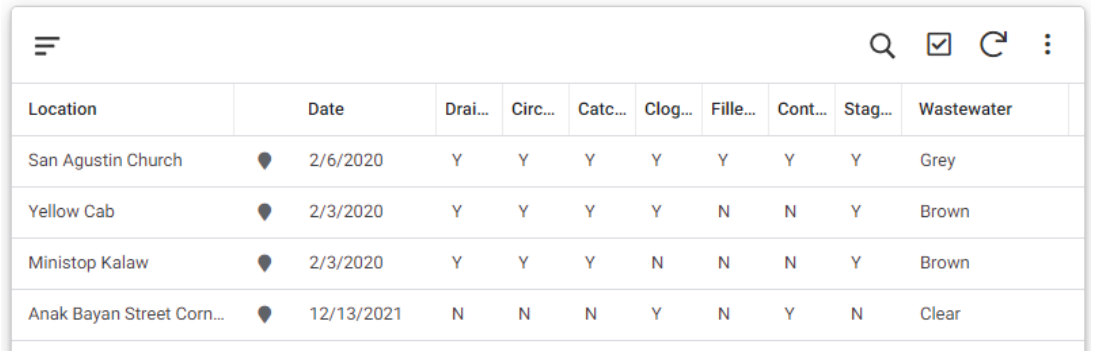


Figure 32. AppSheet Interface: Add New Record - Details, Image, and GPS. The features in the fieldwork tool correspond to the columns of the field checklist.

The data can also be viewed either through the AppSheet mobile application itself or the Google Sheets backend data. The access rights can be defined based on the role of the employee. Figure 33 shows an example data interface on AppSheet.



Location	Date	Drai...	Circ...	Catc...	Clog...	Fille...	Cont...	Stag...	Wastewater
San Agustin Church	2/6/2020	Y	Y	Y	Y	Y	Y	Y	Grey
Yellow Cab	2/3/2020	Y	Y	Y	Y	N	N	Y	Brown
Ministop Kalaw	2/3/2020	Y	Y	Y	N	N	N	Y	Brown
Anak Bayan Street Corn...	12/13/2021	N	N	N	Y	N	Y	N	Clear

Figure 33. AppSheet Interface: Detailed Data View. All information contained in the MMDA-SWMO's Consolidated Report is present in the fieldwork tool.

V.C. Data Product 4: Dashboard Tool on Data Studio

Data Studio is an open-source dashboard tool provided by Google LLC as part of their Google Cloud Platform suite. It easily connects data from spreadsheets and other data sources. This data product was created to supplement the Fieldwork Tool in its digitization objective, and enable the MMDA-SWMO to have seamless and timely generation of reports. The reports may be more valuable to the heads and directors as it presents summary statistics based on a specified timeframe. The reports, as well as the linked data, are easily downloadable and updated in real-time.

Three links will be provided to the MMDA-SWMO. The first is the **Data Studio dashboard link** (accessible via this link: https://bit.ly/dashboard-tool), which leads to two pages of customizable reports: 1) a visualization of the division's historical data from January 2019 up to March 2021, as shown in Figure 34, and 2) a blank dashboard template for visualizing their inspection and water sampling operations when fieldwork resumes. The historical dashboard displays the actual features from the raw data such as the locations and counts of each task, but with additional derived features such as the inspection status, type of the sites, municipal districts, and distribution of tasks and locations, among other things. The dashboard template of the future data includes all of the features present in the historical dashboard, as well as new features such as the assigned team or division, and personnel. Figure 35 depicts a sample report that will be generated when new data is submitted.

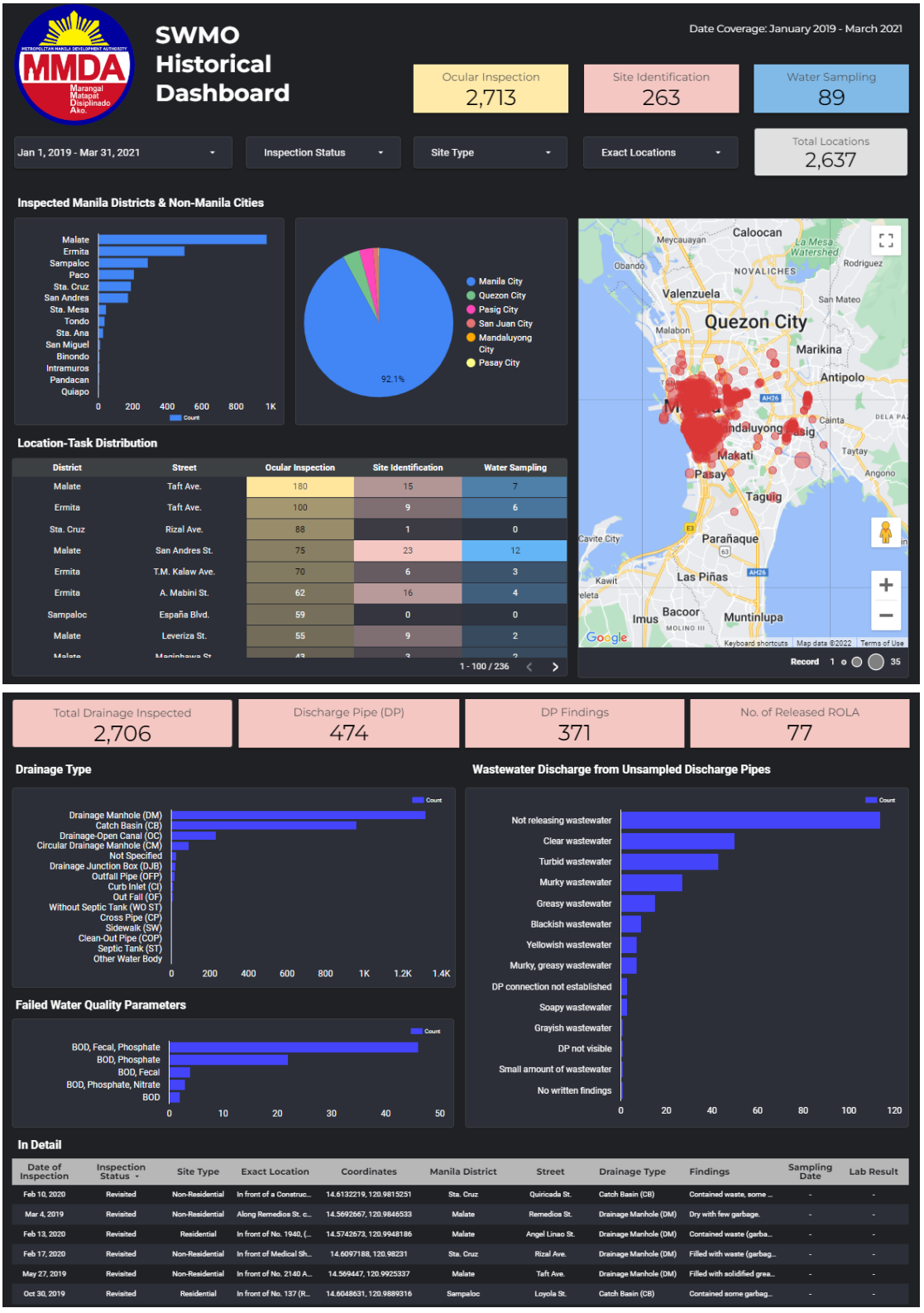


Figure 34. MMDA-SWMO Historical Dashboard (2019-2021). The Dashboard Tool allows for quick visualization and report generation of the division's historical data.

