2022

A Machine learning model to predict non-revenue water with severely unbalanced classes.

Muriithi, Patrick Kimani
Strathmore School of Computing and Engineering
Strathmore University

A Machine Learning Model to Predict Non-Revenue Water with Severely Unbalanced Classes

By

Patrick Kimani Muriithi

136303

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Technology at Strathmore University

School of Computing and Engineering Sciences

Strathmore University

Nairobi, Kenya

October 2022

Declaration and Approval

I, Patrick Kimani Muriithi declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Student Name: Patrick Kimani Muriithi

Sign:_____ Date: 19/7/2022

Approval

The thesis of Patrick Kimani Muriithi was reviewed and approved for examination by the following:

Dr. Henry Muchiri,
Lecturer, School of Computing & Engineering Sciences,
Strathmore University.

Dr. Julius Butime,
Dean, School of Computing & Engineering Sciences,
Strathmore University.

Dr. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University.

Abstract

Every household, industry, institution, organization needs clean water for existence. In Kenya, water is used for human consumption, production, and agriculture. The consumption of water, therefore, contributes to the overall growth of the economy through water bills.

The term non-revenue water (NRW) is defined as water produced and 'lost' before it reaches the customers. NRW is also described as the difference in volume reaching the final consumer for billing and the initial volume released into the distribution network. Based on the assessment of the Public-Private Infrastructure Advisory Facility (PPIF), an organization that fosters inter-agency cooperation to curbing NRW, physical losses are the main causes of NRW. As per PPIF, most NRW emanates from physical losses, including burst pipes that are often a result of poor maintenance. Besides physical losses, PPIF notes other numerous sources of NRW, especially commercial losses arising from the manner billing data is handled throughout the billing process. The main issues related to this cause include under-registration of customers' meters' reading, data handling errors, theft, and illegal connections. Other causes of NRW include unbilled authorized consumption such as water used for firefighting, utilities for operational purposes, and water provided to specific groups for free. Therefore, non-revenue water risks the country's revenue collection, which can lead to slow economic growth.

This research proposes development of a machine learning model that will be used by water service providers. The model will be able to assist the WSP companies to reduce non-revenue water by predicting water consumption of different customers. To achieve these objectives, we intend to focus on providing tools and methods that will guide the WSPs on reducing the non-revenue water. Our model was trained with 2 years consumption dataset of Nairobi County.

The model developed was able to predict customer monthly consumption with percentage accuracy of 95%.

Table of Contents

Table of Contents

List of Figures

Abbreviations and Symbols

AMR: Automated Meter Reading

L.R: Logistic regression

ML: Machine Learning

NCWSC: Nairobi City Water and Sewerages Company

NRW: Non-Revenue Water

O.S: Oversampling

PPIF: Public-Private Infrastructure Advisory Facility

R.F: Random forest

SCADA: Supervisory Control and Data Acquisition

SMOTE: Synthetic Minority Oversampling Technique

U.N: United Nations

U.S: Under sampling

WASREB: Water Services and Irrigation Board

WSP: Water Service Providers

Definition of Terms

Authorized Consumption: It is water used for firefighting, in utilities for operational purposes, and water provided to specific groups for free.

Class: A list of data set allocation attributes and their values

Distribution network: The entire infrastructure comprising mainly pipes and channels used by water service providers to distribute water to the consumers.

Imbalanced data: A classification problem where classes lack equal representation.

Non-revenue water: Water that has been produced, and 'lost' before it reaches the customers

PPIF: (Public-Private Infrastructure Advisory), an organization that fosters inter-agency cooperation to curbing NRW

Utility Company: Refers to a water service provider.

Acknowledgement

I would like to acknowledge and show my gratitude to my supervisor Dr. Henry Muchiri who made this work possible. His guidance and advice carried me through all the stages of writing my paper.

A special thanks to my family and friends for their continuous support and understanding the moral support was crucial while writing my dissertation. My family has been supportive financially, emotionally and spiritually.

Lastly, I would like to thank the almighty for watching over me and answering my prayer requests. All steps taken and undertaken during this process has been facilitated by the almighty. With God I have overcome all obstacles and managed to deliver.

Chapter 1: Introduction

1.1 Background

Every household, industry, institution, organization needs clean water for existence. In Kenya, water is used in human consumption in production and agriculture. The consumption of water, therefore, contributes to the overall growth of the economy through water bills.

NRW ranks significantly high as one of the main challenges facing water service providers in Kenya. More than a third of all water service providers lost nearly half their volume of water to commercial and physical losses as of 2015(U.N. Habitat, 2015). Based on WASREB estimates, the target for Non-Revenue Water stands below thirty percent, and only eight of eighty-six WSPs met the said target. A simple comparison between the years 2014 and 2015 indicates an increase in NRW, which outweighs a concurrent marginal increase in water production during the same period. Such results indicate less volume of consumption water despite the increase in production, which amplifies the imminent threat of NRW to water services providers and water consumers.

Current statistics indicate that 45% of the water produced in the urban sector is lost through NRW, translating to losses over KES 10 billion annually (U.N. Habitat, 2015). Such crippling losses have continued to mar the operations and sustainability of most water service providers in the country. In the large cities, the astronomical losses have made it hard for WASREB to meet the consumer demands by crippling the board's ability to repair, operate, expand and set up a new water supply network. Effectively, this raises the tally of people without access to clean water, with the current number standing at 16 million people (Ng'etich, 2015). In Nairobi, the local water utility Nairobi City Water and Sewerage Company (hereafter called the water utility) struggles to distribute scarce water resources and meet a rapidly increasing demand.

Common methods applied in detecting non-revenue water by utility companies include manual methods that include physical investigation teams by utility companies and police, transient test-based methods, and pressure sensor placement algorithm. These methods are less effective in developing countries like Kenya.

Utility losses remain a key challenge in the utility industry. Utility losses pose a significant threat to the operations of the utility and the overall long-term sustainability and profitability. As a result, Barandouzi et al. (2012) argue that utility companies have turned towards machine learning algorithms and advanced analytics to monitor consumption patterns that allow easy identification of utility theft. However, this shift is not without its share of problems. In the quest for advanced utility consumption management, one key challenge that arises is the vast array of data produced and the vastness of its distribution. However, when properly adopted and used, ML algorithms can significantly cut on the above challenge by cutting on fraudulent transactions to around 1-2% of the total number of observations, thereby reducing data imbalances. ML algorithms also reign on sudden changes in the normal transactions, such as sudden drop/increase in consumption amount, according to Soldevila et al. (2017). Nonetheless, ML algorithms tend to produce unsatisfactory classifiers when faced with imbalanced data sets and sudden drop/increase in consumption.

1.2. Problem Statement

On a site visit in Nairobi water, the non-revenue water department manager made it clear that NRW was a menace that was causing over 40% revenue loss in Nairobi County only. Nairobi water was using manual excel computation to compare between the forecasted consumption with the actual consumption. This thesis proposed a better and accurate way to carry out the comparison.

Besides the manual methods applied by the water utility companies in Kenya, many computational techniques can be used in NRW detection, for example, Bayesian system identification and machine learning. However, they are less effective since they highly depend on how balanced the data is, for they learn by example though most of the real-world problems have imbalanced data classes. An imbalanced data in this context refers to a classification problem where classes lack equal representation.

In machine learning, imbalanced classification implies a predictive modeling problem where for each class label; the number of examples in the *training dataset* is not *balanced*. Therefore, as Soldevila et al. (2017), states, it means that *class* distributions are never close to equal but rather skewed and biased. For instance, we can talk about a

few theft cases in water utility consumption compared to authorize water consumption. A limitation can also cause an imbalance in data collection, such as cost, privacy, and the effort required in data collection and representation. The proposed ML approach solves this problem through the identification of the rare minority data class therefore, achieving higher overall accuracy and solves missed classification due to sudden changes in the consumption patterns of the water utility.

1.3.Objectives

1.3.1. General Objectives

The main objective is to develop a deep learning model using a convolutional neural network used by water utility companies to cub the NRW.

1.3.2 Specific Objectives

1) To critique non-revenue water detection techniques.
2) To design and implement an ML model using Logistic regression (L.R.).
3) To evaluate the performance of the Model.
4) To provide an accurate process of comparing between the actual consumption and the forecasted consumption.

1.4.Research Questions

1) What are the methods used to forecast consumer consumption volumes?
2) What are the means of collecting and processing utility consumption data to be used in the analysis and detection on NRW?
3) Which are the ways applied to reduce the non-revenue water?

1.5. Justification

The problem of NRW detection in a utility company has been studied over the past decades, and different researchers have sought several solutions. However, there is no direct and cheap method to control the increasing rate of water loss, especially in developing countries due to poor water distribution networks and imbalanced data in most real-world problem cases.

In this study, many benefits will arise from the NRW detection model; water utility companies will reduce the non-revenue water hence expanding and improving their services. This study provides a better and affordable solution that can be used across the world. Compared to other methods that have been used in the detection of apparent water loss, using machine learning is better because:

1) The tool is reliable and affordable by any utility company because its independent of the distribution network, which is the biggest challenge in NRW detection of developing countries (Most developing countries have disorganized distribution network)
2) The study pays attention to the losses at the side of the consumers that is leakages after the meter.
3) The tool reduces the time for information access to utility companies since they don't have to wait for reports from investigation teams and well-wishers.
4) The tool is independent of data imbalances as compared to Bayesian system identification.

The study will provide policy makers and management of water service providers with a viable option to curb NRW. It will significantly change how NRW is viewed as well as measures to adopt to reduce it and thus protect the overall viability and sustainability of such enterprises. In The study will focus on apparent water losses, which have not been looked at in most research in the last decade, particularly in the Kenya.

1.6. Scope and Limitation

Detection of NRW is an extremely challenging problem in today's water distribution network and utility companies' everyday operations. NRW comprises physical and apparent water losses, both of which greatly affect the revenue of utility companies worldwide. NRW losses have also made it difficult for utility companies to meet their consumer demands and keep water tariffs reasonable. This dissertation presents a generalized machine learning approach focused on using customer water consumption data patterns to detect illegal water consumers, malfunctioning meters, and leakages after the meter. The scope of this study will be limited to NRW.

Chapter 2: Literature Review

2.1. Introduction

Many approaches have been done in an attempt to solve the problems of NRW losses faced by many of the utility companies all over the world. Review of previous research conducted on reducing non-revenue water is done on this chapter, the methods and approaches adopted by WSPs, and how this knowledge will be used in the development of a deep learning model using a convolutional neural network that water utility companies can use to cub NRW. The chapter will be divided into four sections namely theoretical framework, empirical framework, conceptual framework, and research gap.

2.2 The Water Balance Theory

According to Mastaller and Klingel (2017), understanding the concept of NRW calls for a purposeful and objective view of the water system. Over the years, different countries have come up with various strategies to help achieve this clarity. In the US, the concept of a water audit is used to understand the cost source and magnitude of NRW (Radivojević et al., 2020). The International Water Association (IWA) has developed the water balance concept and which has become adopted by many national water associations across the world.

The water balance theory is based on the concept of water balance developed by IWA. According to Charalambous and Hamilton (2015), understanding the concept of water balance is the foundation on which water losses management techniques are based. Based on this theory, water consumption in a utility system comprises water losses and authorized consumption. Authorized consumption is further broken down into unbilled authorized consumption and billed authorized consumption.

Water losses according to the theory are also categorized as commercial (apparent losses) and real losses. As Mastaller and Klingel (2017) point out, the theory thus holds that non-revenue water is the sum of real losses, commercial losses, and unbilled authorized consumption while revenue water comprises only billed authorized consumption. Therefore, according to IWA, the calculation for NRW can be achieved as follows:

NRW= System Input Volume- Billed Authorized Consumption

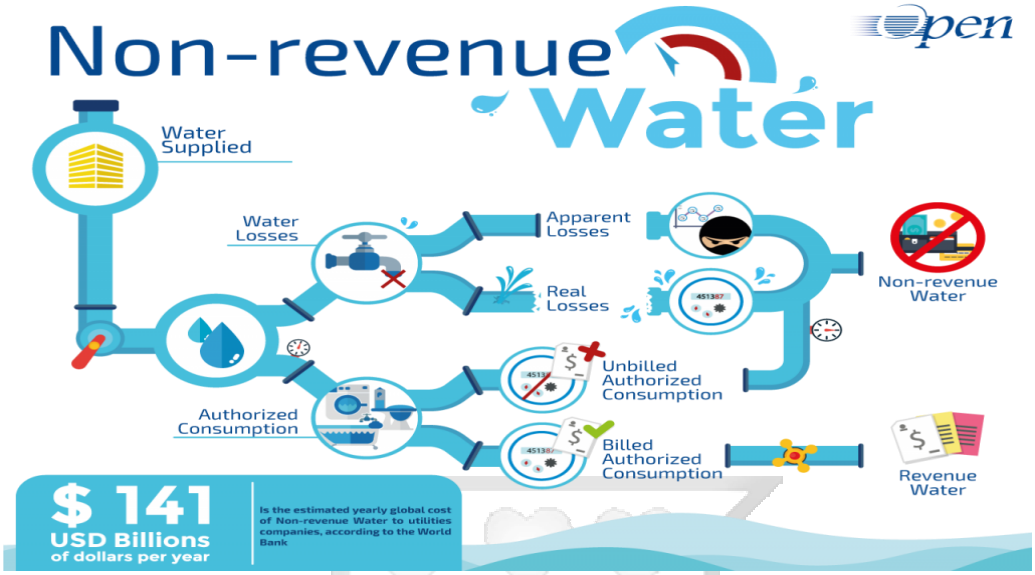


Figure 2.1. Diagrammatical explanation of NRW, its causes, and the estimated value loss globally per year (Liemberger et al., 2006)

2.2.1 Determining System Input Volume

Liemberger et al., (2007) opine that system input volume is equal to the total value of metered system input recorded in the entire system. Meter accuracy should be thoroughly verified by ensuring that discrepancies between temporary measurements and meter readings are addressed and adjustments made to reflect the real situation.

2.2.2 Determining Authorized Consumption

Based on IWA, authorized consumption is defined to include water used in flushing of mains and sewers, filling water tankers, firefighting, and water taken from hydrants. Based on the local practices, these can be considered as billed, metered or unmetered, or unbilled.

2.2.2.1 Billed Metered Consumption

Billed metered consumption must put into consideration data handling and billing errors. The information to calculate the annual billed metered consumption is obtained directly from the billing system with special consideration put into identifying the different consumer categories.

2.2.2.2 Billed Unmetered Consumption

According to Charalambous and Hamilton (2011), this is also obtained from the billing system through identifying and monitoring unmetered domestic water consumers for particular periods.

2.2.2.3 Unbilled metered Consumption

Radivojević et al., (2020) argue that unbilled metered consumption forms an insignificant part of the water balance. Determining the volume uses the same method as used in determining the volume of billed unmetered consumption.

2.2.2.4 Unbilled Unmetered consumption

It marks water used for operational purposes by the utility. According to Mastaller and Klingel (2017), this consumption is overestimated by considering it as a percentage of the overall system input volume or by overestimating it on purpose to cover losses incurred. Unbilled unmetered consumption should be estimated for individual consumption.

2.3 Empirical Review

Numerous studies have been conducted on NRW in an attempt to find a solution to tackle the problem. While significant strides have been made towards reducing water losses, NRW remains a challenge in most countries. According to Amoatey et al., (2018), the problem is exacerbated by poor equipment and infrastructure unmaintained for long periods, increased water fraud, and illegal connections to old buildings.

WSPs in many low-income countries have attempted to reduce water losses by implementing strategies to reduce NRW. According to Kanakoudis and Tsitsifli (2010), the main challenge faced by many of the water utility companies is the reliability of their measurement systems. Further Amoatey et al., (2018), support this finding arguing that the reliability of the measurement system is crucial as it is essential in the evaluation of losses which has a significant impact on the water balance.