Figure 35. Sample Generated Dashboard for MMDA-SWMO's Future Data. This template includes both historical and new features, such as task assignment.

For the upcoming resumption of site inspections and water sampling, we suggest gathering additional data to allow the agency to track and monitor its operations more effectively, such as districts, personnel email addresses, and division assignments, among other things. Data Studio has a *can edit* and *can view* access levels, and the security clearance will be discussed with the MMDA-SWMO.

The second link is to the **Task Form**, which staff must complete when self-assigning or assigning tasks to teammates. It is powered by Google Forms, an open-source survey administration software included in Google Docs Editors suite. If provided with the link, any email provider, including Gmail, Yahoo Mail, Outlook, and

enterprise emails, can access the Task Form. The Google Form also has its own basic visualization of summary statistics that anyone with *edit access* can view, but the charts can only be copied one at a time (see Figure 36). *Edit access* is also allowed to update the questions and details in the Task Form. For notification and approval functions, we added an add-on feature called *Form Approvals* to the Task Form. Following consultation with the MMDA-SWMO, the security clearance and approval order will be finalized.

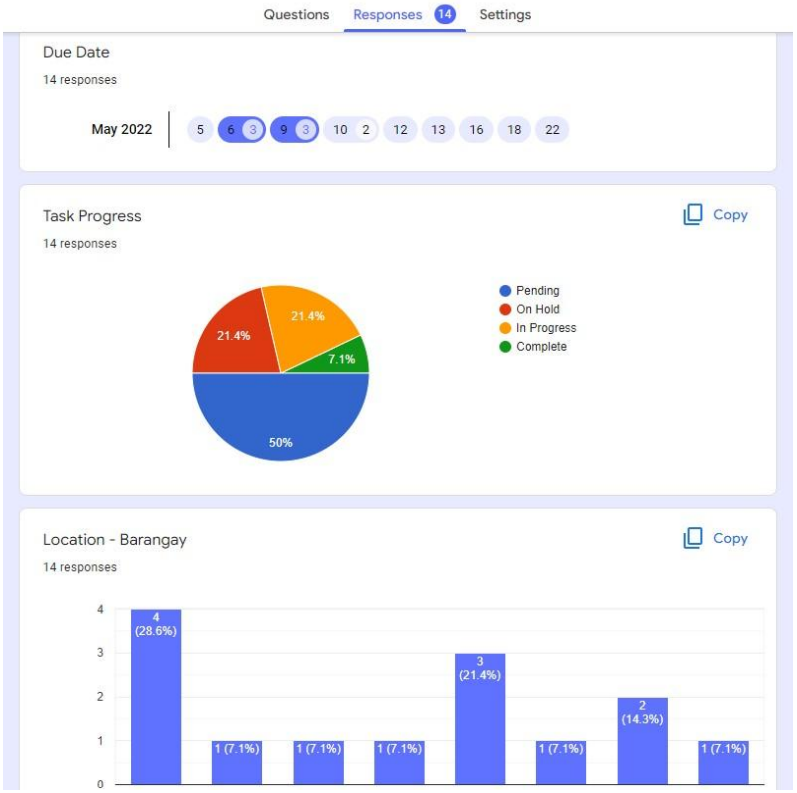
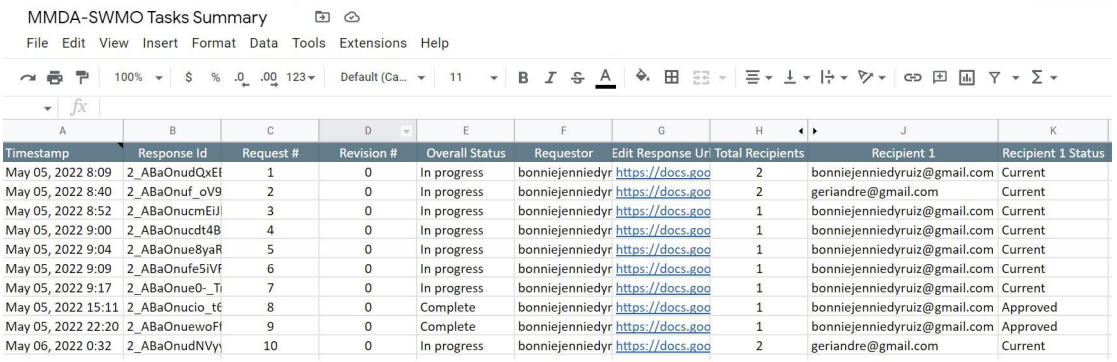


Figure 36. Sample Task Form Responses via Google Forms. This application provides a basic visualization of summary statistics to anyone with edit access.

The responses from the Task Form are saved to **MMDA-SWMO Data** Google Sheet that is linked to the Data Studio. Once submitted, the responses will appear in real-time on the Google Sheet, then on the Data Studio reports ten to fifteen minutes later, or as soon as someone with edit access refreshes them. As shown in Figure 37,

the Google Sheet has two tabs that cannot be edited or deleted. Its data is updated every time a Task Form is submitted or resubmitted, or when a task action, e.g., accepted, completed, declined, and so on, is selected via the Form Approval link included in the notification email (see Figure 38). The Google Sheet also has *edit* and *view* access levels, though the former is insignificant because any changes made directly on the sheet will not be reflected in the Data Studio reports. The security clearance has yet to be finalized with the MMDA-SWMO.


MMDA-SWMO Tasks Summary

File Edit View Insert Format Data Tools Extensions Help

Timestamp	Response Id	Request #	Revision #	Overall Status	Requestor	Edit Response Url	Total Recipients	Recipient 1	Recipient 1 Status
May 05, 2022 8:09	2_ABaOnudQxEI	1	0	In progress	bonniejenniedyr	https://docs.goo	2	bonniejenniedyruiz@gmail.com	Current
May 05, 2022 8:40	2_ABaOnuf_oV9	2	0	In progress	bonniejenniedyr	https://docs.goo	2	geriandre@gmail.com	Current
May 05, 2022 8:52	2_ABaOnucmEiJ	3	0	In progress	bonniejenniedyr	https://docs.goo	1	bonniejenniedyruiz@gmail.com	Current
May 05, 2022 9:00	2_ABaOnucdt4B	4	0	In progress	bonniejenniedyr	https://docs.goo	1	bonniejenniedyruiz@gmail.com	Current
May 05, 2022 9:04	2_ABaOnue8yaR	5	0	In progress	bonniejenniedyr	https://docs.goo	1	bonniejenniedyruiz@gmail.com	Current
May 05, 2022 9:09	2_ABaOnufe5iVF	6	0	In progress	bonniejenniedyr	https://docs.goo	1	bonniejenniedyruiz@gmail.com	Current
May 05, 2022 9:17	2_ABaOnue0-_Ti	7	0	In progress	bonniejenniedyr	https://docs.goo	1	bonniejenniedyruiz@gmail.com	Current
May 05, 2022 15:11	2_ABaOnucio_t6	8	0	Complete	bonniejenniedyr	https://docs.goo	1	bonniejenniedyruiz@gmail.com	Approved
May 05, 2022 22:20	2_ABaOnuewoFf	9	0	Complete	bonniejenniedyr	https://docs.goo	1	bonniejenniedyruiz@gmail.com	Approved
May 06, 2022 0:32	2_ABaOnudNVy	10	0	In progress	bonniejenniedyr	https://docs.goo	2	geriandre@gmail.com	Current

Figure 37. MMDA-SWMO Tasks Summary on Google Sheets. The task form responses will appear in real-time on Google Sheets and Data Studio.

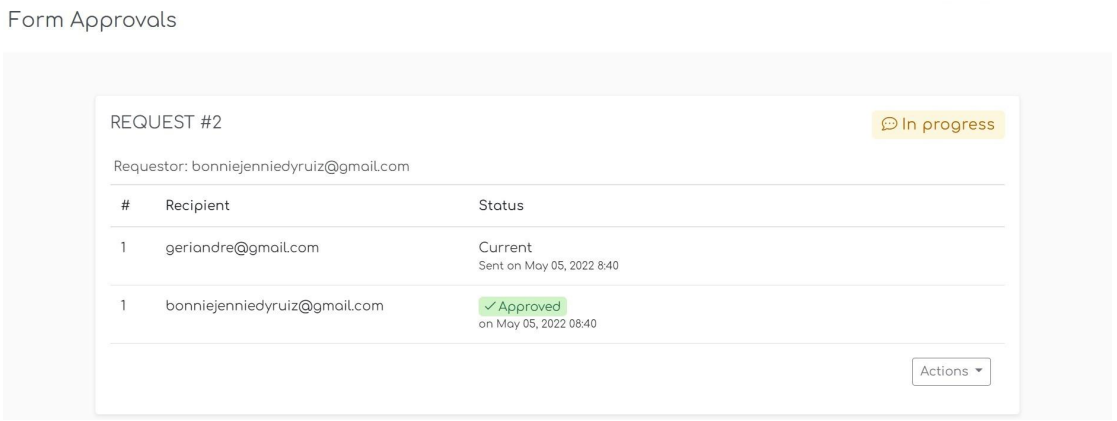


Figure 38. Task Approval Status and Actions. The notification email contains a Form Approval link, from which the user can select task actions.

The three links described are mobile-friendly, but a good internet connection and a fairly wide mobile screen are required to appreciate the Data Studio charts. For

ease of maintenance, we intend to combine the AppSheet and the Data Studio linked Google Sheets into a single file. The AppSheet serves as the Fieldwork Tool for the field team, while the Data Studio generates reports for the entire division. A user onboarding session is part of our team's commitment to MMDA-SWMO.

VI. Conclusions

VI.A. Summary of Findings

Textual descriptions of the inspected sites and their associated wastewater discharge have been valuable in determining the important characteristics of potential water sampling sites. Of all the text vectorization techniques implemented, the TF-IDF representation proved to be the best model to use due to its excellent performance, interpretability, and short training time. With Gradient Boosting Method, the text-based classification model achieved a 99% test accuracy and a 94% F1-Score.

In addition to the presence of discharge pipes in a drainage site, the SHAP interpretability method revealed the most influential tokens in predicting a site's eligibility for water sample collection and potential non-compliance with water quality requirements, which aligns with the clustering results and what the MMDA-SWMO has determined from past inspection and water sampling activities. The significant features identified were '*release', 'inspection' 'turbid', 'murky', 'greasy'*, and *'foul'*.

Through the Geospatial Recommender Systems, we generated two sets of potential non-compliant locations from the previously inspected and identified water sampling sites: 1) top 100 drainage sites without discharge pipes that have a 9% higher likelihood of containing turbid, murky, or greasy wastewater and various wastes such as garbage, sludge, grease, sand, and soil, and; 2) top 100 drainage sites with discharge pipes that have a 2% higher likelihood of failing to meet water quality standards when compared to existing MMDA-SWMO procedures. For the first set, we utilized the top features extracted through SHAP, with Hamming Distance as the best metric; for the second set, we filtered sites with *'dp'* or '*discharge pipes*' in their textual descriptions,

ensuring that the primary requirement of a water sampling site is met, with Cosine and Euclidean Distance as the best metrics.

VI.B. Recommendation

Our Capstone Team was able to propose solutions to address the MMDA-SWMO's operational challenges related to its functions for the Manila Bay Rehabilitation Project, demonstrating the usefulness and effectiveness of employing data science techniques to improve organizational or business functions such as coordination, scheduling, reporting, task prioritization, and resource allocation, among others. We recommend that the agency employ the four data products that we developed as decision-support tools: 1) a Water Quality Prediction Model, 2) a Geospatial Recommender System, 3) a Fieldwork Tool, and 4) a Dashboard Tool.

Based on the Water Quality Prediction Model results, we recommend that the MMDA-SWMO collect water samples from a drainage site with an *inspected* discharge pipe *releasing* wastewater that is visibly *turbid* or *murky*, *greasy*, and has a *foul odor* because it is highly likely to fail the laboratory tests, which check for BOD, fecal coliform, phosphate, and nitrate contents.