A 2010 study on non-revenue water management conducted in Kampala, Uganda found that dilapidated infrastructure was the key cause of NRW (Mons, 2010). A similar study was conducted in 2008, in Accra, Ghana. The study was on the management of NRW

where the main objective was analyses of existing NRW management strategies. After the study levels of NRW were found to be determined by meter under-registration, leakage control, and meter reading inaccuracies, network pressures, and billing inaccuracies (Yeboah, 2008). A 2003 study in the US, at the Tampa Water Department of Florida, sought to identify the causes of an eight-month increase in NRW. The study identified meter data accuracy, meter calibration, and testing, billing system, and accounting procedures as important areas of study (Brian, 2003).

2.3.1 Current NRW Detection Techniques

Accurate, quick, and affordable water loss detection is the key to mitigating all problems and consequences associated with non-revenue water losses. At times some of these techniques can be used concurrently to get better results.

2.3.1.1 Transient Test-Based Techniques (TTBT)

This technique solves the problem of pointing out illegal branches in the water distribution network utilizing steady-state pressure and discharge measurements mainly because such systems are not active according to the regular schedule (Meniconi et al., 2011). Illegal branches can wipe out a large volume of water in the distribution network, which negatively affects the utility company's revenue.

As Atef et al., (2016) point out, illegal branch connections in a water distribution network can be located by analyzing pressure signals through wavelet analysis and branch size, which is related to the reflection coefficient of the connection of the illegal branch (Xin et al., 2014). A case study carried out at the Water Engineering Laboratory of the University of Perugia, Italy, showed that TTBT allows detection of illegal water branches in the distribution network irrespective of whether they are active or inactive. TTBT water theft detection technique is advantageous because the possibility and precision in the localization do not depend on branch functionality conditions. Furthermore, the evaluation of the reflection coefficient at the branch allows the evaluation of the branch size.

2.3.1.2 Meter Tampering Detection Techniques

Several definitions exist for the term meter tampering depending on the different contexts of use. In the case of the water service provision, meter tampering is generally defined as the fraudulent, deliberate manipulation of water reading meters to interfere with consumption data. Ogutu, Okuth and Lall, (2017) support this view stating that meter tampering implies a discrepancy between the actual volume of water consumed and the volume recorded in the utility company's meter readings.

Essential meter manipulation indicates weak consumption control mechanisms, thus resulting in lost revenue. To overcome meter tampering using the ML algorithms, data from the utility company's databases is mined to generate and programme the algorithms (Soldevila et al., 2017). The generated algorithms mainly aim to detect the main types of consumption patterns: Sudden drops, abnormally low consumption, and progressive drops.

2.3.2 Using machine Learning (ML) techniques to detect utility theft

ML is a sub-field of artificial intelligence that deals with the design and development of algorithms that seek to create self-teaching or learning computers. The learning process involves the automatic extraction of rules and patterns from data. ML is closely related to fields like data mining, statistics, and pattern recognition (Hu et al., 2021).

A large number of ML algorithms have been used in utility theft detection modeling. Li et al. (2015) classify them as supervised, semi-supervised, and unsupervised learning methods. A supervised ML algorithm's task is inferring a function from labeled training data. Unsupervised learning is an ML task that draws inferences from data sets consisting of input data without labeled responses (Gao and Ren, 2014). Semi-supervised learning uses both the concepts of supervised and unsupervised learning wherein, in typical circumstances, small volumes of labeled data are combined with a large volume of unlabeled data.

One study in Egypt on NRW stands out, the study uses neural networks to capture the parameters affecting non-revenue water in Egypt (Mona Rafat Elkharbotly...2022). This study aims to offer findings that will lay the basis for sustainable water resources management in Egypt.

2.4 Conceptual Framework

A key measure used to determine the efficiency of any WSP is the extent of NRW reported. Every water utility company strives to reduce the level of NRW as lower levels are considered desirable and are achievable through concerted input including reducing water theft, eliminating billing and accounting errors, and rehabilitating the infrastructure. Through a measured and coordinated interaction of these factors, it is possible to develop a deep learning model using a convolutional neural network that water utility companies can use to cub NRW

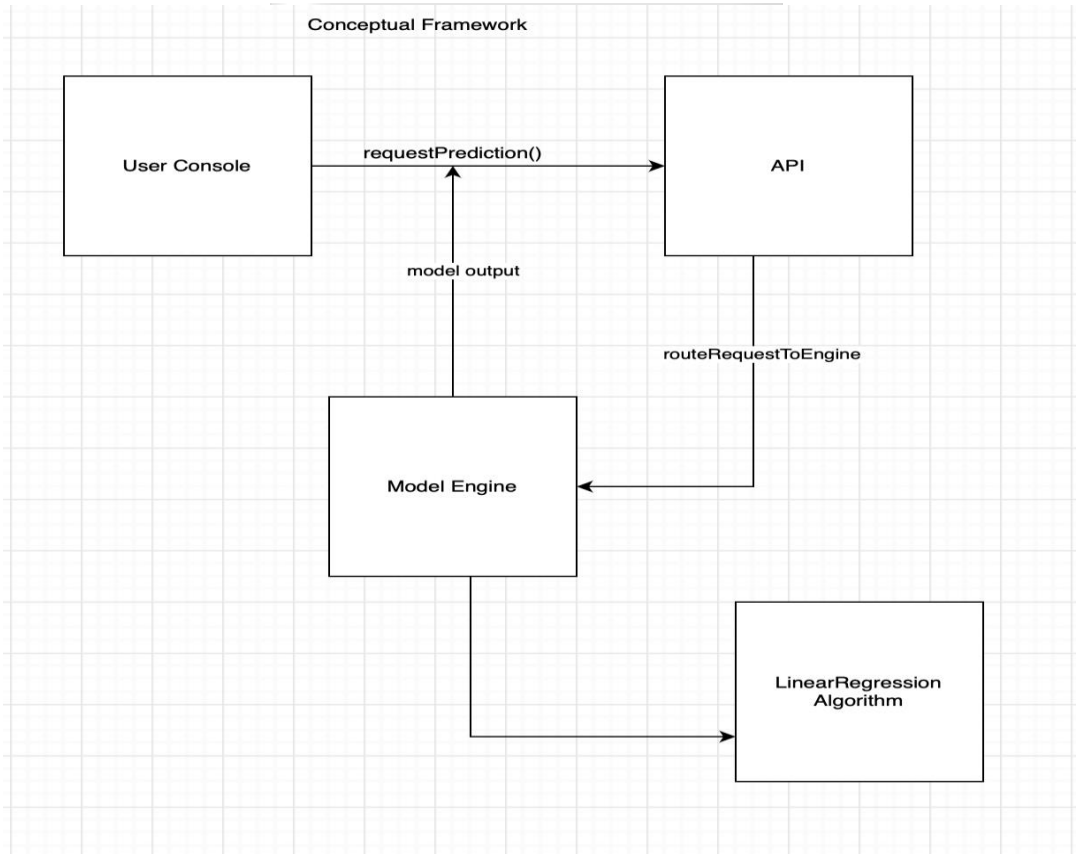


Figure 2.1. A Diagram showing conceptual framework.

2.5 Research Gap

Besides the numerous studies conducted on curbing NRW, the problem persists and plagues many water utility providers globally. The problem is enabled in part by the current techniques used to curb NRW. As identified early, most of the techniques are manual e.g. using excel in consumption comparison which puts them at the risk of inaccuracies, manipulation and decreases their efficacy in detecting and curbing NRW. Thus this remains a major gap as water providers are still experiencing the challenge of non-revenue water, Nairobi water being a case study. The proposed solution which seeks to use machine learning thus seeks to solve this knowledge gap.

Chapter 3: Research Design and Methodology

3.1 Introduction

This section explains the research design and methodology used for this research project. Combinations of experimental design and historical research model have been adopted as the research design for this research project. Rapid application development (RAD) will be used as the system development methodology.

3.2. Research Design and System Development Approach

In this study, we will use multiple research methods. This research method integrates both quantitative and qualitative research strategies within one single project. It takes advantage of using more than one research method as a way to explore a research problem. It also uses mixed data like text, numbers, and additional means involving statistics and text analysis.

3.2.1 Quantitative method

Xin et al. (2014) opine that the quantitative method ought to emphasize objective measurements and analyze the data collected in a mathematical, statistical, or numerical manner. The data can be collected through questionnaires, polls, and surveys or by manipulating pre-existing statistical data using computational techniques. A similar point of view is held by Farley et al. (2008), who states that the main focus of quantitative research is collecting and generalizing numerical data to explain a certain phenomenon or across categories of people.

Descriptive research design will be used in this project. This research design will involve collection of data from WSPs, analysis and interpretation of the data to use it as the basis of building a convolutional neural network model to solve the identified problem.

3.2.2 Qualitative method

The quantitative method is a scientific research method that involves investigating to obtain answers to a certain question. The quantitative research method is systematic. It follows pre-laid procedures to collect evidence and develop findings that previously did not exist and apply to a wider scope other than the one defined by the current study.

For the purpose of this research project, a historical research model will be used. Historical research model uses past events in order to understand present patterns and predict future occurrences. Through a historical model, the convolutional neural network model will be trained with past data obtained from the WSPs to identify current water usage patterns and these can be utilized to reduce NRW. Based on the past consumption data, the model will also be trained to make predictions on what to expect in future in terms of consumption patterns which will make it possible to keep track of NRW.

3.2.3 Why use the mixed research method

Many reasons have been discussed for using mixed-method research, and these include:

1) Weakness in quantitative and qualitative research. Imauzumi and Junkosha (1987) aver that quantitative research experiences challenges in understanding the data setting or context from which it is collected. Qualitative research is deemed weak in its parts due to its susceptibility to biases and its lack of support for statistical analysis and generalization. The use of mixed methods is thus preferred as it offsets these shortcomings by allowing a researcher to carry out analysis and exploration in the same study.
2) As mixed research methods are multifaceted, researchers can use various tools to gather data that is comprehensive. The use of a variety of tools thus provides a broader perspective of the overall research problem.
3) Mixed research methods provide a broad scope of results since they can be either statistical analysis or observations. As Wu and Sage (2008) assert, the broadness of the results means that the results can be validated within the study through additional evidence and support for the findings.
4) The use of both inductive and deductive thinking and reasoning allows both numbers and words for communication. Hence, the results and findings can be understood by a wide range of audiences.
5) The combined methodology is a means to minimize the researcher's personal bias.

3.3 System Analysis and System Design

The prototype will be developed using rapid application development approach. RAD software development methodology that uses minimal planning in favor of rapid prototyping.

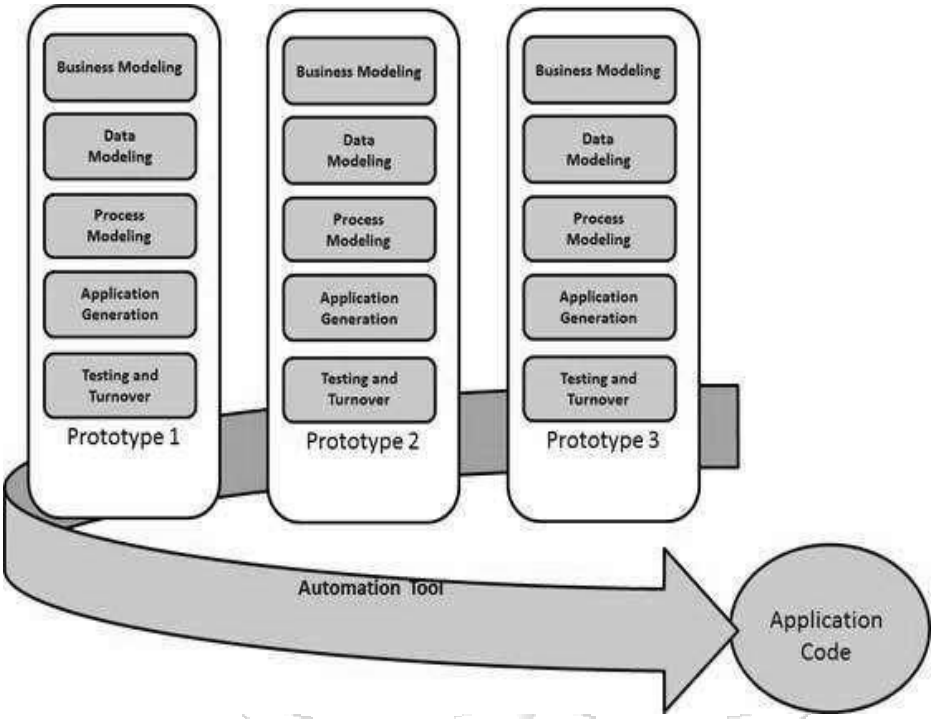


Figure 3.1. An illustration describing the RAD model in detail

3.4 System Implementation

Linear regression Machine learning models were developed and trained using the same data set to evaluate its accuracy. These models include;

1) **Logistic Regression (LR).** The model is used when the dataset to be analyzed with one or multiple independent variables likely to influence the outcome. According to Depuru et al. (2011), a dichotomous variable is used to measure the outcome whereby the possible outcomes are only two. Similarly, Almeida et al. (2014) aver that logistic regression is applied in predicting binary outcomes such as Yes/No, 1/0 given a set of independent variables. Given logistic regression is

a form of regression analysis, it is a predictive analysis. According to Depuru et al. (2011), logistic regression can be used for data description or explaining the relationship between a dependent binary variable and one or more interval, nominal, ordinal, or ratio-level independent variables.

3.5 System Testing

Consumption data of customers over a period of two years will be ingested into the database. The model engine will consume the data split it into training and testing datasets then run the modeling. After which a frontend information system will be used to send prediction request to the engine. Furthermore, an investigation team will follow up on customers who have been classified as fraudulent customers all while using the model to see whether they are detected or flagged.

3.6 Target Population and Sampling

The target population for this research project will be customers of the NCWSC, a WSP in Kenya. The customers whose data will be collected will include all customers who are known to consume water legally, those who are known to consume water fraudulently as well as entities or customers who consume water for free but legally.

3.7 Data Collection

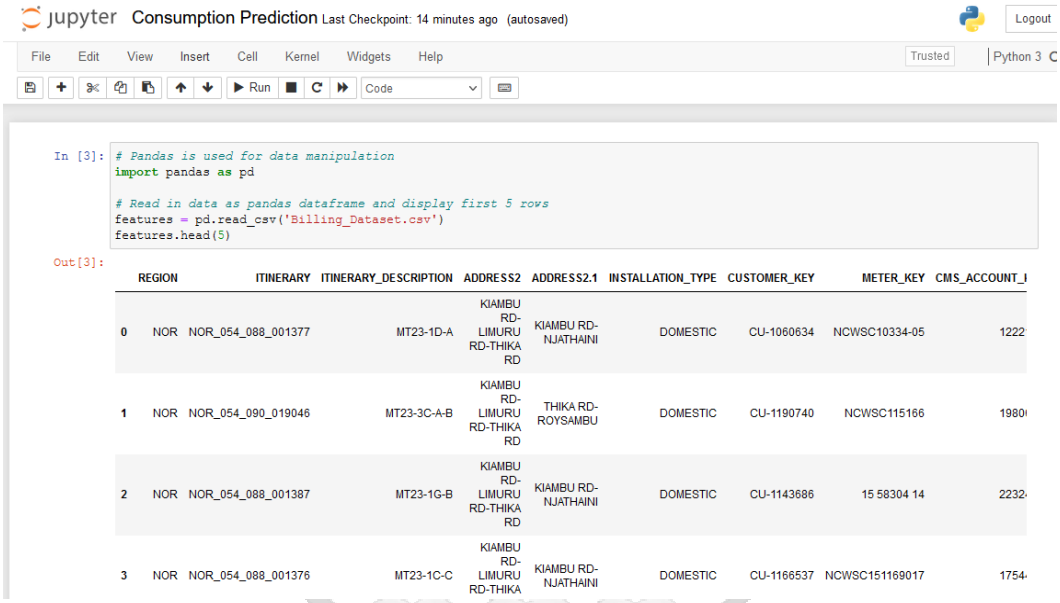
Consumption data for customers in different blocks around Nairobi will be extracted from the billing system at Nairobi City Water and Sewerage Company (NCWSC); this will include all consumption details for all customers; that is, customers that have been caught in the act of using water illegally and those that have never been caught before. Data protection act will be adhered to where the dataset shared will only be used purposely for research purpose.