Using the results of Geospatial Recommender Systems, we can assist MMDA-SWMO, as well as the MMDA-FCSMO, make decisions about area prioritization and resource allocation for its key field tasks: site inspection and water sample collection. We recommend that the MMDA-SWMO, in collaboration with the MMDA-FCSMO, prioritize the *'Top 100 Recommended Sites for Reinspection & Cleaning'* provided in Appendix 1, as these are determined the potential hotspots of unclean drainage systems, with Singalong Street and Dama de Noche Street in Malate leading the list. We further recommend that MMDA-SWMO prioritize the 123 unique

sites from the combined *'Top 100 Recommended Sites for Water Sample Collection'*, provided in Appendix 2 and 3, as these are determined the potential hotspots of non-compliant wastewater from discharge pipes, with A. Mabini Street and Taft Avenue, which cut through both Malate and Ermita, topping the list.

References

Aldhyani, T. H. H., Al-Yaari, M., Alkahtani, H., & Maashi, M. (2020). Water Quality Prediction Using Artificial Intelligence Algorithms. *Applied Bionics and Biomechanics*, 2020, 1–12. https://doi.org/10.1155/2020/6659314.

Alipio, M. I. (2021). Towards Developing A Classification Model For Water Potability In Philippine Rural Areas. *ASEAN Engineering Journal*, 10(2). https://doi.org/10.11113/aej.v10.16594.

Brooke, Sophia. (2018). How Data Science Is Enabling Better Decision-making. *Towards Data Science*. https://towardsdatascience.com/how-data-science-is-enabling-better-decision-making

Chen, F. L., Yang, B. C., Peng, S. Y., & Lin, T. C. (2020). Applying a deployment strategy and data analysis model for water quality continuous monitoring and management. *International Journal of Distributed Sensor Networks*, *16*(6), 155014772092982. https://doi.org/10.1177/1550147720929825.

Department of Environment and Natural Resources. (Updated 2022). RA 9275 – The Philippine Clean Water Act. *EMB*. https://r12.emb.gov.ph/ra-9275-the-philippine-clean-water-act/.

Google. (Accessed 2022). AppSheet. *Google AppSheet.* https://about.appsheet.com/home/.

Google. (Accessed 2022). Data Studio. *Google Marketing Platform.* https://marketingplatform.google.com/about/data-studio/.

Google. (Accessed 2022). Data Studio Overview. *Google Data Studio.* https://datastudio.google.com/overview.

He, H. Bai, Y., Garcia, E.A., Li, S. (2008). Adasyn: adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks*, IEEE World Congress on Computational Intelligence, 1322–1328. IEEE, 2008. https://ieeexplore.ieee.org/document/4633969.

Kononenko, I., Kukar, M. (2007). Symbolic Learning. *Machine Learning and Data Mining*, Chapter 9. https://www.sciencedirect.com/science/article/pii/B9781904275213500095.

Kumpel, E., MacLeod, C., Stuart, K., Cock-Esteb, A., Khush, R., & Peletz, R. (2020).

From data to decisions: understanding information flows within regulatory water quality monitoring programs. *NPJ Clean Water*, *3*(1). https://www.nature.com/articles/s41545-020-00084-0.

Manoiu, V. M., Wójcicka, K. K., Craciun, A. I., Akman, C., Akman, E. (2022). Water Quality and Water Pollution in Time of COVID-19: Positive and Negative Repercussions. *MDPI*. https://www.mdpi.com/2073-4441/14/7/1124/pdf.

Maulion, A. F. (2020). Rapid Water Quality Assessment of Bongoy River in Odiongan, Romblon Using Macro Invertebrates and Fecal Coliform Presence as Bioindicators. *International Journal of Sciences: Basic and Applied Research,* 2020, 69-79. https://www.gssrr.org/index.php/JournalOfBasicAndApplied/article/view/11820.

Molnar, Christoph. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. *GitHub.* https://christophm.github.io/interpretable-ml-book/

Müller, A., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists (1st ed.). O'Reilly Media.

Nazario, Dhel. (2020). Program for Manila Bay Rehabilitation Launched Online. *Manila Bulletin.*https://mb.com.ph/2020/06/06/program-for-manila-bay-rehabilitation-launched-online/

Pennington, J., Socher, R., Manning, C.D. (2014). GloVe: Global Vectors for Word Representation. *Stanford University.* https://nlp.stanford.edu/pubs/glove.pdf.

Presidential Communications Operations Office. (2019). Administrative Order No. 16, s. 2019. *Official Gazette.* https://www.officialgazette.gov.ph/2019/02/19/administrative-order-no-16-s- 2019/.

Senate of the Philippines. (2013). Legarda: Manila Bay Rehab to Benefit 23 Million Filipinos. *Legacy Senate of the Philippines.* http://legacy.senate.gov.ph/press_release/2013/1008_legarda1.asp.

Zlaoui, Khalil. (2021). Interpretable Machine Learning using SHAP: theory and applications. *Towards Data Science.* https://towardsdatascience.com/Interpretable-machine-learning-using-shap-theory-and-applications-26c12f7a7f1a

Appendix

Appendix 1.

Top 100 Recommended Sites for Reinspection & Cleaning (Hamming Distance)