A list of customers that have been caught in the act before will be extracted from the illegal use management unit of NCWSC plus the time when they were caught. Also, customer GPS coordinates data will be extracted from the GPS mapping systems at NCWSC.

3.8 Data Analysis

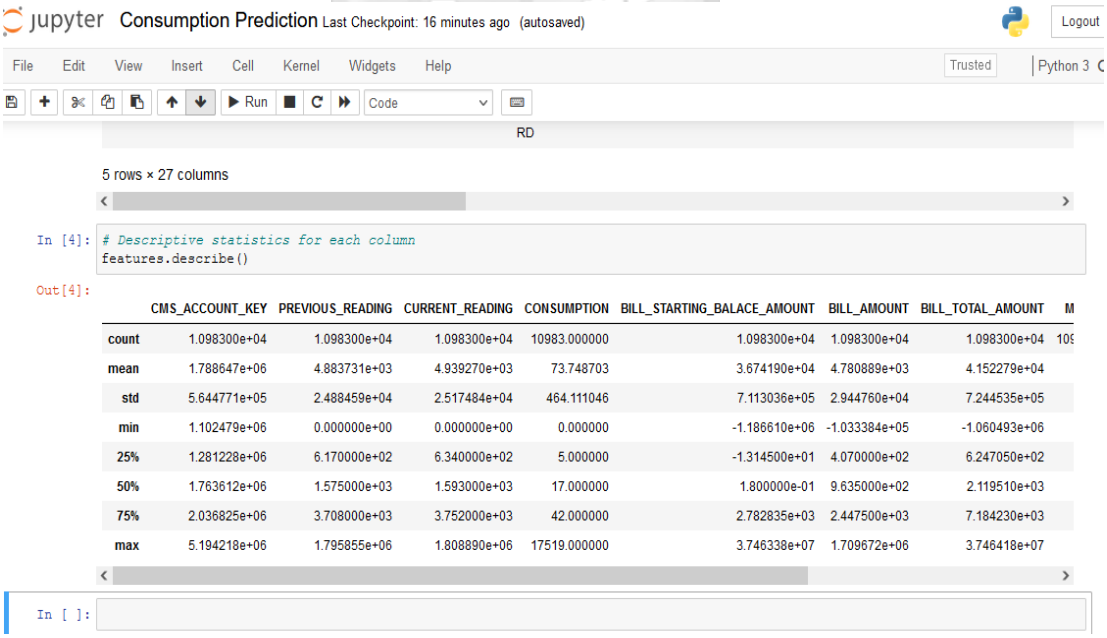
Before the consumption data is used for building, training, and evaluating the model, it will go through the following data processing criterion to make it more suitable for ML algorithm training and better classification results.

1) Establish data collection mechanisms. Since data is extracted from different systems, SQL Queries will be developed depending on the customer reference number as a unique I.D. to each consumer. Here we expect three (3) files (that is, consumption details' file, GPS coordinates' file, and the file for customers who have ever been caught in the act of using water illegally before), each containing a customer reference column which will be used to merge all the files into one file. The merging will be done using python scripts.
2) Format data to make it consistent. Since the data will be aggregated from different sources or databases that different people may manually update, it's worth making sure that all variables within a given attribute are consistently written. Therefore, we have to make sure that the input formats are the same across the entire dataset and our resultant file is in CSV format. This will be implemented using python scripts.
3) Data reduction. Here we will simply remove records (data objects) with negative consumption values, fewer representatives, and records without GPS coordinates are removed to make predictions more accurate. This will also be done using python and excel scripts.
4) Data cleaning Missing values can tangibly reduce our prediction approach; we will substitute them with zero (0) for consumption amount or with consumption mean values.



	REGION	ITINERARY	ITINERARY_DESCRIPTION	ADDRESS2	ADDRESS2.1	INSTALLATION_TYPE	CUSTOMER_KEY	METER_KEY	CMS_ACCOUNT_I
0	NOR	NOR_054_088_001377	MT23-1D-A	KIAMBU RD-LIMURU RD-THIKA RD	KIAMBU RD-NJATHAINI	DOMESTIC	CU-1060634	NCWSC10334-05	1222
1	NOR	NOR_054_090_019046	MT23-3C-A-B	KIAMBU RD-LIMURU RD-THIKA RD	THIKA RD-ROYSAMBU	DOMESTIC	CU-1190740	NCWSC115166	1980
2	NOR	NOR_054_088_001387	MT23-1G-B	KIAMBU RD-LIMURU RD-THIKA RD	KIAMBU RD-NJATHAINI	DOMESTIC	CU-1143686	15 58304 14	2232
3	NOR	NOR_054_088_001376	MT23-1C-C	KIAMBU RD-LIMURU RD-THIKA	KIAMBU RD-NJATHAINI	DOMESTIC	CU-1166537	NCWSC151169017	1754

Figure 2.2. Python Excel scripts showing data processing before data analysis



	CMS_ACCOUNT_KEY	PREVIOUS_READING	CURRENT_READING	CONSUMPTION	BILL_STARTING_BALACE_AMOUNT	BILL_AMOUNT	BILL_TOTAL_AMOUNT	M
count	1.098300e+04	1.098300e+04	1.098300e+04	10983.000000	1.098300e+04	1.098300e+04	1.098300e+04	109
mean	1.788647e+06	4.883731e+03	4.939270e+03	73.748703	3.674190e+04	4.780889e+03	4.152279e+04	
std	5.644771e+05	2.488459e+04	2.517484e+04	464.111046	7.113036e+05	2.944760e+04	7.244535e+05	
min	1.102479e+06	0.000000e+00	0.000000e+00	0.000000	-1.186610e+06	-1.033384e+05	-1.060493e+06	
25%	1.281228e+06	6.170000e+02	6.340000e+02	5.000000	-1.314500e+01	4.070000e+02	6.247050e+02	
50%	1.763612e+06	1.575000e+03	1.593000e+03	17.000000	1.800000e-01	9.635000e+02	2.119510e+03	
75%	2.036825e+06	3.708000e+03	3.752000e+03	42.000000	2.782835e+03	2.447500e+03	7.184230e+03	
max	5.194218e+06	1.795855e+06	1.808890e+06	17519.000000	3.746338e+07	1.709672e+06	3.746418e+07	

Figure 3.3. Data description using jupyter

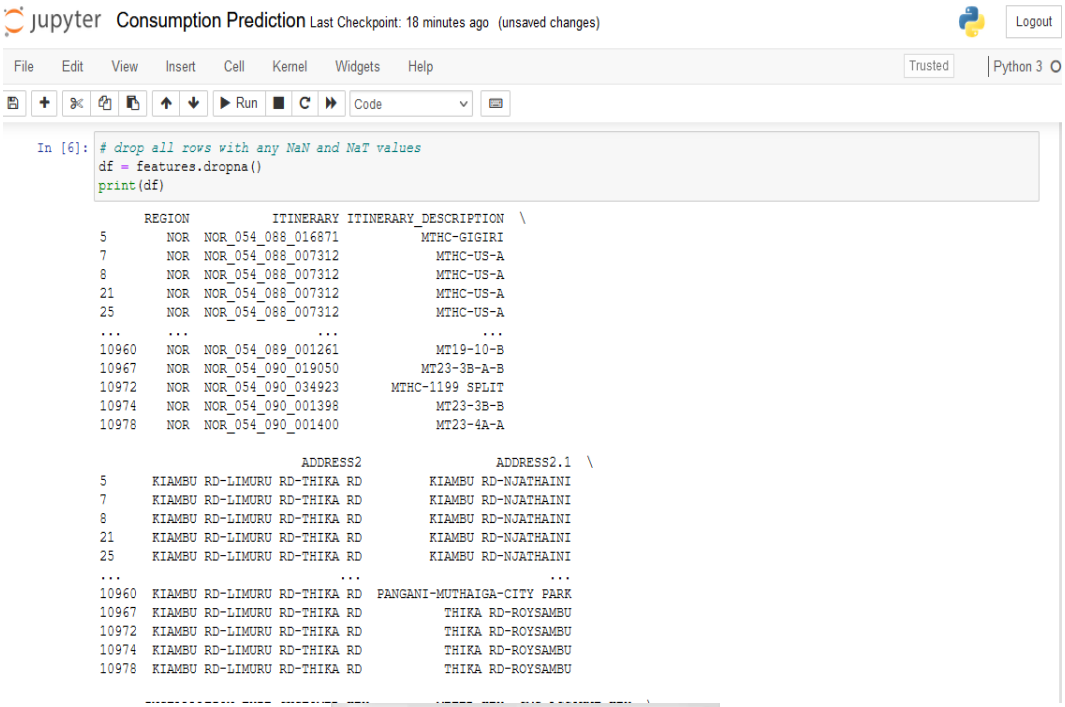


Figure 3.4. Drop null columns in data cleaning

Data decomposition; Here we will add extra data columns to cater to the seasons and classes (honest or non-honest utility consumers). And other factors that affect the consumption of water, like the existence of other water sources in the area and levels of development.

5) Data re-scaling the data will be normalized or standardized before applying it to ML model training to reduce the dimensions and avoid the situations of some consumption values over-weighting others. This will be done during model development.
6) Discretize data. Here data will be divided into a range of months to determine the consumption details according to seasons. This will then give more effective predictions

3.9 Research Quality

The performance the ML model will be assessed using mean square error and mean absolute error.

3.10 Ethical Approval

Ethical approval was sought from the Strathmore Ethical Approval Committee for the approval of this research project. Approval reference number was SU-IERC1226/21. Data collection permits will also be sought from NCWSC allowing for the collection of data from NCWSC.

Chapter 4: System Analysis and Design

4.1 Introduction

In this chapter, the overall architecture and detailed design of the system are described while considering various requirements. The chapter depicts the overall system architecture, describes the various components making up the system, and succinctly the interactions between users and the system components.

The chapter relies on various design diagrams, which are developed to achieve the illustrations, including the use case diagram, class diagrams and sequence diagram.

Based on the earlier identified research gap, the system is intended to work by considering several aspects of non-revenue water, among them leakage location, the difference in customer consumption, illegal connection hotspots, and volume of authorized non-revenue consumption. The system is designed to capture consumption data automatically to avoid human manipulation, use machine learning algorithms to analyze the data, and accurately determine the location. The algorithm is trained to use obtained data to detect and determine the volume of non-revenue, identify NRW patterns, identify the corresponding cause of the NRW and accurately identify the zone of the NRW consumption.

4.2 Requirement Analysis

The current research seeks to utilize the efficiency of machine learning to develop a model capable of detecting non-revenue water, thereby reducing water loss and revenue loss for water service providers while increasing the reliability of water supply across most cities. Based on the earlier established objectives, the system needs to fulfill several requirements if the proposed model is successful. From an interview done with Nairobi Water and Sewerage Company, it came out clearly that non-revenue water was an issue that needed to be addressed. The client provided the following requirements.

4.2.1 Functional Requirements

1. The system should capture consumption data.
2. The system should analyze obtained consumption data.

3. The system should allow retrieval of historical data, patterns, analysis, and reports per consumer, NRW type, and location and consumption type.
4. The system should provide data security, redundancy, and integrity by ensuring data is not easily manipulated, thus guaranteeing the model's effectiveness.

4.2.2 Non-Functional Requirements

i. The system interface should be easy to use with proper UI/UX designs
ii. The system should provide requested data and reports and perform functions requested in less than 5 minutes.
iii. The system should have minimal points of failure and should provide ease of recovery

Based on the research findings analyzed in the literature review, NRW remains a significant problem plaguing WSPs globally, even in developed countries. The problem is significantly rampant in developing countries where factors such as dilapidated water distribution infrastructure, illegal connections, and lack of an effective system to track water consumption only serve to exacerbate the problem. The current system is designed with all these factors in mind by ensuring that proper system analysis techniques are applied when selecting the most appropriate system design and architecture.

4.3 System Architecture

The system is built to collect raw consumption data from various consumption sources and store it in a CSV file. Once the raw data is obtained, it undergoes pre-processing to remove redundancies and noise in the data, including incomplete values and false results. The cleaned data is then classified based on attributes such as users, location, and consumption volume, among other attributes. Once the data is classified, the algorithms allow data manipulation to identify patterns, inconsistencies, and variances that provide key insight into water consumption and NRW. The results of the manipulation operation are saved into the database automatically, and the user is prompted to select whether they need to generate reports. Once the reports are generated, they are saved into the database for faster retrieval in the consequent operations

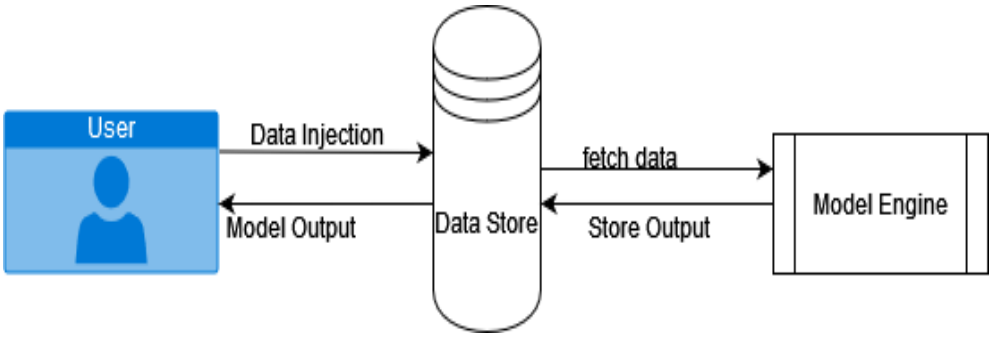


Figure 4.1. An illustration of the system architecture

4.4 Use Case Diagram

The interaction between users and the system is illustrated with a use case diagram. The users of the system referred to as actors interact with the system by performing certain queries in the system and the system responds by giving out results to the queries. Use case diagrams also perform an important role of showing the functionalities that a system should have and operations that the user can perform on the system.

There are several use cases that the identified model will fulfill. Below are the use cases as illustrated in the diagram above.

1. Collect consumption data- water consumption data is collected and stored in the data store.
2. Data Cleaning and manipulation- The user defines and sets the parameters used to clean, classify, authenticate and manipulate captured data.
3. Reports- Obtained by performing a search, defining search parameters, and identifying the type of report to be returned after the search.
4. NRW Identification- The specific type of NRW is obtained through analyzing consumption data, identifying patterns, and comparing them to present data for each type of NRW. The location of the NRW is obtained through filtering location consumption data.
5. Modify- This allows the user to add users, locations and define new consumption parameters and new consumption types.
6. Train model- The user trains the model using training data.

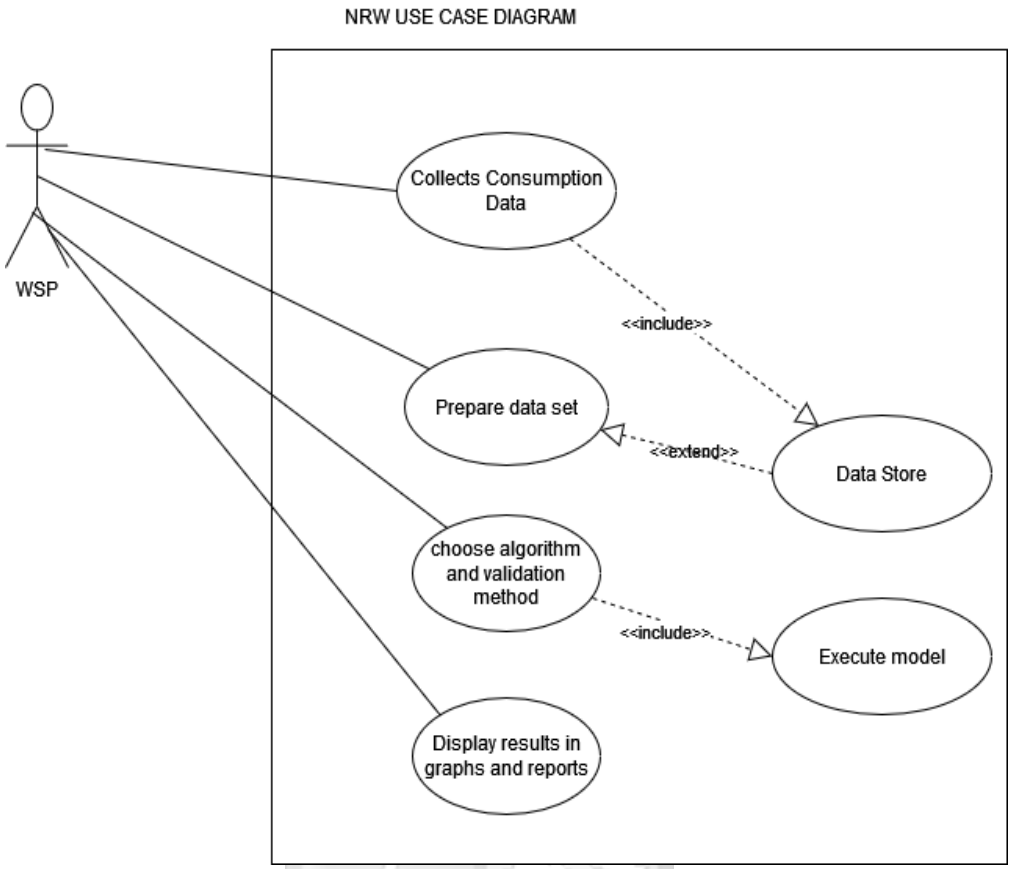


Figure 4.2. A representation of the NRW use case Diagram

4.5 Sequence Diagram

Sequence diagrams illustrate the message interactions between objects in a system. In a system, objects interact by exchanging messages with each other, represented by horizontal arrows in sequence diagrams indicating the flow of messages from the sender to the recipient. The interaction between the administrator and the various objects is illustrated in the following sequence diagram. The administrator starts by defining the parameters for data collection and how the data is to be stored in the database. Once the data is captured and stored, the administrator defines the data cleaning and manipulation parameters. The result of the process is then stored in the database. Next, the administration queries the database by inputting specific search parameters or keywords. The search results are returned to the administrator, after which they are asked to select the format they want the result displayed. The returned results are then saved into the

database. If further action on the results is needed, the administrator proceeds to enter the query while saving the results to the database.

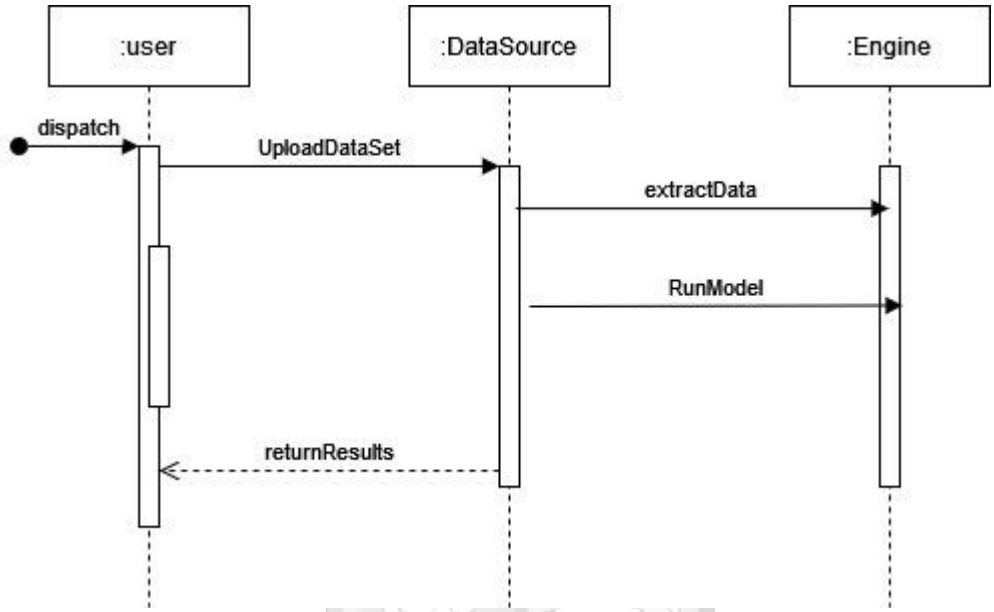


Figure 4.3. Sequence Diagram illustrating message interactions between object in a system

4.6 Database Schema

It defines the relationship among entities and the actual entities stored in the database. In other terms, a database schema is a diagrammatic description of the database using schema diagrams. The user type table contains information about different types of water consumers. The NRW type contains information about the different types of NRW. The user type table has a relationship with the NRW type table.

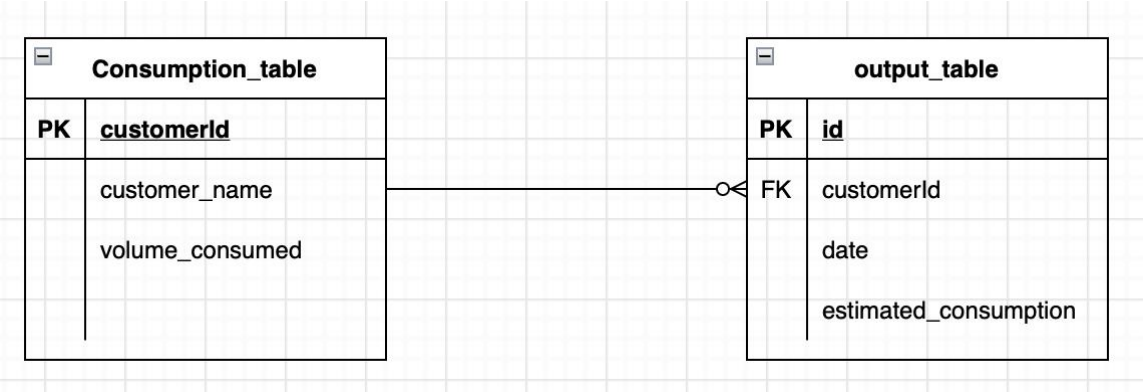


Figure 4.4. An Illustration of the proposed Database schema

Chapter 5: System Implementation and Testing

5.1 Introduction

The software, documentation, and operating procedures of a system are developed and tested during the implementation stage. This section explains how the model was developed, tested, and validated. It begins by detailing how the set covering the problem model was developed. The model is then tested to see if the output values are correct. Experiments mentioned in Chapter 3 were used to establish the ideal arrangement of features to further confirm the researcher's technique. The model's application in detecting causes of NRW is described in the chapter's last part.

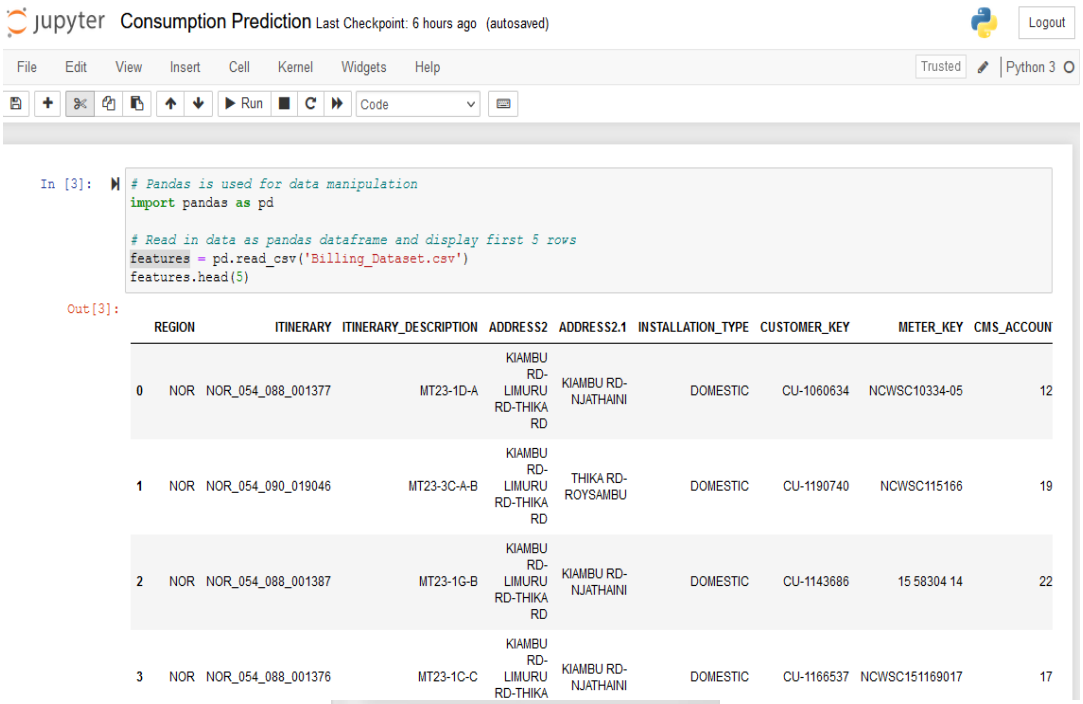
5.2 Hardware and Software Environment

The model was built using a linear regression statistical model, a data science statistical technique. Jupyter tool with the aid of python, pandas, and sklearn library was utilized to manipulate and analyze data. It has data structures and methods that may be used to manipulate numerical tables and time series. Because the requisite libraries were readily available, (Python) was chosen as the model's default programming language. Below is a highlight of the hardware and software specifications used for the model.

Pandas were chosen in data analysis and manipulation as it contains numerous libraries needed to achieve this objective. In particular, sklearn provides data structures and operations for manipulating the numerical figures crucial in calculations to determine the type of NRW.

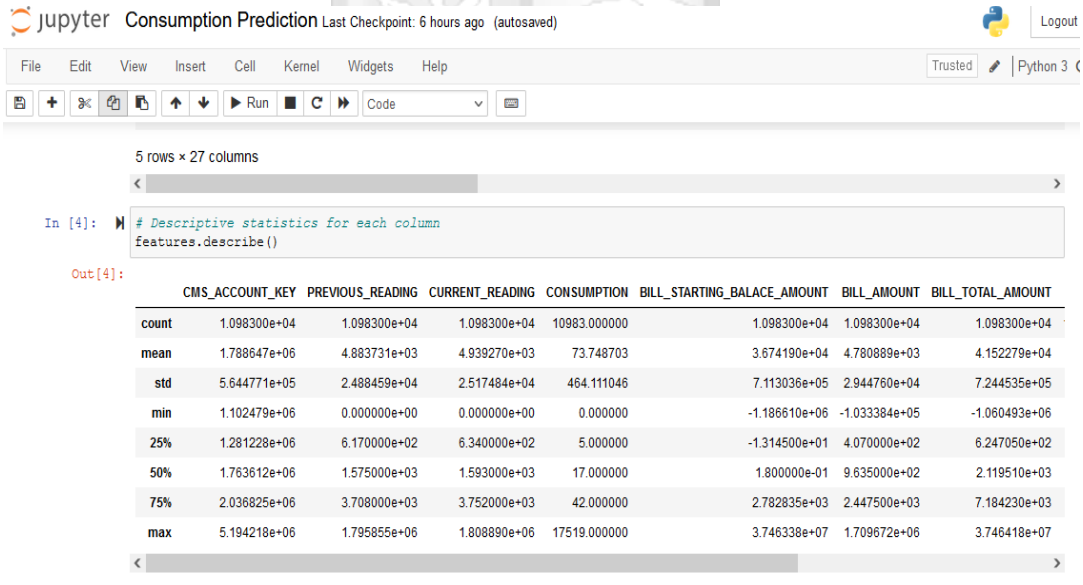
5.2.1 Model developments process

The first step taken towards developing the model was data cleaning using the python programming language and sklearn library. The datasets used were extracted from a data source shared by the Nairobi water. It consisted of over 8000 customers in a certain zone. The below figures illustrate the methods used to clean and fit the dataset in readiness to be used in the model.



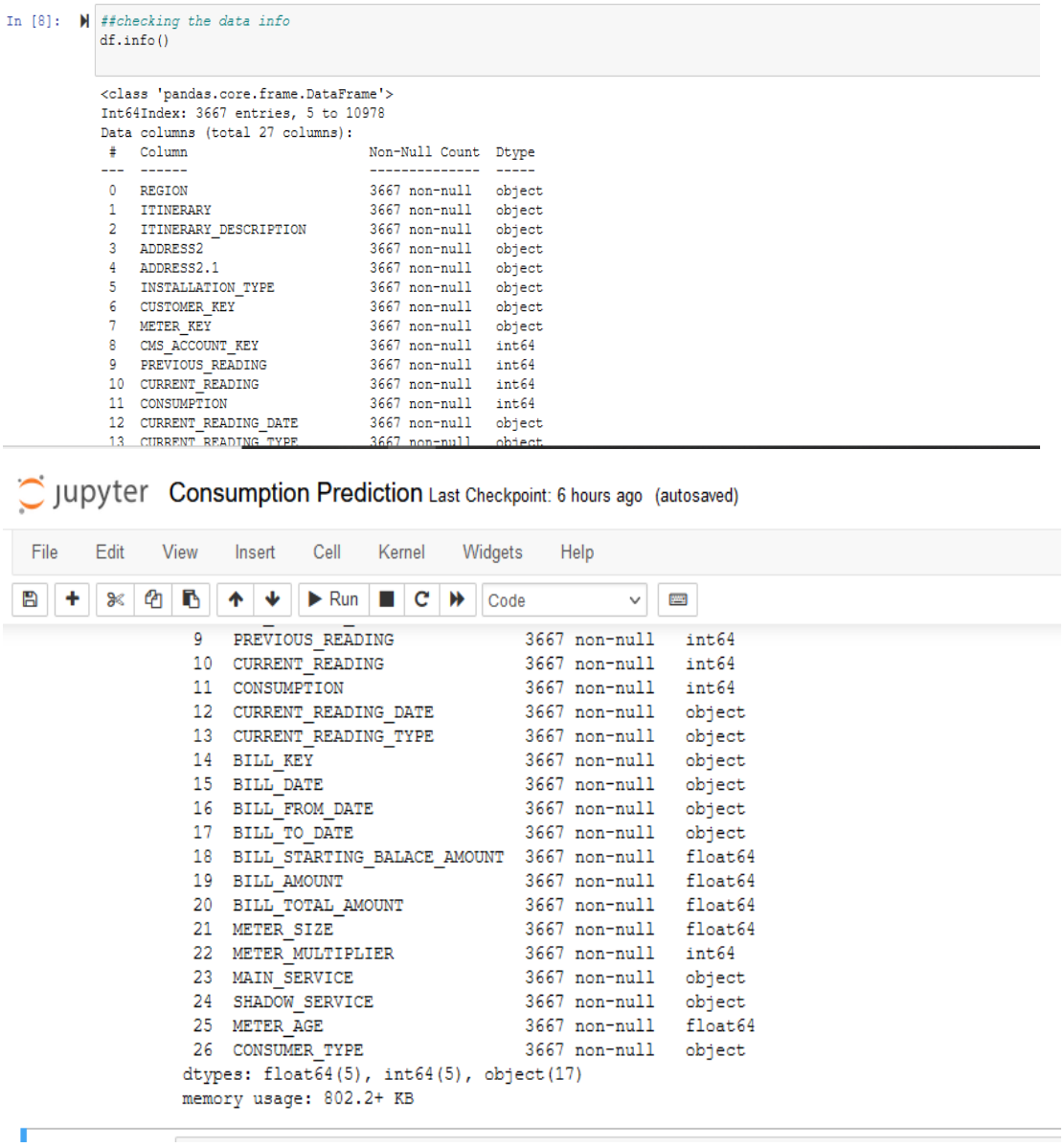
	REGION	ITINERARY	ITINERARY_DESCRIPTION	ADDRESS2	ADDRESS2.1	INSTALLATION_TYPE	CUSTOMER_KEY	METER_KEY	CMS_ACCOUN
0	NOR	NOR_054_088_001377	MT23-1D-A	KIAMBU RD-LIMURU RD-THIKA RD	KIAMBU RD-NJATHAINI	DOMESTIC	CU-1060634	NCWSC10334-05	12
1	NOR	NOR_054_090_019046	MT23-3C-A-B	KIAMBU RD-LIMURU RD-THIKA RD	THIKA RD-ROYSAMBU	DOMESTIC	CU-1190740	NCWSC115166	19
2	NOR	NOR_054_088_001387	MT23-1G-B	KIAMBU RD-LIMURU RD-THIKA RD	KIAMBU RD-NJATHAINI	DOMESTIC	CU-1143686	15 58304 14	22
3	NOR	NOR_054_088_001376	MT23-1C-C	KIAMBU RD-LIMURU RD-THIKA	KIAMBU RD-NJATHAINI	DOMESTIC	CU-1166537	NCWSC151169017	17