1. In front of No. 2219, Singalong St., Manila (1st Catch basin from the right side of establishment)
2. In front of No. 2219, Singalong St., Manila (2nd Catch basin from the right side of establishment)
3. In front of No. 2008 (Residential House), Kasoy St., Alley I, Manila (2nd Open canal from the right side of establishment)
4. In front of No. 2008 (Residential House), Kasoy St., Alley I, Manila (1st Open canal from the left side of establishment)
5. In front of No. 2181, PSA Bike Shop, Singalong St., Manila
6. At the side of PSA Bike Shop, Singalong St., corner Ermin St., Manila
7. In front of Betamia General Merchandise, Angel Linao St., Malate, Manila
8. In front of No. 2329, (Residential House), Dama de Noche St., Malate, Manila
9. In front of No. 2328 - B, (Residential House), Dama de Noche St., Malate, Manila
10. In front of No. 2334, (Residential House), Dama de Noche St., Malate, Manila
11. In front of Vacant Lot Pedro Gil St., Manila
12. Beside No. 2323, (Residential House), Dama de Noche St., Malate, Manila
13. In front of Vacant Lot, Pedro Gil St., Manila
14. In front of No. 2355, Piso Net, Dama de Noche St., Malate, Manila
15. In front of No. 2323, (Residential House), Dama de Noche St., Malate, Manila
16. In front of an Apartment Building, Pres. Quirino Ave., corner Anakbayan St., Malaite, Manila
17. In front of No. 1164, Fullspech Building, Pres. Quirino Ave., corner Anakbayan St., Malate, Manila
18. In front of No. 2341 - A, (Residential House), Dama de Noche St., Malate, Manila
19. Across PAGCOR Bldg. M.H. Del Pilar St., (30 meters from Pedro Gil St.)
20. In front of No. L-9B Khumba Enterprise Inc., Amang Rodriguez Ave., Brgy. Manggahan , Pasig City
21. In front of #1951 (Residential House) Dagonoy St., Sta. Ana, Manila
22. In front of No. 1836-A (Residential House), San Pedro St., Manila
23. Beside O.C.T Motorcycle Shop and Merchandise, Dagonoy St., Sta. Ana, Manila
24. In front of No. 1850 (Residential House), San Pedro St., Manila
25. In front of God's Child Trading, Dagonoy St., Sta. Ana, Manila
26. MRI Food Terminal, Sta. Ana, Manila
27. In front of No. 1256 (Residential House), Dalandan St., San Andres, Manila.
28. In front of No. 1214, (Residential House), Francisco St., Sta. Ana, Manila
29. In front of a Residential House along Arellano St., corner Leyte St., Manila
30. In front of a Residential House along Leyte St., corner Arellano St., Manila
31. In front of No. 1105-A (Residential House), Dagonoy St., Malate, Manila
32. In front of Nouveau Marketing Inc. along Eagle St., E. Rodriguez Ave., Pasig City
33. In front of No.2393-A (Residential House), F. Muñoz St., Corner Dagonoy St., Singalong, Malate, Manila
34. In front of Vista Residences, Laon-Laan St., Sta. Cruz Manila
35. In front of a Construction Site, Laon-Laan St., Sta. Cruz, Manila - Drainage Manhole
36. In front of a Construction Site, Laon-Laan St., Sta. Cruz, Manila - Catch Basin
37. In front of a Construction site, Laon-Laan St., Sta. Cruz, Manila
38. At the left side of No. 1111 (Residential House), F. Muñoz corner Dagonoy St., Singalong, Malate, Manila - 2nd Line Canal
39. At the left side of No. 1111 (Residential House), F. Muñoz corner Dagonoy St., Singalong, Malate, Manila - 4th Line Canal
40. At the left side of No. 1111 (Residential House), F. Muñoz corner Dagonoy St., Singalong, Malate, Manila - 3rd Line Canal
41. At the left side of No. 1111 (Residential House), F. Muñoz corner Dagonoy St., Singalong, Malate, Manila - 1st Line Canal
42. In front of No.115, between Building 3 & 4 (Residential House), West Zamora St., Pandacan, Manila
43. In front of No. 104, between Building 3&4 (Residential House), West Zamora St., Pandacan, Manila
44. In front of a gray gate between Building 3 & 4 (Residential House), West Zamora St., Pandacan, Manila
45. At the side of No. 86, (Building 3 - Apartment), Barangay 841, West Zamora, Pandacan, Manila
46. Across No. 86, (Building 3 Apartment), West Zamora St. Pandacan Manila.
47. In front of PLDT (Left side) along C5 E. Rodriguez corner Corporal Cruz, Pasig City
48. In front of PLDT (Right side) along C5 E. Rodriguez corner Corporal Cruz, Pasig City
49. In front of No. 2017 (Residential House), F. Muñoz St., Paco, Manila
50. In front of No. 2061, Sari-sari Store, F. Muñoz St., Paco, Manila
51. In front of No 2061, Mini Store, F. Muñoz St., Paco, Manila
52. In front of No. 2011 (Residential house), F. Muñoz St., Paco, Manila
53. In front of No. 1998 – A1 (Residential House), F. Muñoz St., Paco, Manila
54. In front of No. 1275 Residential House along Mataas na Lupa, Brgy. 736 Zone 80 San Andres, Manila
55. In front of No. 1998 – A (Residential House), F. Muñoz St., Paco, Manila
56. In front of No. 1281 Sari Sari Store along Mataas na Lupa, Brgy. 736 Zone 80 San Andres, Manila
57. In front of No. 1273 Residential House along Mataas na Lupa, Brgy. 736 Zone 80 San Andres, Manila
58. In front of No. 1992 (Residential House), F. Muñoz St., Paco, Manila

59. In front of No. 1993 (Residential House), F. Muñoz St., Paco, Manila (1st Catch Basin)
60. In front of Jireh Bakery along Mataas na Lupa, Brgy. 736 Zone 80 San Andres, Manila
61. In front of No. 1993 (Residential House), F. Muñoz St., Paco, Manila (1st Drainage manhole)
62. Beside Edith and Cristy Sisig & Tapsilog, F. Muñoz St., Paco, Manila
63. Beside No. 1255 Apartment along Mataas na Lupa, Brgy. 736 Zone 80 San Andres, Manila
64. In front of 2005 (Residential House), F. Muñoz St., Paco, Manila
65. In front of No. 1245 Apartment Mataas na Lupa, Brgy. 736 Zone 80 San Andres, Manila
66. Right side corner of No. 2439, (Sari-sari Store), Arellano St., San Andres Bukid, Manila
67. 3rd Door from the left side of No. 2004-C (Residential House), Kasoy St., Manila
68. In front of No. 2004-C (Residential House), Kasoy St., Alley II, Manila
69. In front of No. 2006-B (Residential House), Kasoy St., Alley II, Manila
70. In front of No.2383 (Sari-Sari Store), Dagonoy St., Malate, Manila
71. In front of No. 2007 Residence House/Sari-Sari Store Anak Bayan St., San andres Bukd, Manila
72. In front of No. 2007 Residence House/Sari-Sari Store Anak Bayan St., San Andres Bukid, Manila
73. In front of No. 2391, (Sari-sari Store), Arellano St., San Andres Bukid, Manila
74. In front of No. 2030 (Sari-Sari Store), Bayabas Alley, Anak Bayan St., San Andres Bukid, Manila
75. In front of Lotto Outlet, Angel Linao St., Manila
76. In front of No. 2010 Sari-sari Store, Angel Linao St., Manila
77. Across Lotto Outlet, Angel Linao St., Manila
78. At the side of Lotto Outlet, Angel Linao St., San Andres Bukid, Manila
79. In front of No. 2026, (Residential House) Angel Linao St., San Andres Bukid, Manila
80. In front of No. 1422 Delight Construction Supply Pedro Gil St., Paco, Manila - b. Drainage manhole
81. In front of No. 1422 Delight Construction Supply Pedro Gil St., Paco, Manila - a. Catch Basin
82. In front of No. 1418 Furnicom Furniture Pedro Gil St., Paco, Manila
83. In front of Arthur's Cut Rite, Taft Ave., Malate, Manila
84. In front of a Barber Shop along Mataas na Lupa St., Paco, Manila
85. Across Karl and Annie Store, F. Muñoz St., corner Mataas na Lupa St., Paco, Manila
86. In front of No. 1273 (Residential House), Mataas na Lupa St., Paco, Manila.
87. In front of No. 1269 (Residential House), Mataas na Lupa St., Paco, Manila.
88. In front of No. 1281-A (Residential House), Mataas na Lupa St., Paco, Manila.
89. In front of No. 1259 (Residential House with Sari-Sari Store), Mataas na Lupa St., Paco, Manila.
90. In front of Karl and Annie Store, F. Muñoz St., corner Mataas na Lupa St., Paco, Manila
91. In front of a soft drink and beer Dealer along Mataas na Lupa St., Paco, Manila.
92. In front of Residential House with Eatery along Mataas na Lupa St., Paco, Manila.
93. In front of No. 1281 (Residential House), Mataas na Lupa St., Paco, Manila.
94. In front of Central College of the Philippines – Gate 2 (right side), Aurora Blvd., Quezon City
95. In front of No. 1997 Apartment, Angel Linao St., Manila
96. In front of Residential House Gate I, Parola, Manila - 1st Drainage Manhole
97. In front of Residential House Gate I, Parola, Manila - 2nd Drainage Manhole
98. Across MPD Station 12 Gate 10 Brgy. 275 Zone 25, Parola, Manila
99. Across Iglesia ni Cristo Gate 9, Parola, Manila
100. In front of Jhoana Store Ana Sarmiento St., san Andres Bukid, Manila

Appendix 2.

Top 100 Recommended Sites for Water Sample Collection (Cosine Distance)

1. In front of Residential House, No. 17-D, G. Araneta Ave., Sta. Mesa, Quezon City
2. In front of Residential House No. 2172A, Tuazon Compound, District 5, Brgy. 709, Zone 78 (Left Side)
3. In front of Residential House No. 2172I, Tuazon Compound, District 5, Brgy. 709, Zone 78
4. In front of Residential House No. 2172J, Tuazon Compound, District 5, Brgy. 709, Zone 78
5. In front of No. 2140 Apostolic Nunciature, Taft Avenue, Manila (2nd drainage from the right of the center gate of the establishment)
6. In front of Cowboy Grill, A. Mabini St cor. Arsenio Herrera St. Ermita, Manila
7. At the side of Little Ceasar's Pizza, U.N Ave., Corner Alhambra St., Ermita, Manila
8. In front of No. 1661, Sari-sari Store, Sagrada Familia St., Manila
9. In front of Gram Care Diagnostics Center, Mabini St., Ermita, Manila
10. At the side of Starbucks at Bayview Park Hotel Manila, Roxas Blvd., Service Road Cor. UN Ave., Ermita, Manila
11. In front of Décor Modern Furniture and Lightning along Araneta Ave., corner Palanza St., Quezon City
12. In front of Red Dot Restaurant and Bar, A. Mabini St.cor. Arsenio Herrera St., Ermita, Manila
13. In front of The Contemporary Hotel along Araneta Ave., Quezon City
14. In front of US Dontics Dental Supply Ramon Magsaysay Blvd., Manila
15. In front of the MRF of Bgy.712, Maginhawa St.,
16. At the side of Smile to Go Milk Tea, Morayta St., Sampaloc, Manila
17. Along A. Mabini St., in front of First Mission Center of Skills Development, Inc.
18. In front of Bgy. Hall of Bgy. 713, Maginhawa St.,
19. A. Mabini St., Ermita, Manila in front of Calle Bar
20. University of the East Ramon Magsaysay (UERM) Memorial Hospital

21. San Andres St., cor. Mariposa St., Malate Manila across Jollibee Fast Food
22. In front of No. 2073 (Residential House), Smith St., Corner San Andres St., Malate, Manila (1st Catch Basin)
23. In front of Bgy. Hall of Bgy. 708, Maginhawa St.,
24. Across Kasaganaan St. cor. Agno St., Bgy.708, Maginhawa St.,
25. No. 1411 Mabini St., Ermita, Mla.
26. In front of One Archer Condominium Building along Fidel A. Reyes St., District 5, Brgy. 709, Zone 78
27. A. Mabini St. Cor. Sta. Monica St., Ermita Manila in front of Good Rate Money Changer (1st Drainage manhole)
28. In front of Dunkin' Donuts, Taft Ave., Malate, Manila - Drainage Manhole
29. In front of Residential Bldg. beside Beauzitl Bldg., San Andres St., Ermita, Manila (4th Drainage junction box)
30. No. 1444 Mabini St., Ermita, Manila
31. In front of Mutsarap Fried, G. Araneta Ave., Sta. Mesa, Quezon City
32. In front of no. 2276-F, Int. 3, Brgy. 717, Malate, Manila
33. In front of no. 2278-E, Int.3, Brgy. 717, Malate, Manila
34. In front of no. 2280-B, Int. 3, Brgy. 717, Malate, Manila
35. In front of no. 2288 Int 39-D, Brgy. 717, Malate, Manila
36. In front of no. 2288 Int 39-B, Brgy. 717 Malate, Manila
37. In front of temple of Heaven Property, 2160 along Fidel A. Reyes St., District 5, Brgy. 709, Zone 78
38. In front of Unit no. 1 Agoncillo Townhomes, Agoncillo St., Malate, Manila
39. In front of 2288 Interior 20, Ramirez Compound, Barangay 718
40. In front of KB24 Restaurant, Ronis Bldg., San Marcelino St., Corner San Andres St., Malate, Manila
41. In front of Hotel 2016 Manila, A. Mabini St. Malate, Mla
42. Along Leon Guinto St., cor. Gen. Malvar, in front of Philippine Women's University, School of Fine Arts and Design
43. Beside Alpan Automotive Auto cars Inc., G. Araneta Ave., Quezon City
44. In front of No. 299 N. Domingo St., (Ongoing Construction), Brgy. Ermitanyo, Quezon City
45. In front of No. 2018 (Private Property), Smith St., Malate, Manila
46. In front of KB4 Restaurant, No. 961 San Andres St., Malate, Manila
47. Along Guerrero St., beside Shalom Hotel
48. In front of Twenty 202 Bldg., Taft Ave., Manila (right side)
49. SM Sta.Mesa, Manila
50. In front of Aling Banang Eatery, N. Domingo St., Brgy., Corazon De Jesus, San Juan City - Circular Drainage manhole without cover
51. In front of No. 2281 (Residential House), Syquia St., Brgy. 874 Sta. Ana, Manila
52. In front of OSM Building, Pedro Gil St., Ermita, Manila
53. In front of Mini-stop Convenience Store, along Aurora Blvd., Quezon City
54. In front of Kasara Resort Residences along Eagle St., E. Rodriguez Ave., Pasig City
55. In front of Northpark Noodles, G. Araneta Ave., Quezon City
56. In front of No.651, Bgy 713, Maginhawa St.,
57. In front of No. 382 (Residential House), Minerva St., Malacañang, Manila (Catch Basin)
58. In front of No. 378 (Town House), Minerva St., Malacañang, Manila (Circular Drainage Manhole)
59. In front of No. 393 (Residential House), Minerva St., Malacañang, Manila (left side)
60. Across No. 382 (Residential House), Minerva St., Malacañang, Manila
61. In front of No. 1416 (Residential House), Agoncillo St., Manila
62. In front of No. 1433 (Residential House), Agoncillo St. Manila
63. In front of Gea Mel Eatery, Angel Linao St., corner San Andres, Manila (1st Catch basin with steel grating cover from the right side of establishment)
64. In front of No. 1108, Yougo Chinese Store, San Marcelino St., corner Natividad Lopez St., Manila
65. In front of No. 10 City Shutter, Valencia St., Quezon City
66. In front of No. 19 - Residential House, N. Domingo St., Quezon City (right side)
67. In front of No. 3, The Laundry House, N. Domingo St., corner Valencia St., Quezon City
68. Along A. Mabini St. cor. San Andres St., at the side of Malate Church
69. In front of Panda Construction Supply Inc., G. Araneta Ave., Quezon City
70. In front of Jason Store and Eatery, Benitez St., corner Pedro Gil St., Manila
71. In front of PBCOM, G. Araneta Ave., Quezon City
72. In front of Myxgen Audio Video Morayta St., Sampaloc, Manila
73. In front of Sam & Mas Eatery, Loyola St., Sampaloc, Manila (2nd Catch Basin from the right-side of establishment)
74. In front of No. 21 New Manila Condominium, N. Domingo St., Quezon City
75. A. Mabini St., Ermita, Manila in front of Bayad Center
76. Quirino Ave. Cor. Sargon St., Cluster 1, District 5, Brgy. 707, Zone 78
77. In front of Nitz Restaurant, Leon Guinto St., Corner J. Nakpil St., Manila
78. In front of 2288 Interior 27, Barangay 718
79. In front of 2288 Interior 29, Barangay 718
80. In front of RLMC - left side along P. Antonio St., Brgy. Ugong, Pasig City
81. In front of Lotto Phil., Charity Sweepstakes Office, N. Domingo St., corner Lt. Artiaga St., San Juan City
82. In front of No. 1461 Yakult Main Office, Agoncillo St., corner Escoda St., Manila
83. In front of No. 20 Alberione Center Disciples of the Divine Master along Araneta Ave., Quezon City
84. Across Barangay Hall of Brgy. 696 Zone 76 District V, P. Hidalgo St., Ermita, Manila
85. In front of Shell Gasoline Station, N. Domingo, San Juan City - left side
86. In front of no. 1563-D, (Residential House), Antonio Isip St., Brgy. 814, Zone 88, Dist. V, Paco, Manila (right side)
87. In front of no. 1563-D, (Residential House), Antonio Isip St., Brgy. 814, Zone 88, Dist. V, Paco, Manila (left side)
88. In front of PGA Cars Body and Paint Center, J. Cruz St., Brgy. Ugong, Pasig City
89. Aurora Residences, Quezon City
90. Gen. Malvar St., cor. Guillermo St., Malate Manila in front of St. Paul Malvar Convent
91. In front of Suncoast Brand International Corp. & Covert Garden Inc. J. Cruz St., Brgy. Ugong, Pasig City

92. In front of BPI (Left Side), N. Domingo St., Brgy. Pasdeña, San Juan City
93. In front of Apt Entertainment, J. Cruz St., Brgy. Ugong, Pasig City
94. In front of Metropole Laundry & Dry Cleaners, Valencia St., Quezon City
95. In front of Metrobank, Valencia St., Quezon City
96. Kalentong Market , San Juan City
97. Across No.2266, Bgy. 717, Maginhawa St.,
98. In front of Cosmopolitan Memorial Chapel, G. Araneta Ave., Quezon City
99. In front of 7/11 Convenience Store, Nakpil St., corner Leon Guinto St., Manila
100. In the middle of Purok 1 Alley between Brgys. 704 and 705

Appendix 3.

Top 100 Recommended Sites for Water Sample Collection (Euclidean Distance)

1. In front of Décor Modern Furniture and Lightning along Araneta Ave., corner Palanza St., Quezon City
2. Along A. Mabini St., in front of First Mission Center of Skills Development, Inc.
3. In front of Hotel 2016 Manila, A. Mabini St. Malate, Mla
4. In front of KB24 Restaurant, Ronis Bldg., San Marcelino St., Corner San Andres St., Malate, Manila
5. In front of No. 299 N. Domingo St., (Ongoing Construction), Brgy. Ermitanyo, Quezon City
6. In front of Kasara Resort Residences along Eagle St., E. Rodriguez Ave., Pasig City
7. In front of Unit no. 1 Agoncillo Townhomes, Agoncillo St., Malate, Manila
8. A. Mabini St., Ermita, Manila in front of Bayad Center
9. Quirino Ave. Cor. Sargon St., Cluster 1, District 5, Brgy. 707, Zone 78
10. In front of US Dontics Dental Supply Ramon Magsaysay Blvd., Manila
11. In front of Residential Bldg. beside Beauzitl Bldg., San Andres St., Ermita, Manila (4th Drainage junction box)
12. In front of No. 393 (Residential House), Minerva St., Malacañang, Manila (left side)
13. Across No. 382 (Residential House), Minerva St., Malacañang, Manila
14. In front of No. 378 (Town House), Minerva St., Malacañang, Manila (Circular Drainage Manhole)
15. In front of No. 382 (Residential House), Minerva St., Malacañang, Manila (Catch Basin)
16. In front of 2288 Interior 20, Ramirez Compound, Barangay 718
17. In front of No. 1433 (Residential House), Agoncillo St. Manila
18. In front of No. 1416 (Residential House), Agoncillo St., Manila
19. In front of No. 20 Alberione Center Disciples of the Divine Master along Araneta Ave., Quezon City
20. In front of Jason Store and Eatery, Benitez St., corner Pedro Gil St., Manila
21. In front of Shell Gasoline Station, N. Domingo, San Juan City - left side
22. In front of No. 1108, Yougo Chinese Store, San Marcelino St., corner Natividad Lopez St., Manila
23. In front of Mutsarap Fried, G. Araneta Ave., Sta. Mesa, Quezon City
24. In front of Hizon Bldg., Brgy. 720
25. In front of KB4 Restaurant, No. 961 San Andres St., Malate, Manila
26. In front of The Contemporary Hotel along Araneta Ave., Quezon City
27. In front of No. 10 City Shutter, Valencia St., Quezon City
28. In front of No. 2018 (Private Property), Smith St., Malate, Manila
29. In front of Northpark Noodles, G. Araneta Ave., Quezon City
30. Residential House near Lubiran Bridge
31. Beside Alpan Automotive Auto cars Inc., G. Araneta Ave., Quezon City
32. In front of RLMC - left side along P. Antonio St., Brgy. Ugong, Pasig City
33. In front of House No. 2188 along Asuncion St., District 5, Brgy. 721, Zone 78
34. In front of House No. 2164 along Asuncion St., District 5, Brgy. 721, Zone 78
35. In front of Aling Banang Eatery, N. Domingo St., Brgy., Corazon De Jesus, San Juan City - Circular Drainage manhole without cover
36. In front of no. 2241, Brgy. 720
37. In front of no. 2237, Brgy. 720
38. In front of No. 2280, Brgy. 720
39. In front of no. 2245, Brgy. 720
40. In front of No. 2140 Apostolic Nunciature, Taft Avenue, Manila (2nd drainage from the right of the center gate of the establishment)
41. Residential House perpendicular to Kalinga St., Sta. Mesa Manila
42. In front of No. 21 New Manila Condominium, N. Domingo St., Quezon City
43. In front of BDO-Brixton Hill (Right Side), G. Araneta Ave., Quezon City
44. In the middle of Purok 1 Alley between Brgys. 704 and 705
45. In front of the MRF of Bgy.712, Maginhawa St.,
46. At the right side of No. 408 Pag-asa Steel – Gate 3, Amang Rodriguez Ave., Pasig City
47. Aurora Residences, Quezon City
48. In front of Metrobank, Valencia St., Quezon City
49. In front of Metropole Laundry & Dry Cleaners, Valencia St., Quezon City
50. In front of PICMW along Joe Borris St., Pasig City
51. No. 1411 Mabini St., Ermita, Mla.
52. In front of Intership Navigation Training Center Engineering Laboratory Culinary Center
53. In front of PBCOM, G. Araneta Ave., Quezon City

54. Bagong Lipunan St. cor. Kalayaan St., Bgy. 711
55. In front of a Residential House (no. 740), along Balingkit St., Malate, Manila
56. In front of No. 11 - Residential House, Valencia St., Quezon City
57. In front of No. 7 - Residential House, Valencia St., Quezon City
58. Across Kasaganaan St. cor. Agno St., Bgy.708, Maginhawa St.,
59. In front of Bgy. Hall of Bgy. 708, Maginhawa St.,
60. In front of Salvador S. Anonuevo Grains Retailer, Arellano St., San Andres, Bukid, Manila
61. In front of Unit 16 - Valencia Hills Towers, Valencia St., Quezon City
62. In front of Unit 11 - SEACOM Valencia Hills Towers, Valencia St., Quezon City
63. In front of Unit 20 - New York Laundry, Valencia Hills Towers, Valencia St., Quezon City
64. In front of Panda Construction Supply Inc., G. Araneta Ave., Quezon City
65. In front of No. 19 - Residential House, N. Domingo St., Quezon City (right side)
66. Kalentong Market , San Juan City
67. In front of Gea Mel Eatery, Angel Linao St., corner San Andres, Manila (1st Catch basin with steel grating cover from the right side of establishment)
68. In front of No. 2073 (Residential House), Smith St., Corner San Andres St., Malate, Manila (1st Catch Basin)
69. San Andres St., cor. Mariposa St., Malate Manila across Jollibee Fast Food
70. No. 1444 Mabini St., Ermita, Manila
71. In front of Race World OTB along Rizal Ave., corner S. Herrera St., Manila
72. Along Guerrero St., beside Shalom Hotel
73. In front of Residential House, No. 17-D, G. Araneta Ave., Sta. Mesa, Quezon City
74. In front of Cosmopolitan Memorial Chapel, G. Araneta Ave., Quezon City
75. In front of No. 3, The Laundry House, N. Domingo St., corner Valencia St., Quezon City
76. A. Mabini St. Cor. Sta. Monica St., Ermita Manila in front of Good Rate Money Changer (1st Drainage manhole)
77. Across Sunway International Manpower Services Inc., Smith St., corner Quirino Ave., San Andres Bukid, Manila
78. In front of Mang Inasal, Gelino St., corner Dagupan St., Sampaloc, Manila
79. In front of No. 1972- Int. 8A (Residential House), Mataas na Lupa St., Malate, Manila
80. In front of Super Saver Apartelle, G. Araneta Ave., Quezon City
81. Along Leon Guinto St., cor. Gen. Malvar, in front of Philippine Women's University, School of Fine Arts and Design
82. In front of Los Churreros along Joe Borris St., Pasig City
83. In front of Brgy. 740, Zone 80 (Brgy. Hall), Anak Bayan St., San Andres Bukid, Manila
84. No. 1987 (Residential House) Anak Bayan St., San Andres Bukid, Manila
85. A. Mabini St., Ermita, Manila in front of Calle Bar
86. In front of No. 1461 Yakult Main Office, Agoncillo St., corner Escoda St., Manila
87. In front of no. 1563-D, (Residential House), Antonio Isip St., Brgy. 814, Zone 88, Dist. V, Paco, Manila (right side)
88. In front of no. 1563-D, (Residential House), Antonio Isip St., Brgy. 814, Zone 88, Dist. V, Paco, Manila (left side)
89. In front of No. 2281 (Residential House), Syquia St., Brgy. 874 Sta. Ana, Manila
90. In front of Lotto Phil., Charity Sweepstakes Office, N. Domingo St., corner Lt. Artiaga St., San Juan City
91. In front of Nitz Restaurant, Leon Guinto St., Corner J. Nakpil St., Manila
92. In front of No. 1438 Transorient Maritime Agencies, Inc. Agoncillo St., Manila
93. In front of BPI (Left Side), N. Domingo St., Brgy. Pasdeña, San Juan City
94. In front of Twenty 202 Bldg., Taft Ave., Manila (right side)
95. Gen. Malvar St., cor. Guillermo St., Malate Manila in front of ST. Paul Malvar Convent
96. In front of Suncoast Brand International Corp. & Covert Garden Inc. J. Cruz St., Brgy. Ugong, Pasig City
97. Across Barangay Hall of Brgy. 696 Zone 76 District V, P. Hidalgo St., Ermita, Manila
98. In front of Bgy. Hall of Bgy. 713, Maginhawa St.,
99. Across No.2266, Bgy. 717, Maginhawa St.,
100. In front of OSM Building, Pedro Gil St., Ermita, Manila