Figure 5.1. An illustration of dataset cleaning and fitting methods



	CMS_ACCOUNT_KEY	PREVIOUS_READING	CURRENT_READING	CONSUMPTION	BILL_STARTING_BALACE_AMOUNT	BILL_AMOUNT	BILL_TOTAL_AMOUNT
count	1.098300e+04	1.098300e+04	1.098300e+04	10983.000000	1.098300e+04	1.098300e+04	1.098300e+04
mean	1.788647e+06	4.883731e+03	4.939270e+03	73.748703	3.674190e+04	4.780889e+03	4.152279e+04
std	5.644771e+05	2.488459e+04	2.517484e+04	464.111046	7.113036e+05	2.944760e+04	7.244535e+05
min	1.102479e+06	0.000000e+00	0.000000e+00	0.000000	-1.186610e+06	-1.033384e+05	-1.060493e+06
25%	1.281228e+06	6.170000e+02	6.340000e+02	5.000000	-1.314500e+01	4.070000e+02	6.247050e+02
50%	1.763612e+06	1.575000e+03	1.593000e+03	17.000000	1.800000e-01	9.635000e+02	2.119510e+03
75%	2.036825e+06	3.708000e+03	3.752000e+03	42.000000	2.782835e+03	2.447500e+03	7.184230e+03
max	5.194218e+06	1.795855e+06	1.808890e+06	17519.000000	3.746338e+07	1.709672e+06	3.746418e+07

Figure 5.2. Data description



```
In [8]: ##checking the data info
        df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 3667 entries, 5 to 10978
Data columns (total 27 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   REGION                      3667 non-null   object
 1   ITINERARY                   3667 non-null   object
 2   ITINERARY_DESCRIPTION       3667 non-null   object
 3   ADDRESS2                    3667 non-null   object
 4   ADDRESS2.1                  3667 non-null   object
 5   INSTALLATION_TYPE           3667 non-null   object
 6   CUSTOMER_KEY                3667 non-null   object
 7   METER_KEY                   3667 non-null   object
 8   CMS_ACCOUNT_KEY             3667 non-null   int64
 9   PREVIOUS_READING            3667 non-null   int64
 10  CURRENT_READING             3667 non-null   int64
 11  CONSUMPTION                 3667 non-null   int64
 12  CURRENT_READING_DATE        3667 non-null   object
 13  CURRENT_READING_TYPE        3667 non-null   object
```


jupyter Consumption Prediction Last Checkpoint: 6 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Run Code

```
 9   PREVIOUS_READING              3667 non-null   int64
 10  CURRENT_READING               3667 non-null   int64
 11  CONSUMPTION                   3667 non-null   int64
 12  CURRENT_READING_DATE          3667 non-null   object
 13  CURRENT_READING_TYPE          3667 non-null   object
 14  BILL_KEY                      3667 non-null   object
 15  BILL_DATE                     3667 non-null   object
 16  BILL_FROM_DATE                3667 non-null   object
 17  BILL_TO_DATE                  3667 non-null   object
 18  BILL_STARTING_BALACE_AMOUNT   3667 non-null   float64
 19  BILL_AMOUNT                   3667 non-null   float64
 20  BILL_TOTAL_AMOUNT             3667 non-null   float64
 21  METER_SIZE                    3667 non-null   float64
 22  METER_MULTIPLIER              3667 non-null   int64
 23  MAIN_SERVICE                  3667 non-null   object
 24  SHADOW_SERVICE                3667 non-null   object
 25  METER_AGE                     3667 non-null   float64
 26  CONSUMER_TYPE                 3667 non-null   object
dtypes: float64(5), int64(5), object(17)
memory usage: 802.2+ KB
```

Figure 5.3. An illustration of how the model will check for data types

Check for nulls, study the missing values, and choose the best method to deal with them. We choose to drop them.

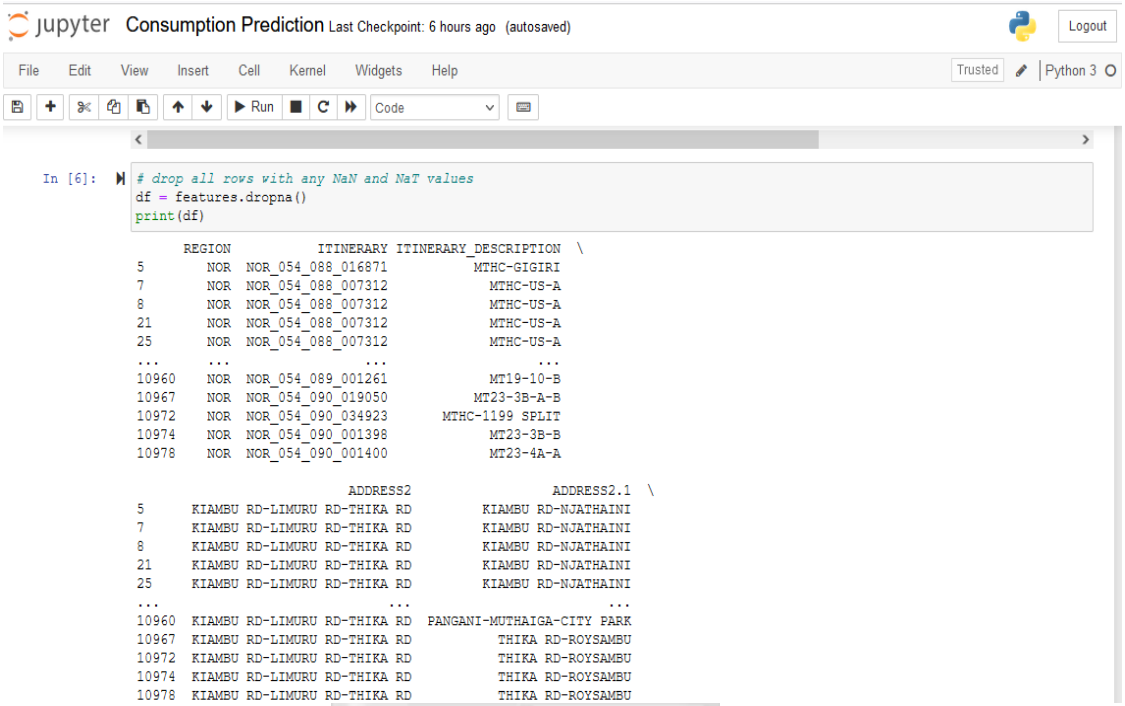


Figure 5.4. An illustration of how the model will handle missing values and nulls

5.2.2 Splitting data into Training and Testing data

We created both the training and test sets in a one-liner, bypassing to train_test_split () the modeling. By using testsize = 0.2 it means that the ration of training dataset to testing dataset is 4:1. DataFrame along with the fraction of the examples that should be included in the testing set. As before, we also set a random_state so that the results are reproducible; every time we run the code, the same instances will be included in the training and testing sets, respectively. The method returns a tuple with two DataFrames containing the training and testing examples.

```
In [12]: df1 = df[["CUSTOMER_KEY", "ADDRESS2", "CONSUMER_TYPE", "BILL_DATE", "CONSUMPTION"]]
         df1.head(5)
```

Out[12]:

	CUSTOMER_KEY	ADDRESS2	CONSUMER_TYPE	BILL_DATE	CONSUMPTION
5	CU-1218165	KIAMBU RD-LIMURU RD-THIKA RD	High Consumer	1-Feb-22	3474
7	CU-1141714	KIAMBU RD-LIMURU RD-THIKA RD	High Consumer	3-Feb-22	67
8	CU-1179212	KIAMBU RD-LIMURU RD-THIKA RD	High Consumer	3-Feb-22	122
21	CU-1141571	KIAMBU RD-LIMURU RD-THIKA RD	High Consumer	3-Feb-22	56
25	CU-1124193	KIAMBU RD-LIMURU RD-THIKA RD	High Consumer	3-Feb-22	1

```
In [ ]:
```

jupyter Consumption Prediction Last Checkpoint: 6 hours ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

```
21              SEW        8.6    High Consumer
25              SEW        5.9    High Consumer
...             ...        ...              ...
10960           SEW        3.7  Normal Consumer
10967           SEW        1.7  Normal Consumer
10972           SEW        1.6  Normal Consumer
10974           SEW        1.7  Normal Consumer
10978           SEW        5.0  Normal Consumer

[3667 rows x 27 columns]
```

```
In [7]: from sklearn.model_selection import train_test_split

        training_data, testing_data = train_test_split(df, test_size=0.2, random_state=25)

        print(f"No. of training examples: {training_data.shape[0]}")
        print(f"No. of testing examples: {testing_data.shape[0]}")

No. of training examples: 2933
No. of testing examples: 734
```

Figure 5.5. An Illustration of the process of data splitting

5.3 Feature Engineering

Feature engineering or feature extraction uses domain knowledge to extract features from raw data. The motivation is to use these extra features to improve the quality of results from a machine learning process, compared with supplying only the raw data to the machine learning process (Boehmke, Bradley; Greenwell, Brandon (2019). In this section, we aim to extract features from our raw dataset.

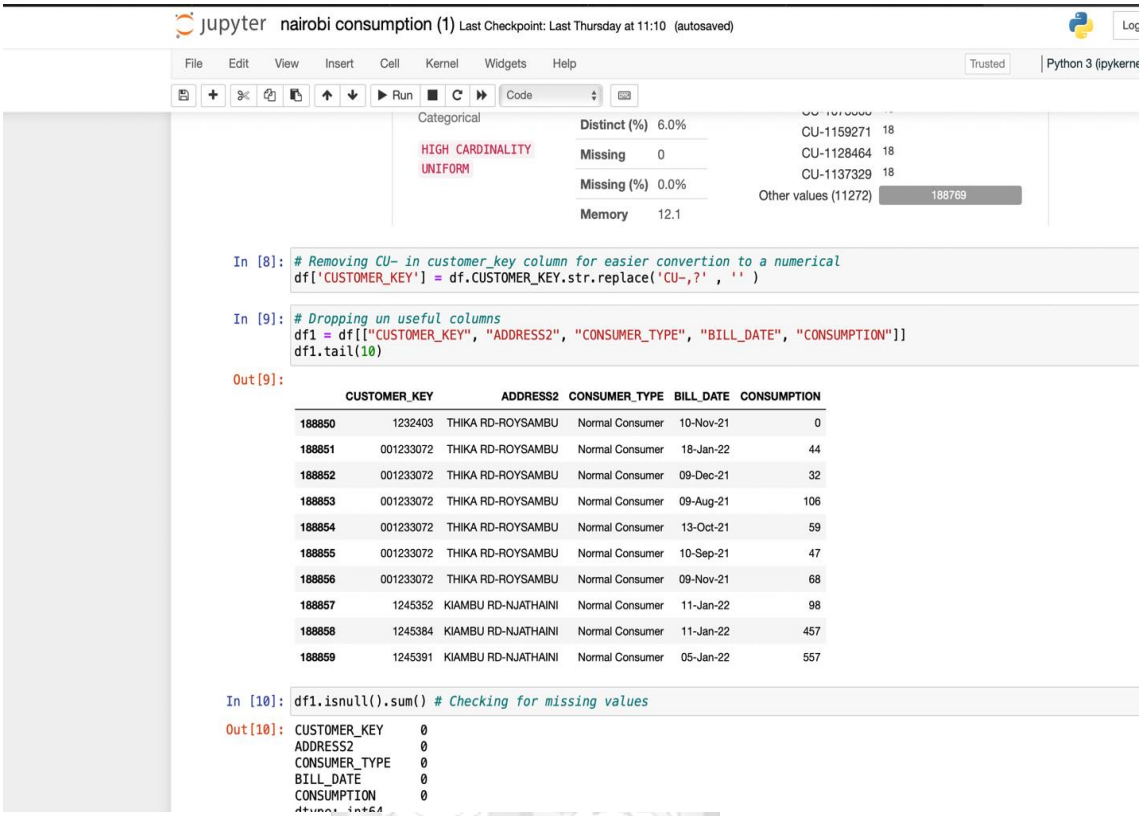


Figure 5.6. The process of extracting features from raw data

In the above code snippet, we have removed some elements in the customer key column for easier conversion to numerical, dropped columns that were not useful, and checked missing values.

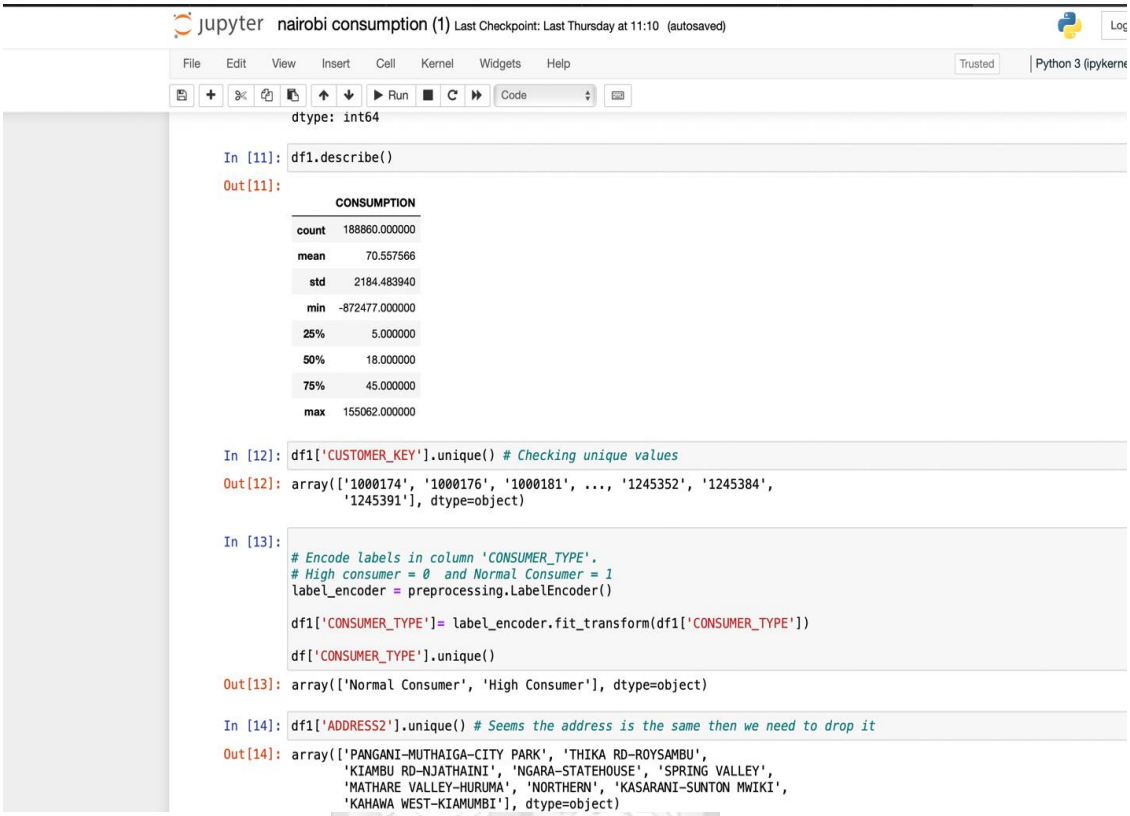


Figure 5.7. Extraction of feature from raw data by checking customer key uniqueness

From the above snippet, we need to check the columns customer key's uniqueness and drop the address since it won't be used in our model. Next, we need to convert the independent variables into integers that regressions models can understand.

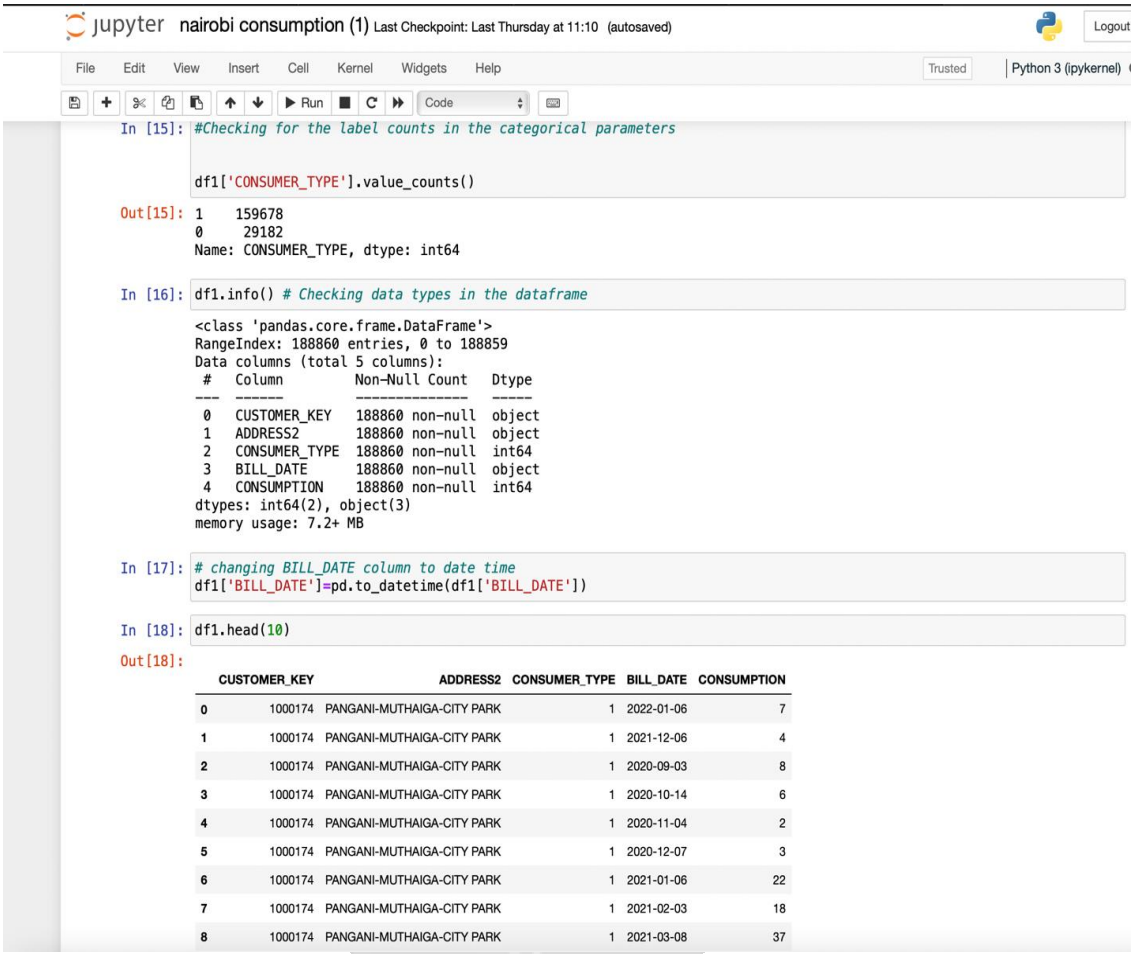


Figure 5.8. Conversion of independent variables into integers

After that, we separate the independent variables from the dependent variables. Then split the dataset into train and test data (80% train and 20% test)

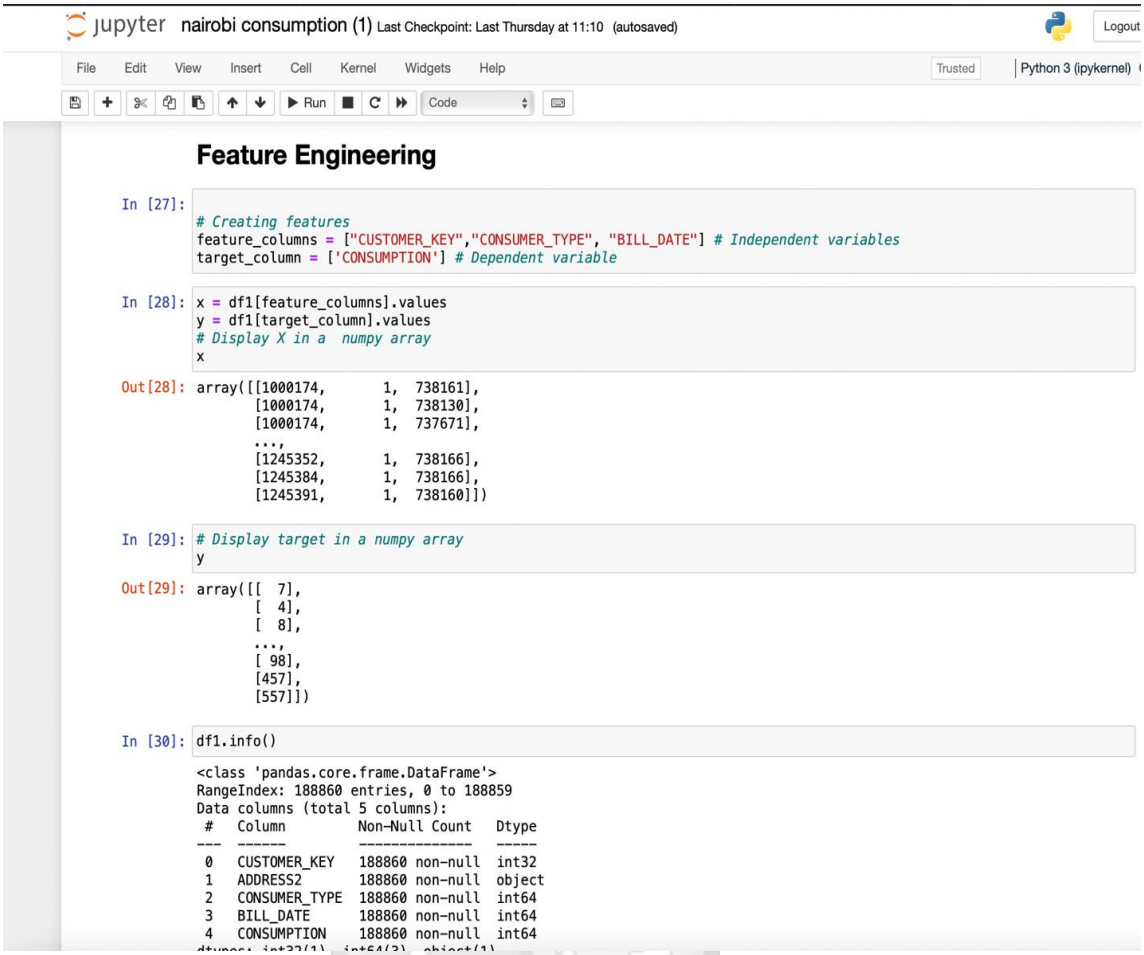


Figure 5.9. Separating independent variables from dependent variables

Get the correlation of each Feature in the dataset and plot the results.

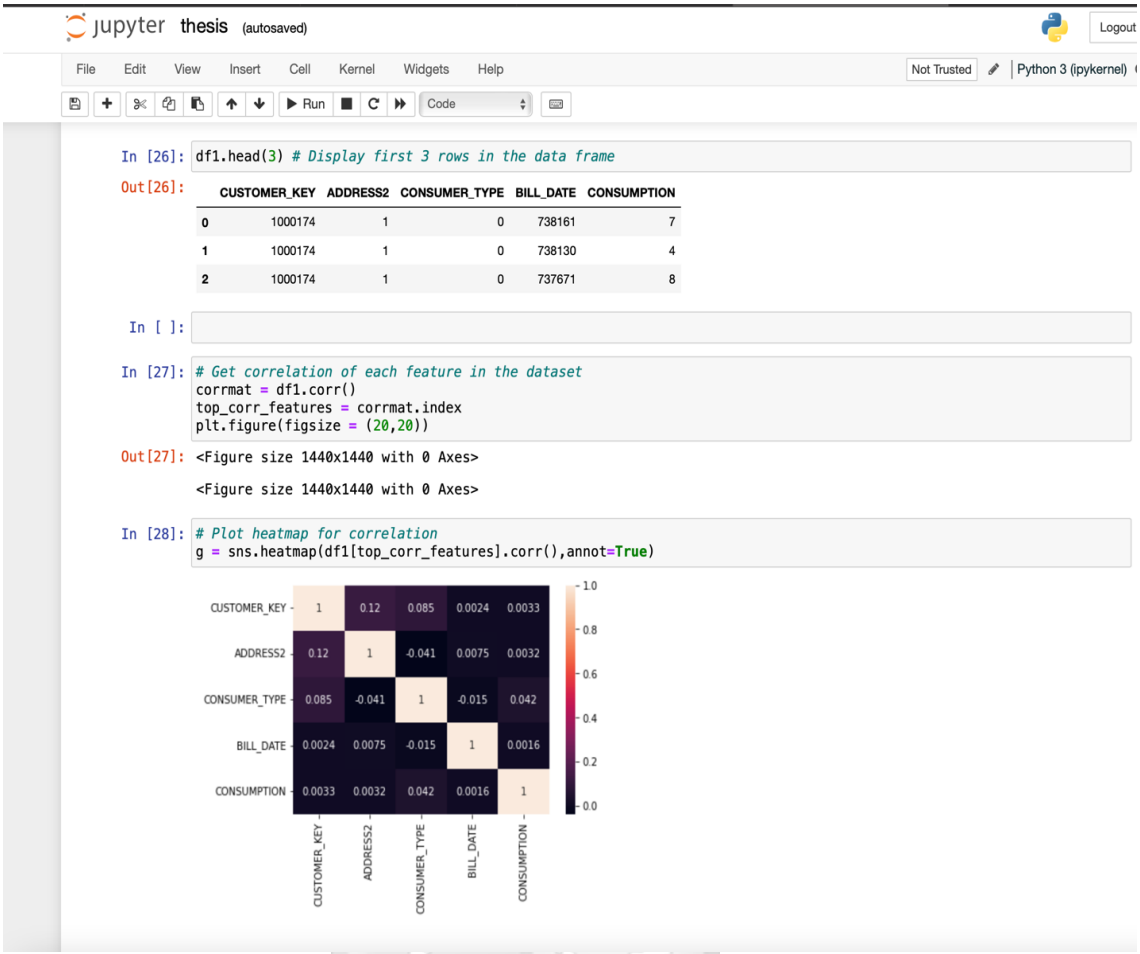


Figure 5.10. Establishing Correlation between features in dataset and plotting the results

5.4 Implementation and Setup of Linear Regression Algorithm

After successfully cleaning, preparing, and extracting features from our dataset, it is now time we implement our model. The aim is to train our model using linear regression.

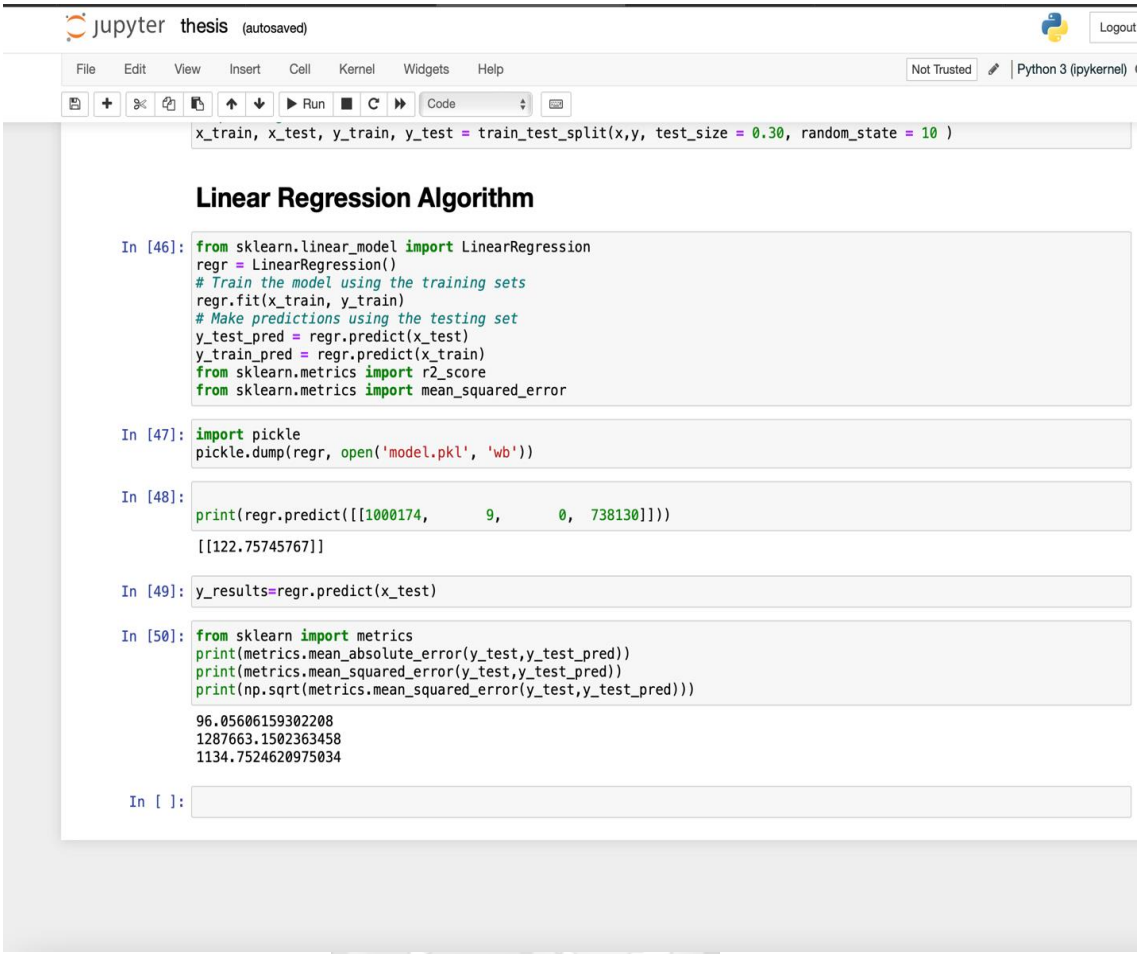


Figure 5.11. An illustration of the process of training the model using linear regression

We use the sklearn library, which features linear regression capabilities. Then, we use a pickle for packaging our model into a production-ready model.

5.5 Testing the Model

For the model to be considered operationally viable, it must be tested to achieve the desired results. The model results showed high accuracy in the customer consumption prediction; accuracy of 95% was achieved. We have prepared a prototype frontend page (angular) to call an API developed in flask (python language), making regular calls to our model for prediction. The result is a numerical prediction of the water consumption of the given customer.

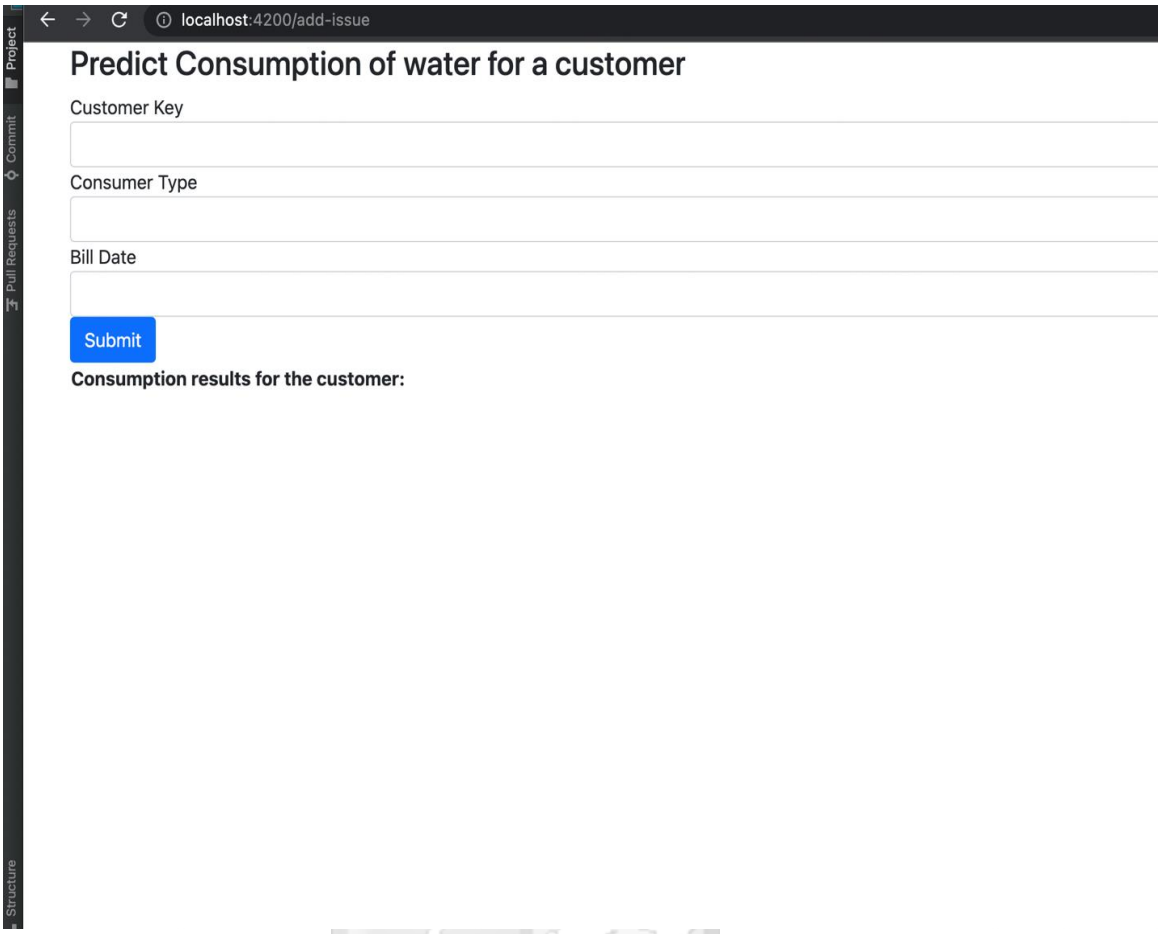


Figure 5.12. A screen caption of code for training the model

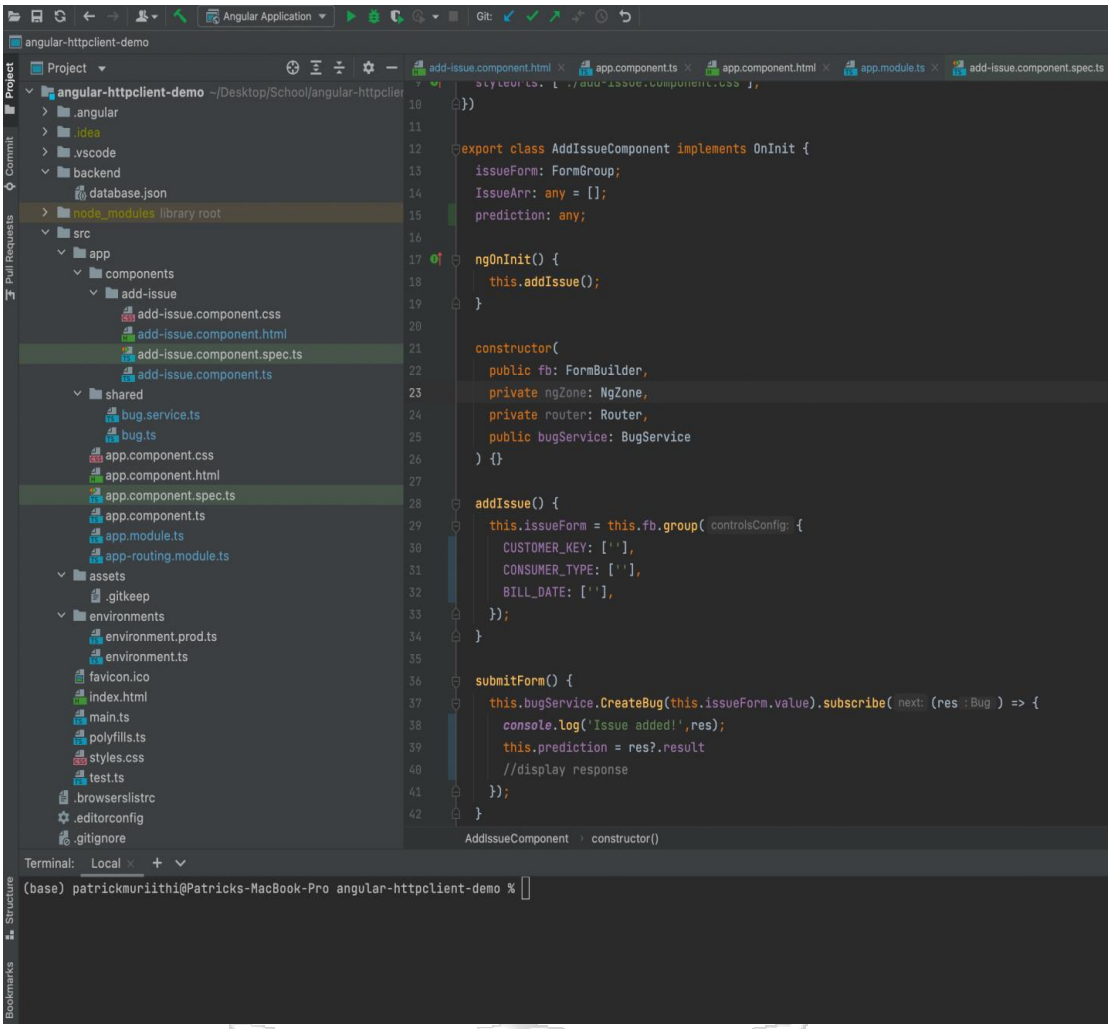


Figure 5.13. Frontend page to consume the model API

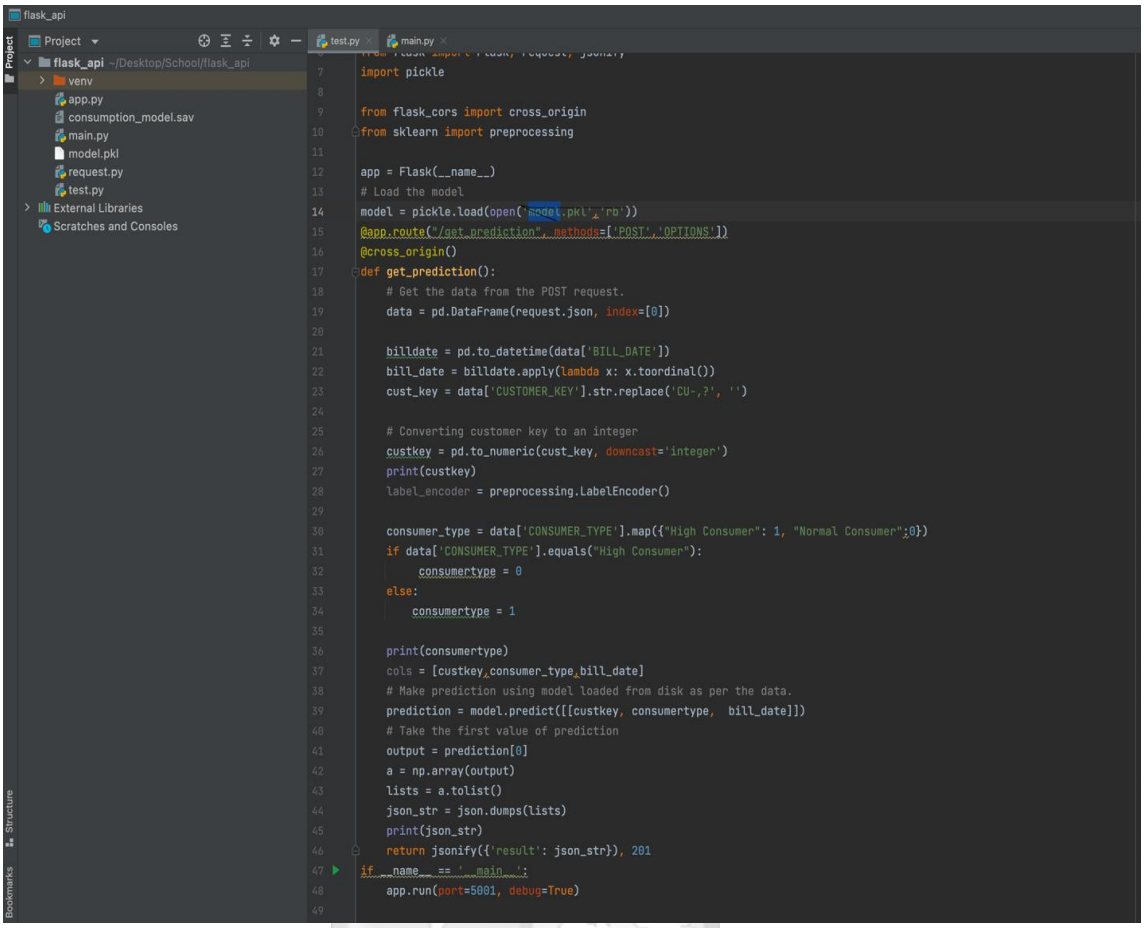


Figure 5.14. API Implementing the model

The model is also tested to ensure that it provides accurate data based on the input data and prioritizes results based on the operator's needs. During these tests, previous consumption data was used as the test data. The test data is taken in its raw form to mimic a live working environment where the data received is unbalanced and may have inconsistencies. Therefore, the model must be able to make the most out of the available input data by identifying patterns and inconsistencies that signal NRW consumption.

The tests carried on the model ensure that bugs are identified either in the units or within the model as a whole that would otherwise hamper the intended optimal performance of the model. The model was also tested to ensure that it meets all the functional and non-functional requirements stipulated by the user. Once it is established that the model meets the requirements, its usability is tested. The key focus in usability testing was to ensure that the model could perform all the functions as needed by the user in an easy, smooth, and user-friendly manner. Usability testing seeks to establish

whether the model is user-friendly and that users can easily use and learn to use the model. For the model to be considered successful, it must achieve a pass in each test, where a pass means that it performed as expected. The below table provides a summary of the tests carried out and what a pass in each test means.

Chapter 6: Discussions

6.1 Introduction

The research findings are discussed in connection with the objectives in this chapter. It explains how this was accomplished, starting with the literature study and continuing through the design and execution stages. It also goes over the parameters utilized to gauge the implementation's success. There's also a summary of the findings and some recommendations for further research.

6.2 Collecting, cleaning, manipulating and processing data from WSP

One of the key objectives of this research was to establish an effective methodology that water service providers could utilize to collect and use the data collected to form part of the decision-making process. As noted in the literature review, collecting data by WSPs is often a daunting challenge due to the vastness of the data to be collected and other challenges related to incompleteness and inconsistencies in the data collected. It was established that more focus should be on how to work on the data collected in its unbalanced form to make predictions and conclusions rather than trying to improve the data itself, as it would be a difficult and expensive endeavor. Nonetheless, it was also noted that transiting to digital meters could help the process of data collection by eliminating data losses and data biases arising from human inefficiencies. Besides, it would significantly improve the process of cleaning, manipulating, and processing data.

6.3 Conduct an analysis and a survey on methods for unbalanced class machine learning

As mentioned earlier, the data collected by WSPs is extremely unbalanced. This realization prompted the need to establish the most efficient methodology that can be used to work with such kind of data to develop a machine learning model. A deep dive into the different existing methods was conducted with emphasis on analyzing the advantages and shortcomings of each. After a thorough analysis, it was established that each method has its advantages over other methods. It was also established that one way the shortcomings could be overcome is by formulating a model that implements a combination of the methodologies. However, such an endeavor is limited in its applicability. It results in overly complex systems that are untenable in terms of cost of

operation and the skillset required operating them. Furthermore, it was noted that implementing more than one methodology results in other shortcomings that eventually affect the system's overall effectiveness.

In establishing the current model, much focus was given to the machine learning models, as they were more computationally efficient when dealing with vast data in real-world operations. As established in the literature review, the 4 main methodologies that provide better results when dealing with unbalanced class data in real-world operations are under-sampling, oversampling, SMOTE, and random forest. The previous studies noted that a combination of these methodologies provides an optimal solution with fewer shortcomings and does not increase the cost of operation and the complexity of the resultant system.

6.4 To design, develop and implement a Machine Learning model

This was the primary objective of this research paper. The need to develop such a model is in the backdrop of the failure of the existing systems to solve the problem of NRW that has plagued WSPs for years. Based on the literature review, it was noted that most of the existing systems rely on the physical presence of field officers to identify sources of the NRW, such as leakages and illegal connections. Further, it was established that the existing systems do not have mechanisms to remotely localize NRW, thus needing extra field officers, which has become a financial burden for WSPs. For those WSPs that dared to implement better systems, the cost of running such systems, the skill requirement, and the overall bulkiness and complexity made such systems inoperable in the long term, thus leaving the NRW problem unsolved for long.

According to numerous research studies conducted, it was established that any system attempting to solve this problem must be simple and easy to use, foremost to eliminate the human skills that most of the previous systems posed. Further, it was noted that any such system must be able to manipulate and process vast amounts of unbalanced data without adding to the system's complexity. Besides, the resultant system should be able to localize NRW sources and provide a prediction of the type of NRW and offer real-time monitoring, notifications, and reports. It was established that such a system must implement well-defined and developed machine learning algorithms to perform deep searches, analyze patterns and report accordingly

6.5 Testing the developed ML model

Testing is a prerequisite for the acceptance of systems. Testing is done to ascertain that the system can effectively achieve what it said it could do and how it is supposed to do it. A literature review notes that several tests can be conducted on the system depending on its intended use. However, there are some common tests that all systems must go through; among them is unit testing, where the functioning of each of the system's units is tested; user testing, where the ease of use is evaluated and functional testing, where the system is evaluated against the functional and non-functional requirements established during the customer requirements phase of the project.

6.6 Model Validity and advantages to the current systems

The model created proves to be a feasible solution to the problem of NRW. Despite its geographical location, it requires minimal resources to fully implement and can be easily adapted to any WSP. The model is designed to work effectively in any condition, provided the minimum operational requirements are met. It also provides numerous advantages over the currents systems, among the reduction of human data manipulation and automation of the process of NRW identification besides allowing for seamless and real-time monitoring of water distribution infrastructure.

6.7 Research Results and Contributions

The model showed high accuracy of 95%. The successful implementation of the model provides a breakthrough that can be replicated to monitor other critical supply infrastructures such as oil and electricity with slight modifications to adapt it to the workings of these sectors. Therefore, such a fete is significant as it would help reduce thefts and revenue losses in those areas, which are especially rampant in developing countries, particularly in Kenya, where losses from water and electricity theft run into billions of shillings. The model further is an accomplishment in computer algorithms as the world continuously shifts to developing machine learning algorithms that can help solve the outstanding challenges that threaten the wellbeing of people, resources, and the environment.

Chapter 7: Conclusion and Recommendations

7.1 Conclusion

This research study was conducted to develop a machine learning model that can detect NRW in a water distribution infrastructure. The need for such a model arose from the failure of the previous system implemented to address the challenge effectively and holistically. As noted in other sections of this research study, previously implemented systems were marred with inefficiencies and shortcomings that deemed them unfeasible to operate in the long term without threatening the long term existence of WSPs. However, it was noted that significant progress had been made in addressing the problem of NRW, but the main areas of concern for WSPs remain unaddressed by the current systems.

In response to the prevailing situation, this research study sought to provide a structured approach that can be implemented to solve the problem summarily. The approach suggested was the formulation, development, and implementation of a machine learning model capable of learning over time to make critical decisions related to identifying different types and sources of NRW and providing recommendations on how much can be addressed in real-time. The model developed can work with the unbalanced data collected from different consumers, clean it, analyze and manipulate it to identify patterns that can then be used to identify different NRWs. Besides, the model needed to be able to localize the NRW to reduce the need for field officers, thus steeply reducing the cost of operating the system. In addition, the real-time monitoring of the consumption data provides for rapid reaction to cases of NRW, which other systems don't provide. Effectively, the model provides an effective way to curb NRW, protects WSPs from revenue losses, and ensures that consumers are protected from shortages occasioned by a lack of resources due to lost revenue.

7.2 Recommendations

Based on the developed model, the researcher notes some recommendations can be made towards helping solve the problem of NRW effectively and help in the progression of research in this field.

i. WSPs should strive towards replacing manual meters with digital meters to ensure ease of data collection, reduce consumer interference, and cases of data manipulation by officials.

ii. WSPs should establish research departments within their organization. This could go a long way in developing feasible plans of action to deal with numerous challenges facing them currently and in the future.

iii. In implementing the current model, WSPs must be aware the effectiveness of such solutions is entirely based on their organizational operations, such as having an effective data collection mechanism and having swift response teams to address issues raised by the system. Most importantly, it must be noted that the system only identifies what could be a probable NRW but implementing decisions remains the role of the management. It is recommended that WSPs formulate proper strategies and guidelines that complement the systems to achieve success.

7.3 Areas of Future Research

Challenges keep evolving, which calls for researchers to keep up with these challenges by carrying out continuous research. In line with this view, research in this field must continue to seek further solutions to the problem of NRW. As is, people have often found ways to bypass solutions implemented, and therefore, we cannot be oblivious of this point when implementing this system. Therefore, this is a rallying call for further research in water distribution and other areas of amenity provision, such as electricity. Due to the similarity in distribution architecture of major utilities such as water, oil, and electricity, it is crucial to note that research breakthroughs in one field can easily be replicated in the other fields.

References

Alexander, J. (2016). Reducing non-revenue water: water. *IMIESA, 41*(9), 37-40.

Almeida, F., Brennan, M., Joseph, P., Whitfield, S., Dray, S., & Paschoalini, A. (2014). On the acoustic filtering of the pipe and sensor in a buried plastic water pipe and its effect on leak detection: an experimental investigation. *Sensors, 14*(3), 5595-5610.

Amoatey, P. K., Minke, R., & Steinmetz, H. (2018). Leakage estimation in developing country water networks based on water balance, minimum night flow, and component analysis methods. *Water Practice & Technology, 13*(1), 96-105.

Atef, A., Zayed, T., Hawari, A., Khader, M., & Moselhi, O. (2016). Multi-tier method using infrared photography and GPR to detect and locate water leaks. *Automation in Construction, 61*, 162-170.

Barandouzi, M. A., Mahinthakumar, G., Ranjithan, R., & Brill, E. D. (2012). Probabilistic mapping of water leakage characterizations using a Bayesian approach. In *World Environmental and Water Resources Congress 2012: Crossing Boundaries* (pp. 3248-3256).

Bhat, R. R., Trevizan, R. D., Sengupta, R., Li, X., & Bretas, A. (2016, December). Identifying non-technical power loss via spatial and temporal deep learning. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 272-279). IEEE.

Brian Pickard et al (2003). Reducing Non-Revenue Water: A Myriad of Challenges. Tampa Water Department, Florida, USA.

Charalambous, B., & Hamilton, S. (2011). Water balance-the next stage. *Water utility journal, 1*, 3-10.

Criminisi, A., Fontanazza, C. M., Freni, G., & Loggia, G. L. (2009). Evaluation of the apparent losses caused by water meter under-registration in the intermittent water supply. *Water Science and Technology, 60*(9), 2373-2382.

Depuru, S. S. S. R., Wang, L., & Devabhaktuni, V. (2011, March). Support vector machine-based data classification for detection of electricity theft. In *2011 IEEE/PES Power Systems Conference and Exposition* (pp. 1-8). IEEE.

Farley, M., Wyeth, G., Ghazali, Z. B. M., Istandar, A., Singh, S., Dijk, N., ... & Kirkwood, E. (2008). The manager's non-revenue water handbook: a guide to understanding water losses. *United States of America: United States Agency for International Development (USAID)*.

GAO, B. K., & REN, X. J. (2014). Pipeline Leak Detection Based on Improved Differential Evolution Algorithm and Fuzzy Neural Network. *Control and Instruments in Chemical Industry*, 01.

Gupta, G. (2017). Monitoring Water Distribution Network using Machine Learning.

Hu, X., Han, Y., Yu, B., Geng, Z., & Fan, J. (2021). Novel leakage detection and water loss management of urban water supply network using multiscale neural networks. *Journal of Cleaner Production*, *278*, 123611.

Imaizumi, H. (1987). *U.S. Patent No. 4,677,371*. Washington, DC: U.S. Patent and Trademark Office.

Jensen, H. A., & Jerez, D. J. (2019). A Bayesian model updating approach for detection-related problems in water distribution networks. *Reliability Engineering & System Safety*, *185*, 100-112.

Kanakoudis, V., & Muhammetoglu, H. (2014). Urban water pipe networks management towards non-revenue water reduction: Two case studies from Greece and Turkey. *CLEAN–Soil, Air, Water*, *42*(7), 880-892.

Li, R., Huang, H., Xin, K., & Tao, T. (2015). A review of methods for burst/leakage detection and location in water distribution systems. *Water Science and Technology: Water Supply*, *15*(3), 429-441.

Liemberger, R., Brothers, K., Lambert, A., McKenzie, R., Rizzo, A., & Waldron, T. (2007). Water loss performance indicators. In *Proceedings of IWA Specialised Conference Water Loss 23th-26th September* (pp. 148-160).

Mastaller, M., & Klingel, P. (2017). Adapting the IWA water balance to intermittent water supply and flat-rate tariffs without customer metering. *Journal of Water, Sanitation and Hygiene for Development*, *7*(3), 396-406.

McLaughlin, S., Holbert, B., Fawaz, A., Berthier, R., & Zonouz, S. (2013). A multi-sensor energy theft detection framework for advanced metering infrastructures. *IEEE Journal on Selected Areas in Communications*, *31*(7), 1319-1330.

Meniconi, S., Brunone, B., & Ferrante, M. (2011). In-line pipe device checking by short-period analysis of transient tests. *Journal of Hydraulic Engineering*, *137*(7), 713-722.

Monedero Goicoechea, I. L., Biscarri Triviño, F., Guerrero Alonso, J. I., Roldán, M., & León de Mora, C. (2015). An Approach to Detection of Tampering in Water Meters. In *KES 2015: 19th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (2015), p 413-421*. Elsevier.

Mons, J. (2010). Strategies for non-revenue water management in developing countries: a case study of Kampala, Uganda. Kampala, Uganda

Mutikanga, H. E., Sharma, S. K., & Vairavamoorthy, K. (2011). Assessment of apparent losses in urban water systems. *Water and Environment Journal*, *25*(3), 327-335.

Ng'etich, C. K. (2015). *Corporate governance and financial performance of water companies in Kenya* (Doctoral dissertation, University of Nairobi).

Ogutu, G. A., Okuthe, P. K., & Lall, M. (2017). A review of probabilistic modeling of pipeline leakage using Bayesian Networks. *J. Eng. Appl. Sci*, *12*, 3163-3173.

Pérez-Pérez, E. J., López-Estrada, F. R., Valencia-Palomo, G., Torres, L., Puig, V., & Mina-Antonio, J. D. (2021). Leak diagnosis in pipelines using a combined artificial neural network approach. *Control Engineering Practice*, *107*, 104677.

Radivojević, D., Blagojević, B., & Ilić, A. (2020). Water Supply System Performance Improvement in the Town of Pirot Using Water Balance IWA Methodology and Numerical Simulations. *Tehnički vjesnik*, *27*(3), 970-977.

Soldevila, A., Fernandez-Canti, R. M., Blesa, J., Tornil-Sin, S., & Puig, V. (2017). Leak localization in water distribution networks using Bayesian classifiers. *Journal of Process Control*, *55*, 1-9.

Steffelbauer, D., Neumayer, M., Günther, M., & Fuchs-Hanusch, D. (2014). Sensor placement and leakage localization considering demand uncertainties. *Procedia Engineering*, *89*, 1160-1167.

U.N. HABITAT. (2012). Illegal use reduction operation manual. *United Nations Human Settlements Programme Kenya*.

WASREB. (2010). Impact 3: A performance report of Kenya's water services sub-sector.

World Bank. (2006). Using Performance-Based Contracts to Reduce Non-Revenue Water.

Wu, Z. Y., & Sage, P. (2008). Water loss detection via genetic algorithm optimization-based model calibration. In *Water Distribution Systems Analysis Symposium 2006* (pp. 1-11).

Wyatt, A., Richkus, J., & Sy, J. (2016). Using Performance-Based Contracts to Reduce Non-Revenue Water. *(PPIAF Report) International Bank for Reconstruction and Development, The World Bank*.

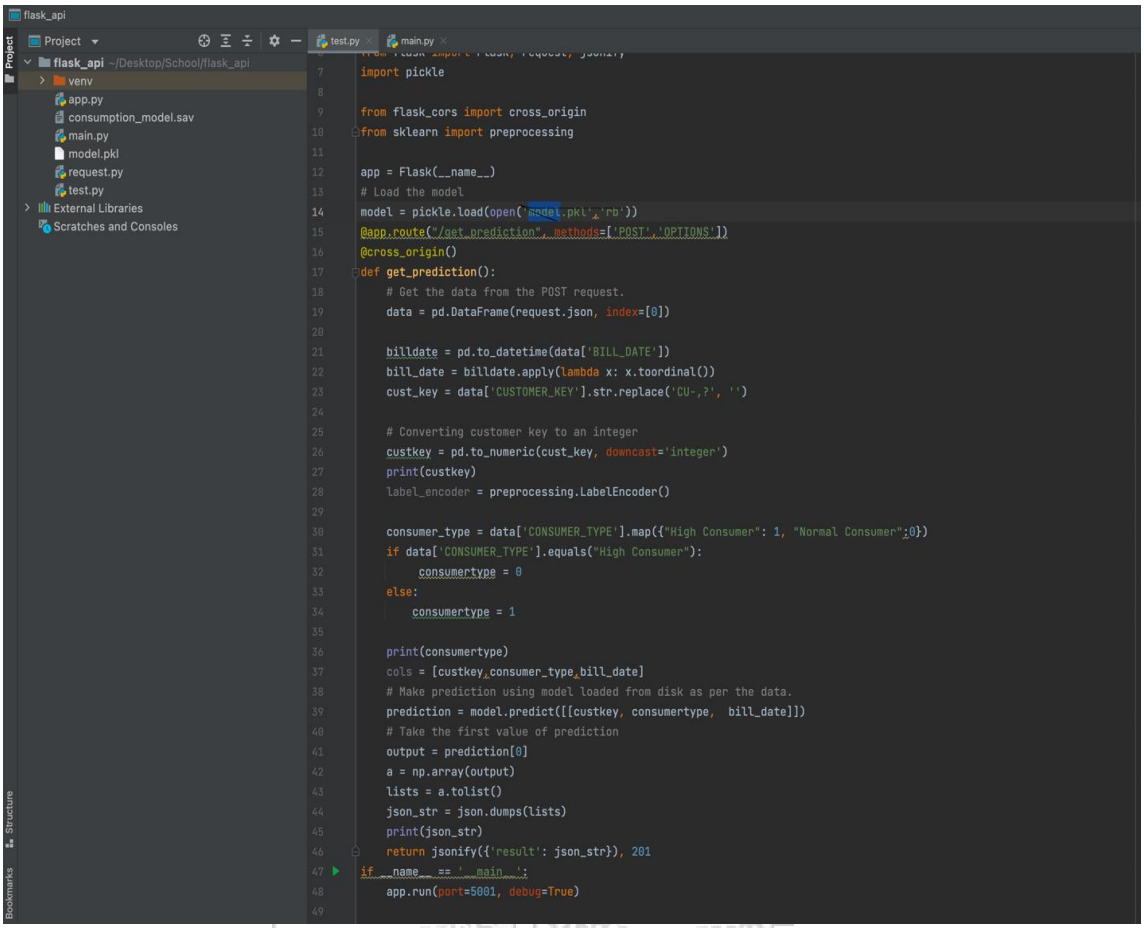
Xin, K., Tao, T., Lu, Y., Xiong, X., & Li, F. (2014). Apparent losses analysis in district metered areas of water distribution systems. *Water resources management*, *28*(3), 683-696.

Yeboah, P. A. (2008). Management of non-revenue water: a case study of the water supply in Accra, Ghana".

Zhou, B., Lau, V., & Wang, X. (2019). Machine-learning-based leakage-event identification for smart water supply systems. *IEEE Internet of Things Journal*, *7*(3), 2277-2292.

Appendices

Appendix A: Sketch Program



Appendix B: Industry Approval

NAIROBI CITY WATER & SEWERAGE COMPANY LTD.
KAMPALA RD, P. O. Box 30656-00100, Nairobi, Kenya
Tel: +254 703 080 000
Email: info@nairobiwater.co.ke
www.nairobiwater.co.ke

NCWSC/HR/TRG.14/VOL.8/001/MMM/ak 28th February, 2021

Patrick Kimani,
Strathmore University,
P.O Box 59857-00200,
Nairobi.
Email: Patrickkimani273@gmail.com

Cell: 0729105862.

Dear Patrick,

RE: RESEARCH ON A MODEL TO DETECT NON- REVENUE WATER WITH UNBALANCED CLASSES USING MACHINE LEARNING

Reference is made to your letter dated 17th February, 2022 on the above-mentioned subject.

Approval is hereby granted to you to collect data from **1st March, 2022 and 30th August, 2022** for your Bachelor degree project titled ***"A model to detect Non- Revenue water with severely unbalanced classes using machine learning"*** *at Nairobi City Water and Sewerage Company limited* in Nairobi City County.

The Non – Revenue Water Manager whose office is located at National Water Offices will assist you with the relevant Data/information in relation to your project.

All findings/information on Company matters should be accorded utmost confidentiality.

Please note upon completion, you will be expected to submit a copy of the findings to the office of the undersigned.

By a copy of this letter the following Officers are hereby informed accordingly.

I. Non -Revenue Manager
II. Research and Development Manager

N.B: Kindly ensure you observe the Government directives on Covid -19 pandemic.

Yours Sincerely,

Eng. Nahason Muguna
Managing Director

Appendix C: Ethical Approval

Strathmore UNIVERSITY

16th February 2022

Mr Muriithi Patrick
patrick.muriithi@strathmore.edu

Dear Mr Muriithi,

RE: A Model to Detect Non-Revenue Water with Severely Unbalanced Classes Using Machine Learning

This is to inform you that SU-IERC has reviewed and **approved** your above **SU- master's** research proposal. Your application reference number is **SU-IERC1226/21.** The approval period is **16th February 2021 to 15th February 2022.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-IERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-IERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-IERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-IERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and also obtain other clearances needed

Yours sincerely,

for: Prof Fred Were,
Chairperson; SU-IERC

STRATHMORE UNIVERSITY INSTITUTIONAL ETHICS REVIEW COMMITTEE (SU-IERC)
16 Feb 2022
TEL: +254 (0)70303400
P.O BOX 59857-00200
NAIROBI-KENYA

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu