MERU UNIVERSITY OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

BACHELOR OF SCIENCE IN DATA SCIENCE

WATERBORNE DISEASES PREDICTION

REGISTRATION NUMBER: CT204/106232/21

NAME OF STUDENT: KIPKOROS VINCENT

A Research Project Submitted in Partial Fulfillment of the Requirements of the Bachelor of Science in Data Science of Meru University of Science and Technology

April, 2025

DECLARATION

This research proposal is my original work prepared with no other than the indicated sources and support, and has not been presented elsewhere for a different or similar assignment.

Student Reg. No. CT204/106232/21

Student Name KIPKOROS VINCENT 2504/2025

DEDICATION

I dedicate this report to my family and friends who have continuously supported and believed in my abilities, contributing to my success.

A special acknowledgment goes to my supervisor for his guidance and insights. Additionally, I am grateful to my mentor Dickson Gikombe for the wisdom and encouragement, which have continuously inspired me to strive for excellence.

Finally, I want to express my gratitude to Meru University of Science and Technology (MUST) for providing me with the foundation and knowledge necessary to succeed.

Thank you all for your unwavering support and belief in me

.

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my family and several individuals Evans, victor, and Amos for their unwavering support and encouragement throughout my studies. Additionally, the entire Meru University fraternity who played a significant role in the successful of my studies.

First and foremost, I extend my sincere appreciation to my course advisor, Mr. Antony Irungu, whose guidance, support, and encouragement was invaluable throughout my studies and even during my proposal writing. His insights and expertise provided me with a deeper understanding of the field and helped me navigate various challenges.

TABLE OF CONTENTS

Contents

LIST OF FIGURES

1 CHAPTER ONE
INTRODUCTION

1.1 Background of study

Water is an essential component of the ecosystem with many health advantages, but when it is highly contaminated, it can present serious health risks, such as transmissible infections through the water. Water quantity and quality are consistently associated with the commonly reported infectious diseases worldwide, thus suggesting that although water is essential to human survival and ecosystem function, it can also act as a source for the spread of diseases. Majority of waterborne diseases cases are seen in places with poor sanitation facilities and restricted areas to clean drinking water (Magana et al. 2020).

The health of the young and old in our society is seriously threatened by the waterborne infections, which are caused by inadequate facilities, a lack of public awareness, and inaccurate predictive models that rely on minimal historical available data. Although drinking polluted water is the main way of that these illnesses are, spread inadequate food handling, a dirty environment, and inadequate hygiene can also spread them. (Landrigan et al. 2020).

Common waterborne illnesses include cholera which is caused by Vibrio cholerae, dysentery induced by Entamoeba histolytica, typhoid caused by Salmonella typhi, and hepatitis A which is caused by virus. These diseases mainly affect slums, refugee's camps, rural communities' dependent of flowing rivers for their daily sustenance, flooded areas and urban areas, particularly those near improperly disposed sewages'. If left untreated, waterborne infections can cause severe dehydration and even death, particularly in vulnerable populations such as children and the elderly (Ali et al., 2020).

Historically, disease management strategies for waterborne diseases have relied on practices such as boiling water, water filtration, quarantines and isolation, chlorination, and public health campaigns. Even with this strategies, they frequently lack efficacy and cost effectiveness, and some of the procedures took a lot of time. As a result, there has been an

increasing demand for early disease detection and timely intervention, which are essential in minimizing the harmful effects of these diseases (Branz et al. 2017; Lantagne and Yates 2018). According to World Health Organization, waterborne disease cause approximately 1.5 million fatalities each year, which accounts for about 3.6% out of the entire global burden of diseases, whereas at least 1.7 billion of people across the globe consume drinking water from unsafe sources. Climate change, population increase and urbanization already pose challenges for water supply system. Better access to water source can lead to better health, particularly for young children who are more vulnerable to diseases associated to water consumption. Waterborne infections pose a serious threat to world health (WHO, 2022).

Technological advancements, particularly in intelligent systems, offer new solutions for addressing this issue. Predictive modeling techniques, which utilize historical and real-time data, enable the forecasting of the most common waterborne diseases outbreaks. These models allow medical specialist and researchers to detect potential problems early, thereby enabling more proactive and efficient disease management decisions (WHO.,2021).

In addition, for appropriate interventions to be developed, it is imperative to identify the underlying causes and the effects of their interactions on the risks of waterborne diseases. However, the complex water health nexus continues to pose issues for human and environmental health because of fragmented interventions and lack of multi-sectoral coordination, which has prevented the development and implementation of sustainable changes in water infrastructure, such as clean water supplies and functional sewage systems. (Ntajal, 2020).

This project aims to develop a machine learning-based predictive model that can forecast waterborne diseases outbreaks in Kenya by analyzing factors such as water quality, population density, sanitation conditions, and historical disease data. Thus enabling timely interventions, reduce disease transmission, and improve public health.

1.2 Motivation for study

The motive under this research is the high rate of deaths caused by waterborne illnesses which is approximately 1.5 million deaths annually, which accounts for about 3.6% of the total global burden of diseases. Despite interventions put in place to address this problems, it has not been effective. This results to use of new interventions like development of machine learning driven predictive models to assess and address this issues ultimately enhancing public health.

1.3 The statement of the Problem

Waterborne illnesses continue to be a major cause of morbidity and mortality, particularly in developing regions with poor sanitation and limited access to clean water. Vulnerable populations, including young children and the elderly, are most at risk. Despite interventions like sanitation facilities, handwashing campaigns, and chlorination of drinking water, these illnesses persist, significantly hindering public health and socioeconomic development. Furthermore, the high cost of clean drinking water in these regions forces many to rely on unsafe water sources, exacerbating the problem.

This project seeks to develop a machine learning-based predictive model to forecast the likelihood of waterborne diseases using clinical indicators such as symptoms, vital signs, and infection duration. This system will enable healthcare professionals to make early, data-driven decisions, improving patient outcomes and public health responses. To effectively address the root causes of waterborne diseases, interventions should focus on improving access to clean water, better waste management, and sanitation. Additionally, governments and NGOs must prioritize environmental cleanliness and raise awareness about the risks of consuming contaminated water.

1.4 Research objectives

1.4.1 General objectives

To develop a machine learning-based predictive model capable of forecasting the likelihood of waterborne disease infections in individuals.

1.4.2 Specific objectives

The objectives of this research project are:

i. To assess the relevance and effectiveness of clinical indicators in the early detection and management of waterborne diseases.
ii. To develop a predictive system that aids healthcare providers in identifying high-risk patients for waterborne diseases, enabling prompt diagnosis and treatment.
iii. To evaluate the predictive performance of the machine learning model using metrics such as accuracy, precision, recall, and F1-score, ensuring the reliability of individual-level disease risk assessments.
iv. To predict the likelihood of waterborne disease infection in individuals based on real-time clinical data.

1.5 Significance of the study

The contribution of this project would be of interest to individual, community, and government. Individuals get knowledge about how to safeguard themselves and their family against diseases spread by water since they are more informed about the importance of clean drinking water and value of good personal hygiene, leading to healthier lifestyle and reduced personal and economic burden of waterborne diseases on families. Additionally, it will also help the communities' and government by promoting a shared understanding on the significance of water quality, sanitation and hygiene. This understanding will guide in public health initiatives, infrastructure investment, and emergency response strategies.

1.6 Scope of the study

This project aims to explore the frequency, causes, impacts, and preventive methods of waterborne illnesses in rural and urban areas in Kenya. the study will take some considerations such as fostering collaboration between multidisciplinary teams while addressing moral, legal and performance enhancing concerns.

1.7 Assumptions in the study

In conducting this research, the following assumptions were made that:

1. Those affected by waterborne illnesses have equal access to healthcare services
2. The current interventions such as water treatment and sanitation improvement are either effective or ineffective, depending on the study's hypothesis.
3. The environmental factors such as temperature and water quality influence the spread of waterborne illnesses remain constant throughout the study period.
4. waterborne illnesses affect various populations similarly, regardless of socioeconomic position, age, or immunity.
5. There is reliable data on diseases caused by water such as infection rate and geographic distribution are accurate and readily available.

1.8 Limitations of study

The study's limitations include its reliance on representativeness and quality of data, potential biases resulting from small demographic samples, difficulties interpreting complex models with outdated training data, ethical issues related to data privacy, the need for rigorous clinical validation, the risks associated with an over-reliance on automation, challenges in managing uncommon conditions, and integration issues with current healthcare systems. Additionally lack of models to help in prediction of the occurrence and ways to tackle this diseases.

2 CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction.

Water is vital for health but can pose serious risks when contaminated, leading to waterborne diseases, particularly in areas with poor hygiene, sanitation, and limited access to clean drinking water. These diseases are a significant health threat, especially to vulnerable populations like children and the elderly. Despite interventions such as improved sanitation, handwashing, safe water storage, and chlorination, waterborne illnesses remain a major public health issue, hindering socioeconomic development. This project aims to use advanced technologies like machine learning predictive modeling to forecast waterborne disease outbreaks. By utilizing historical and real-time data, the project will help raise awareness and analyze the environmental, economic, and social factors influencing disease frequency.

2.2 Waterborne Diseases and Risk Factors.

Manetu et al. (2021) emphasize the need for comprehensive strategies to combat waterborne diseases by improving water supply, sanitation, hygiene practices, and public health interventions. The study highlights the importance of technological innovations, such as low-cost filtration systems and real-time water quality monitoring, to enhance these efforts. Integrating these technologies into existing public health strategies can make communities more resilient to waterborne illnesses. The review also calls for further research to fill gaps in current intervention approaches and develop more effective solutions.

Mutono et al. (2021) stress the importance of better surveillance of health trends in urban Africa, particularly regarding piped water access. The study reveals weak correlations between current study designs and health outcomes, calling for stronger research methodologies. It also underscores the disparity in waterborne disease prevalence, with rural and low-income communities disproportionately affected due to inadequate sanitation and water infrastructure. The article proposes an integrated approach that combines enhanced

water systems, community health education, and targeted treatment programs to address these issues and ensure more effective public health interventions.

2.3 Socio-environmental Factors and Practices.

Mugendi et al. (2023) emphasize the importance of proper water handling, particularly in communities like Yatta, where poor water management practices contribute to higher incidences of food- and water-borne illnesses. The study calls for community-based education campaigns to promote safe water handling, including boiling or treating water, proper storage, and maintaining hygiene. Targeting caregivers, especially those responsible for young children, could significantly reduce illness prevalence in these areas. These interventions are vital to protecting vulnerable groups and improving public health.

Mazuze et al. (2023) highlight the impact of socio-environmental factors such as geography, poverty, education, and infrastructure on vulnerability to waterborne diseases. Rural communities, informal settlements, and areas affected by climate change are at higher risk due to limited access to clean water and sanitation. Inequities in socioeconomic status exacerbate the problem, as lower-income households lack the resources for proper water treatment. The authors advocate for targeted interventions for high-risk groups and emphasize the need for collaborative efforts between local communities, public health organizations, and governments to reduce the prevalence of waterborne diseases.

2.4 Water Supply and Sanitation Infrastructure.

Ferreira et al. (2021) highlight the long-term health benefits of strategically investing in clean water and sanitation infrastructure. Their study found that targeted, efficient spending on water and sanitation significantly reduced hospital admissions, indicating that infrastructure improvements directly impact public health. The research stresses the need for governments and international organizations to prioritize infrastructure improvements in areas most vulnerable to waterborne illnesses, suggesting that coordinated public health campaigns and infrastructure investments can amplify the positive outcomes.

Ibeto and Obaje (2022) examine the role of a high-quality water supply in reducing waterborne diseases at Ibrahim Badamasi Babangida University Lapai. They emphasize the importance of strict water quality standards and routine inspections to ensure public health and safety. The research underscores the value of ongoing monitoring and investment in water quality management, with findings indicating that access to clean water reduces the incidence of waterborne infections. The authors also recommend incorporating water safety education into health awareness programs to further mitigate risks and improve long-term water management strategies.

2.5 Preventive and Intervention Strategies.

Chan et al. (2021) examine the effectiveness of primary preventive measures in preventing waterborne diseases, particularly in areas lacking access to clean water. They emphasize the importance of Water, Sanitation, and Hygiene (WASH) interventions, which have proven effective in preventing disease transmission in vulnerable areas. In addition to WASH, the study highlights the critical role of community education in promoting proper hygiene, sanitation, and water management practices. The authors argue that while education is essential, sustained public health benefits require infrastructure improvements, such as clean water supplies and sewage systems.

Chauque et al. (2023) focus on low-cost point-of-use water treatment technologies as an accessible solution for reducing waterborne diseases in low-resource settings. Technologies like ceramic filters, chlorine disinfection, and solar water disinfection (SODIS) provide affordable ways to enhance home water safety. Furthermore, Meki, Ncube, and Voyi's (2022) review stresses the importance of combining water filtration, immunization campaigns, and hygiene promotion to reduce diarrheal diseases. Their findings suggest that multi-pronged strategies, including vaccination and collaboration, are essential for reducing waterborne illnesses at the community level.

2.6 Climate Change and Environmental Factors.

The study by Manaseh et al. (2023) investigates the factors influencing cholera risk in Nairobi County, emphasizing the critical role of socioeconomic status, environmental conditions, and public health infrastructure. The research highlights that poor WASH (Water, Sanitation, and Hygiene) conditions in informal community's increase cholera risk, especially for marginalized groups with limited access to healthcare. To mitigate these risks, the study recommends focused interventions that address environmental factors like clean water availability and drainage, as well as addressing the socio-economic determinants of health.

Fadhil, Ismail, and Tosepu (2021) examine the relationship between climate change and the prevalence of waterborne illnesses, revealing that factors such as increased rainfall, flooding, and temperature variations significantly influence the spread of waterborne pathogens. They explain how severe weather events, like floods, can contaminate drinking water, while temperature changes affect pathogen survival and reproduction rates. The authors stress the need for further research to understand these dynamics and recommend improved monitoring systems, better water quality management, and adaptive measures in vulnerable communities to combat the health risks associated with climate change.

2.7 Conceptual Theoretical Frameworks and Academic Case Studies.

The article highlights the growing challenges posed by rapid urbanization in sub-Saharan Africa, where poor water and sanitation infrastructure exacerbate the spread of waterborne diseases. The authors propose an innovative approach by dividing metropolitan areas into distinct exposure zones, allowing for a more nuanced understanding of the factors influencing waterborne disease risk. By integrating this zonal analysis with the urban exposome paradigm, which considers all environmental exposures, they aim to improve the design of public health interventions tailored to urban challenges.

The study at Moi University underscores the urgent need for targeted public health initiatives to address the high frequency of waterborne illnesses. The research reveals significant differences in disease incidence among demographic groups, particularly

highlighting gender disparities linked to access to clean water, sanitation, and health education. It calls for a comprehensive strategy involving university administrations, health authorities, and community organizations to improve water quality, sanitation practices, and hygiene education. The study advocates for multifaceted solutions to reduce the prevalence of waterborne diseases and enhance public health in university settings.

2.8 Research Gap.

While many studies focus in the prevention of waterborne illnesses, primarily in the development of inventive and reasonably priced water treatment options for underdeveloped and rural areas. There is a need to fully understand the long-term effectiveness, scalability, and integration of new technologies like water purification systems and low-cost filtration techniques. Additionally, there is need in embracing new technologies like using machine learning models and artificial intelligence models for monitoring the progress of this efforts in suppressing waterborne diseases. This is archived by accurate collection of data which will be used in the development of this models.

2.9 Summary.

The literature underscores the critical role of community education as a preventive tool alongside Water, Sanitation, and Hygiene (WASH) initiatives. Educating communities about proper hygiene, sanitation, and water management significantly enhances their ability to prevent waterborne diseases. Technological innovations such as low-cost filtration methods, water purification systems, and real-time data monitoring of water quality are also promising in addressing these challenges, making communities more resilient to waterborne diseases.

The review stresses the need for governments, NGOs, and international organizations to prioritize infrastructure improvements in the most affected regions. However, it does not explore the potential benefits of advanced technological models, such as artificial intelligence and machine learning, in early disease detection, prediction, and progress monitoring. These innovations could play a pivotal role in the prevention and management of waterborne diseases, especially for vulnerable groups in high-risk communities, enhancing the effectiveness of ongoing public health efforts.

3 CHAPTER THREE
RESEARCH METHODOLOGY

3.1 Introduction

This research aims to develop a predictive model for managing waterborne illnesses in Kenya using data analytics and machine learning. By collecting and analyzing health data, the project seeks to identify patterns, trends, and risk factors to support evidence-based interventions, especially in vulnerable communities. The model will be tested for accuracy and reliability to ensure practical applicability. Ultimately, the goal is to improve public health outcomes through targeted and efficient disease control strategies.

3.2 Research design

This study uses a combination of descriptive and machine learning approaches to analyze waterborne disease patterns in Kenya. It examines the prevalence, distribution, and contributing factors of illness to uncover key insights and provide data-driven recommendations for public health policies and resource allocation, especially for vulnerable populations. By analyzing clinical symptoms and infection trends, the model aims to identify hidden correlations. These insights will support proactive decision-making and targeted disease prevention efforts.

3.3 Data Collection Methods

This research relies entirely on secondary data obtained from credible online sources to support the analysis of waterborne diseases in Kenya. Health records from hospitals provide valuable insights into disease incidence, patient demographics, and treatment outcomes, offering a clear understanding of how waterborne illnesses affect different populations. The integration of these diverse secondary data sources creates a strong foundation for identifying patterns, assessing risk factors, and informing public health interventions aimed at reducing the burden of waterborne diseases.

3.4 Data Preparation

3.4.1 Data Cleaning Methods

Data cleaning ensures the accuracy and reliability of data by removing duplicates, transforming data, and identifying outliers. Duplicate entries are identified and removed because they have the potential to provide erroneous estimates of sickness occurrence. Resolving inconsistencies is a key step in data cleaning, especially when data comes from multiple sources. Hospitals, health centers, and environmental agencies may use different formats, units, or terminology. For example, the same disease might be labeled differently, like "cholera" and "acute diarrhea". Such variations must be standardized to ensure accurate analysis.

3.4.2 Handling Missing Values

It is essential to handle missing values in waterborne diseases data to avoid data gaps that might skew results and lower prediction accuracy. Mean replacement is used for quantitative data, where missing values are substituted with the mean to preserve the general data trend without distorting results. Imputation techniques specific to data types handle this problem. When dealing with categorical data, such the types of water sources, closest neighbor imputation is used to maintain realistic distributions by filling in the gaps with values from related records. When combined, these techniques guarantee an accurate and comprehensive dataset that allows thorough examination of trends in waterborne disease and associated environmental variables.

3.4.3 Data Transformation Procedures

Data transformation techniques are essential in preparing dataset for accurate analysis by enabling comparability across diverse variables. This procedure makes it possible to compare data more directly across scales, which is essential for integrating environmental measurements with variables like illness incidence rates. Normalization methods are also used, especially for data that might vary greatly, such the quantity of rainfall or the population density in impacted areas. Normalization enhances the accuracy and stability of machine learning models used in predictive analysis by scaling these values between 0 and 1, which lessens the influence of extreme values.

3.5 Data Analysis Techniques

3.5.1 Statistical Methods

To gain a basic understanding of the dataset, statistical methods will be utilized to compile and examine the data aspects. To provide a broad overview of the main trends and variability in numerical data, descriptive statistics such as mean, median, and standard deviation will be calculated. Frequency distributions and cross-tabulations will also be used to analyze the relationships between categorical variables in order to look for any potential patterns or connections pertaining to waterborne disease indicators.

3.5.2 Machine Learning Algorithms

Machine learning algorithms such as decision trees, random forests, and logistic regression play a vital role in predicting and understanding waterborne diseases. By analyzing environmental and socioeconomic factors like temperature, rainfall, population density, and infrastructure quality, these models help identify high-risk regions. Decision trees and random forests reveal the most influential variables by modeling complex relationships, while logistic regression estimates the likelihood of outbreaks under specific conditions. These insights enable health authorities to prioritize interventions and develop targeted strategies for preventing the spread of waterborne illnesses. Mohammad, F. et al (2023).

3.5.3 Tools and Software

To accurately predict waterborne illnesses from secondary data sources, a comprehensive approach to data analysis will be implemented using Python. This will include utilizing packages like Pandas and NumPy for effective data manipulation and analysis. Visualization tools such as Matplotlib and Seaborn will be employed to create informative plots that aid in understanding both the data and model results. Machine learning algorithms will be implemented using Scikit-learn to develop predictive models, while SQL will be used for managing databases, ensuring data integrity, and efficiently retrieving and searching large datasets.

3.6 Model Building

3.6.1 Model selection

To determine which machine learning algorithm is best suited for the predictive task, the first step in the model-building process is model selection. Machine learning provides a reliable way to achieve high accuracy in forecasting the presence of waterborne diseases. This model analyses important factors including clinical symptoms, vital signs, and infection duration, to support early detection using machine learning methods like random forests and gradient boosting. In conclusion, this strategy helps to reduce and prevent the spread of waterborne diseases by enabling public health organizations to more effectively allocate resources and proactively identify infected patients.

3.6.2 Training and Testing

The training and testing phases are essential for developing an accurate model for predicting waterborne disease outbreaks. The model is trained using real-time clinical data, such as symptoms, heart rate, blood pressure, and infection duration, to identify correlations between variables. The testing phase, which uses fresh, untested data, is then employed to assess the model’s ability to generalize to real-world situations. Performance indicators like accuracy, precision, recall, and F1 score are used to evaluate the model’s effectiveness in predicting patient status.

3.6.3 Evaluation Metrics

Machine learning models are evaluated using accuracy, precision, recall, and F1-score to assess their effectiveness in predicting waterborne infections and identifying high-risk areas. The F1-score balances precision and recall, while recall measures the model's ability to detect all potential outbreaks. Accuracy reflects the overall success rate, and precision indicates the model's reliability in predicting disease events. Confusion matrices are also used to present the classification results, offering a detailed view of the model's performance. These evaluation tools ensure that the chosen model is reliable and practical for real-world applications in forecasting waterborne illnesses.

3.7 Ethical Considerations

3.7.1 Data Privacy and security

Data security and privacy are critical when creating predictive models for waterborne disease outbreaks, especially when handling sensitive survey data. To protect individual privacy, personally identifiable information is anonymized through deletion or encryption, and only de-identified, aggregated data is used for analysis. Strong security measures, such as encryption, access controls, and regular audits, safeguard stored data and prevent unauthorized access. All procedures comply with local privacy laws and regulations, like the GDPR, ensuring that access is limited to authorized personnel. These safeguards foster trust, encourage participation, and maintain ethical standards, ultimately enhancing the quality and reliability of public health research.

3.7.2 Ethical implications of the research

Ethical considerations are crucial in predicting waterborne disease outbreaks, ensuring respect for participants' autonomy and privacy. Informed consent is prioritized, ensuring participants understand the research goals, scope, and how their data will be used. The study seeks approval from institutional review boards or ethics committees to ensure the research respects individual rights, minimizes harm, and aligns with ethical standards. By adhering to these ethical guidelines, the study upholds integrity, builds participant trust, and enhances the reliability and quality of the findings.

3.8 Limitations

3.8.1 Potential biases

Restricting data collection to specific demographics or regions can introduce selection bias, which may distort results by under-representing high-risk populations and regions. This bias can weaken the model’s predictive power across different contexts. Response bias is also a concern, particularly in self-reported surveys, where inaccurate data may arise from memory issues, social desirability, or misinterpretation of questions. To mitigate these biases, cross-verifying responses, using diverse sampling methods, and incorporating objective data sources can improve the model's accuracy and ensure that high-risk areas for waterborne illnesses are more reliably identified.

3.8.2 Limitations in data and methodology

Data gaps in historical records pose a significant challenge, as they can undermine the reliability of trend analysis and pattern recognition, leading to inaccurate estimates and potentially misguided public health measures. Additionally, the complexity of environmental and socioeconomic factors, such as population density, climate, and sanitation, complicates the identification of clear causal relationships in waterborne disease outbreaks. This multifactorial nature can lead to misinterpretations of the data. To address these challenges, rigorous data validation, combining multiple data sources, and applying advanced statistical techniques are essential to enhance the reliability of studies and ensure more accurate predictions.

3.9 Summary

In conclusion, the research's methodological approach emphasizes the application of a descriptive-analytical methodology to systematically investigate the dynamics of waterborne diseases in Kenya. In order to reflect the complex nature of the problem, the research places a strong emphasis on a thorough approach to data collecting, relying on a variety of sources, such as historical medical records, environmental data, and socioeconomic indicators. To improve the findings' robustness and enable a more thorough comprehension of disease distribution patterns, sophisticated analytical methods like machine learning and geospatial analysis are used. This methodical approach guarantees that every stage is rational and clear, tackling all the details that come with studying waterborne illnesses. The study intends to provide trustworthy insights that may guide public health initiatives and support more successful disease management plans in Kenya by combining a variety of data sources and advanced analysis techniques.

4 CHAPTER 4
RESULTS AND DISCUSSIONS

4.1 Introduction

This study uses clinical variables including symptoms, vital signs, and illness duration together with machine learning and data-driven methodology to forecast waterborne illnesses in Kenya. The study considers and examines important factors including body temperature, fever, blood pressure, and heart rate using only secondary data from reliable internet sources, such as public health records. In order to find trends and evaluate the diagnostic use of these indicators, the technique entails data preparation, exploratory analysis, and predictive modelling.

4.2 Importing required libraries

The libraries are essential for data science and machine learning tasks. pandas is used for data manipulation and analysis, seaborn and matplotlib.pyplot help in creating visualizations. For machine learning, sklearn provides various models and tools, including SVC (Support Vector Classification), MultinomialNB (Naive Bayes), LogisticRegression, RandomForestClassifier, and DecisionTreeClassifier for classification tasks, along with tools like TfidfVectorizer for text feature extraction, train_test_split for data splitting, and StandardScaler for feature scaling. Pipeline allows you to chain preprocessing and modeling steps, while ColumnTransformer applies different transformations to different columns. Evaluation metrics like classification_report and accuracy_score help assess model performance. These libraries streamline the workflow of data preprocessing, modeling, and evaluation.

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report
import matplotlib.pyplot as plt
import seaborn as sns
from imblearn.combine import SMOTETomek
from sklearn.model_selection import train_test_split
from collections import Counter
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import SVC
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from imblearn.pipeline import Pipeline as ImbPipeline
import joblib
```

4.3 Loading the dataset

	id	disease_name	type	symptoms	causes	transmission_mode	treatment	prevention	heart_rate	blood_pressure_systolic	bl
0	1	Tuberculosis	Communicable	Cough, chest pain, weight loss, fever	Mycobacterium tuberculosis bacteria	Airborne	Antibiotics	Vaccination, good hygiene	98	128	
1	2	Lung Cancer	Non-communicable	Persistent cough, chest pain, weight loss	Smoking, genetic factors	NaN	Surgery, chemotherapy, radiation	Avoid smoking, regular screenings	80	123	
2	3	Cholera	Water-borne	Severe diarrhea, dehydration, vomiting	Vibrio cholerae bacteria	Contaminated water	Rehydration, antibiotics	Safe drinking water, sanitation	60	132	
3	4	Influenza	Communicable	Fever, chills, sore throat, muscle aches	Mycobacterium tuberculosis bacteria	Airborne	Antibiotics	Vaccination, good hygiene	77	107	
4	5	Diabetes	Non-communicable	Increased thirst, frequent urination, fatigue	Smoking, genetic factors	NaN	Surgery, chemotherapy, radiation	Avoid smoking, regular screenings	68	116	

4.4 Performing Exploratory Data Analysis

Viewing dataset size and columns (data.shape, data.info(), data.describe()).

Checking for null values (data.isnull().sum()).

Creating scatter plots, pair plots, count plots for visual patterns.

Identifying outliers and potential data errors.

```
data.isnull().sum()
```

```
id                          0
disease_name             2045
type                      108
symptoms                 1985
causes                   2018
transmission_mode        8015
treatment                2018
prevention               2018
heart_rate                  0
blood_pressure_systolic     0
blood_pressure_diastolic    0
fever                       0
body_temperature            0
duration_of_infection       0
infection_status            0
dtype: int64
```

Figure 1

waterborne_diseases [], highlights all the diseases that are caused by contaminated water based on the symptoms a patient exhibit

so as to form a new column that affirms if patient is suffering from waterborne represented by 1 and not represented by 0

```
data = data.dropna()
data.isnull().sum()
```

```
id                          0
disease_name                0
type                        0
symptoms                    0
causes                      0
transmission_mode           0
treatment                   0
prevention                  0
heart_rate                  0
blood_pressure_systolic     0
blood_pressure_diastolic    0
fever                       0
body_temperature            0
duration_of_infection       0
infection_status            0
dtype: int64
```

The table gives a quick statistical summary of your DataFrame's numeric columns (by default). Here's what it shows: count — total number of non-null (non-missing) values in each column. mean — the average value of each column. std — the standard deviation (how spread out the values are). min — the smallest value in the column. 25% — the value at the 25th percentile (first quartile). 50% — the median (middle value). 75% — the value at the 75th percentile (third

quartile). max — the largest value in the column.

	id	heart_rate	blood_pressure_systolic	blood_pressure_diastolic	body_temperature	duration_of_infection
count	11663.000000	11663.000000	11663.000000	11663.000000	11663.000000	11663.000000
mean	10193.373832	79.890251	114.758124	75.099889	38.002555	12.400326
std	5718.926553	11.719070	14.625593	9.028275	1.163330	9.954630
min	1.000000	60.000000	90.000000	60.000000	36.000000	0.000000
25%	5420.000000	70.000000	102.000000	67.000000	37.000000	2.000000
50%	10267.000000	80.000000	115.000000	75.000000	38.000000	12.000000
75%	15112.500000	90.000000	127.000000	83.000000	39.000000	21.000000
max	19999.000000	100.000000	140.000000	90.000000	40.000000	30.000000

The bar chart shows the distribution of the is_waterborne classes in the dataset. Each bar represents the number of samples labeled as either:

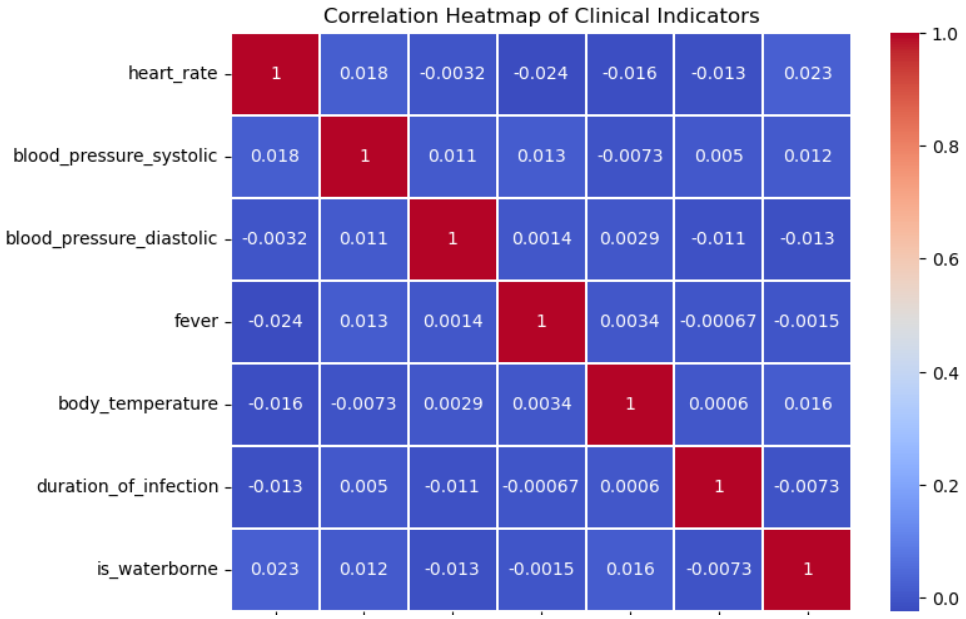
0 → Not Waterborne

1 → Waterborne

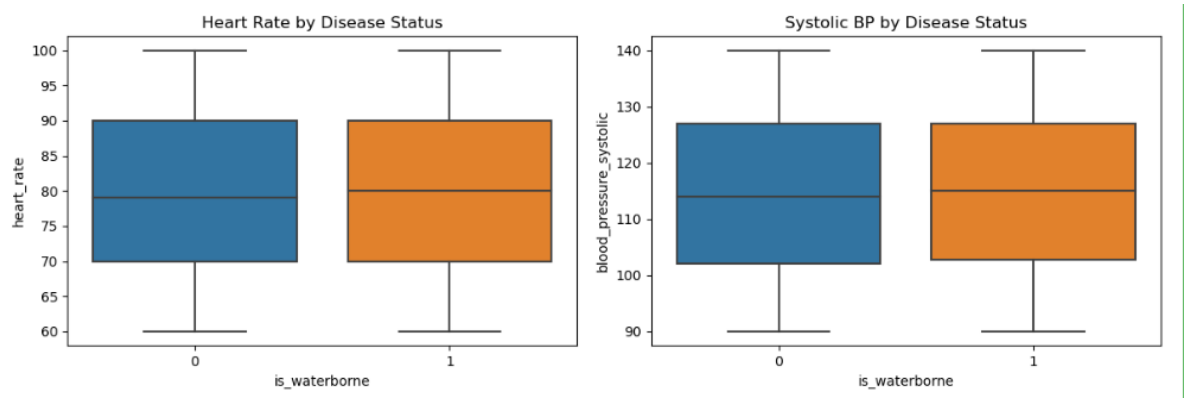
If the bars are roughly the same height, it means the dataset is balanced — both classes have a similar number of records.

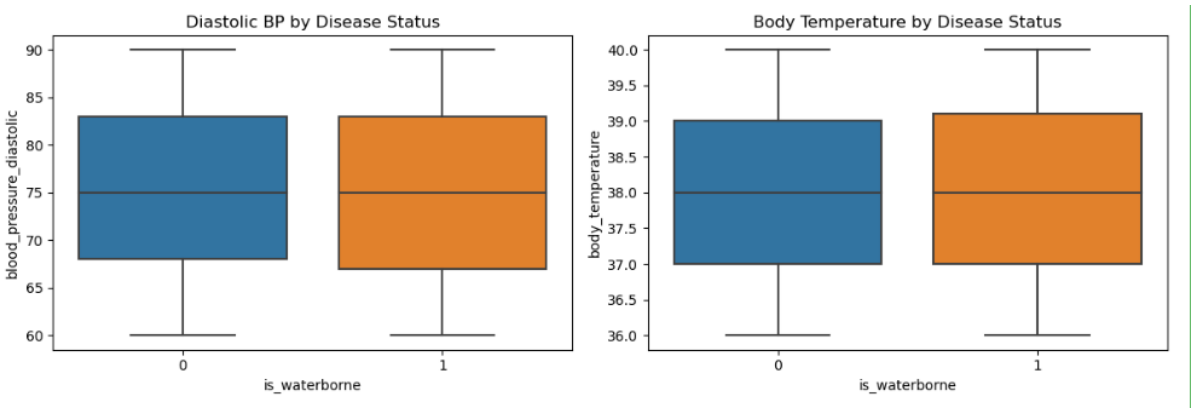
If one bar is much taller than the other, it indicates class imbalance — one class dominates the dataset. This can be important because imbalanced data can cause machine learning models to become biased toward the majority class.

The correlation heatmap illustrates the strength and direction of linear relationships between clinical indicators (such as heart rate, blood pressure, fever, body temperature, and infection duration) and the target variable the presence or absence of a waterborne disease.



The number of times a person's heart beats per minute (bpm). When someone is suffering from waterborne diseases dehydration, fever, or infection stress can elevate the heart rate. Blood Pressure Systolic is the top BP pressure in your arteries when your heart beats. In cases of severe dehydration BP systolic pressure usually drops below 90 mmHg Blood Pressure Diastolic is the low BP pressure in your arteries when your heart rests between beats. Dehydration and fluid loss from infections usually reduce diastolic pressure.





This code selects a specific set of important features from the dataset to keep the analysis clean and focused.

Then, it defines and runs a function visualize_data() that creates a pair plot using Seaborn.

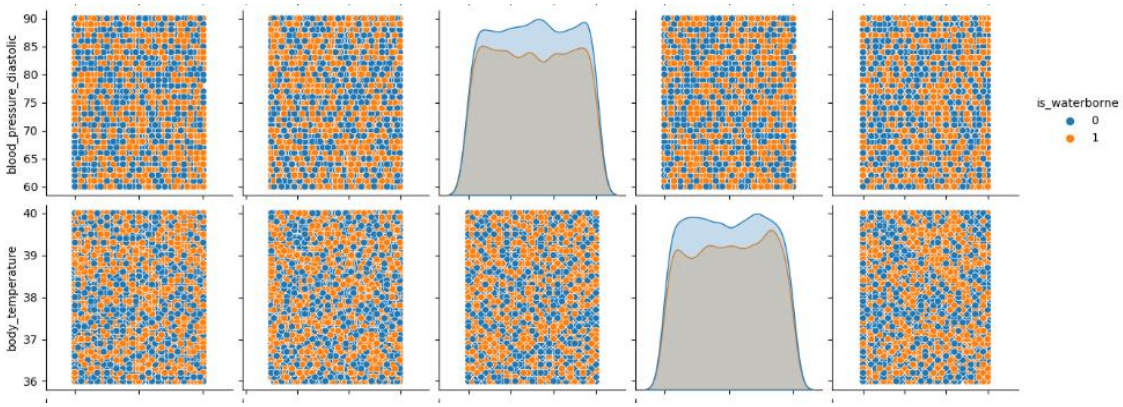
The pair plot shows:

How each feature relates to every other feature?

How the data points are grouped based on the is_waterborne class (using different colors)?

The diagonal of the plot shows smooth distribution curves (KDE) for each individual feature.

To visually explore relationships between features and see if waterborne and non-waterborne cases form clear patterns or clusters.



4.5 Model development

4.5.1 Splitting dataset before training the model

X = data.drop('is_waterborne', axis=1)

y = data ['is_waterborne'],, ## it is the targeted column for prediction

This line of code splits the dataset into training and testing sets.

X_train and y_train → data used to train the model.

X_test and y_test → data used to test the model's performance on unseen data.

The test_size=0.2 means:

→ 20% of the data will be used for testing, and 80% for training.

This code prepares your data for machine learning by splitting it into two parts: one for learning and one for evaluation

4.6 Training Machine Learning model.

I trained and evaluated a Support Vector Machine (SVM) classifier by fitting it to the preprocessed training data and using it to predict labels for the test set. I then assessed the model's performance using a classification report (including precision, recall, and F1-score) along with overall accuracy.

4.7 Classification Report

Precision (per class):

Class 0: 0.99 → Out of all predictions labeled as 0 (Not Waterborne), 99% were correct.

Class 1: 1.00 → Out of all predictions labeled as 1 (Waterborne), 100% were correct.

→ Interpretation: The model makes very few false positives.

Recall (per class):

Class 0: 1.00 → Out of all true 0s in the test set, the model correctly found 100%.

Class 1: 0.99 → Out of all true 1s in the test set, the model correctly identified 99%.

The model almost never misses a real case.

F1-Score (per class):

Class 0: 1.00

Class 1: 0.99

The F1-score balances precision and recall and according to the output the both are extremely high, meaning the model handles both false positives and false negatives very well.

Support (per class):

Class 0: 1259 actual samples in test data.

Class 1: 1074 actual samples in test data.

This shows the number of samples in each class that is useful to judge whether your classes are balanced.

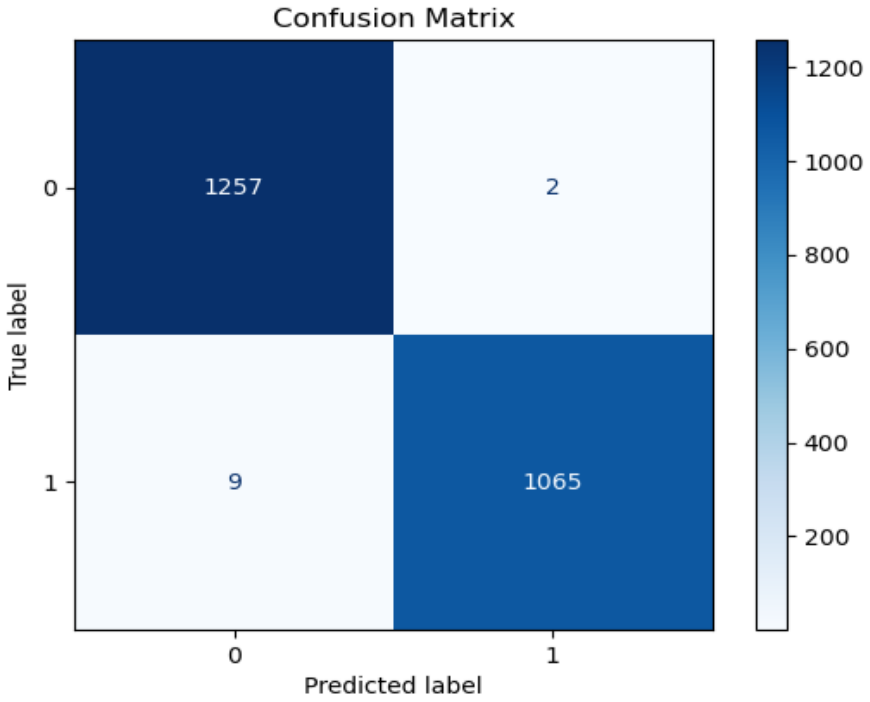
SVM Classification Report:

	precision	recall	f1-score	support
0	0.99	1.00	1.00	1259
1	1.00	0.99	0.99	1074
accuracy			1.00	2333
macro avg	1.00	1.00	1.00	2333
weighted avg	1.00	1.00	1.00	2333

4.8 Confusion Matrix.

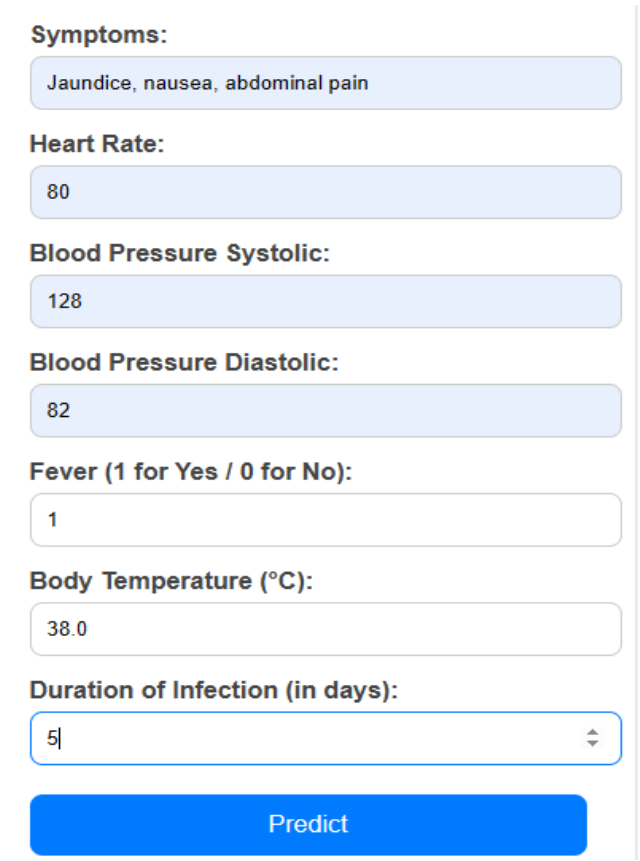
The confusion matrix shows that the Support Vector Machine (SVM) model performed exceptionally well on the test data. Out of 2,333 total samples, the model correctly classified 1,257 non-waterborne cases and 1,065 waterborne cases. Only 2 non-waterborne cases were incorrectly predicted as waterborne (false positives), and 9 waterborne cases were incorrectly classified as non-waterborne (false negatives).

This extremely low number of misclassifications indicates that the model is both highly accurate and reliable for distinguishing between waterborne and non-waterborne cases. The results align with the classification report, which showed near-perfect precision, recall, and F1-scores, confirming the model's strong generalization ability on unseen data.



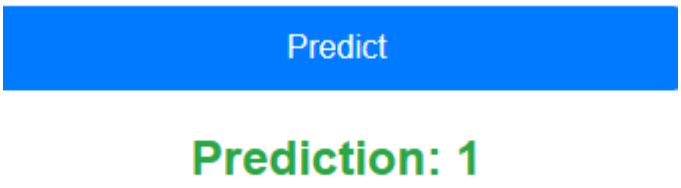
4.9 Model Deployment

4.9.1 Creating a GUI window Example



The output of the predictive model 1 indicates the patient is suffering from waterborne diseases whereas 0 indicates absence of waterborne disease

4.9.2 The output of prediction



5 CHAPTER FIVE
RECOMMENDATION AND CONCLUSION

5.1 Recommendation

In areas where outbreaks of waterborne diseases are common, it is advised that medical facilities include this prediction model into their standard patient screening procedures. The ongoing gathering of high-quality clinical data and recurring retraining to adjust to changing illness patterns are also essential to the model's performance. If public health officials want to improve epidemic planning and response even further, they should think about integrating this approach with community-level surveillance systems. To optimize this prediction tool's practical impact, cooperation among data scientists, medical professionals, and legislators will be crucial.

5.2 Conclusion

Using clinical data to create a machine learning-based prediction model for detecting possible waterborne disease infections provides a proactive way to improve early diagnosis and patient outcomes. Utilizing critical health indicators including fever state, infection duration, symptoms, and vital signs, the model can help medical professionals make prompt and well-informed judgements. This strategy not only improves early intervention capabilities but also eases the strain on medical facilities by facilitating more effective case prioritization. The predictive approach has the potential to significantly reduce the severity and spread of waterborne infections, particularly in vulnerable populations throughout Kenya, if it is completely implemented.

REFERENCES

Agingu, J. B. (2020). Levels and differentials of occurrence of water borne diseases at Moi University, Kenya.S

Alexandre, Z., Rafael, C. D., & Pedro, A. G. (2021). Conceptual frameworks regarding waterborne diseases in sub-Saharan Africa and the need of for a new approach to urban exposomes. Epidemiology and Health, 43, 1–6.

Ali, M., Nelson, A. R., Lopez, A. L., & Sack, D. A. (2020). Updated global burden of cholera in endemic countries. PLoS Neglected Tropical Diseases, 14(9), e0007920.

Branz A, Levine M, Lehmann L, Bastable A, Ali SI, Kadir K, Yates T, Bloom D, Lantagne D (2017) Chlorination of drinking water in emergencies: a review of knowledge to develop recommendations for implementation and research needed. Waterlines 36(1):4–39.

Chan, E. Y. Y., Tong, K. H. Y., Dubois, C., Mc Donnell, K., Kim, J. H., Hung, K. K. C., & Kwok, K. O. (2021). Narrative review of primary preventive interventions against water-borne diseases: scientific evidence of health-EDRM in contexts with inadequate safe drinking water. International Journal of Environmental Research and Public Health, 18(23), 12268.

Chauque, B. J. M., Issufo, M., Benitez, G. B., Cossa, V. C., Chauque, L. G. H., Stauber, C. E., Benetti, A. D., & Rott, M. B. (2023). Why do low-cost point-of-use water treatment technologies succeed or fail in combating waterborne diseases in the field? A systematic review. Journal of Environmental Chemical Engineering, 110575.

Chinebu, T. I., Okafor, K. C., Anoh, K., Uzoeto, H. O., Apeh, V. O., Okafor, I. P., Adebisi, B., & Okoronkwo, C. A. (2024). Smart waterborne disease control for a scalable population using biodynamic model in IoT network. Computers in Biology and Medicine, 181, 109034.

Fadhil, M., Ismail, M., Tosepu, R., & others. (2021). An impact of climatic change on water-borne diseases: a review. IOP Conference Series: Earth and Environmental Science, 755(1), 12081.

Ferreira, D. C., Graziele, I., Marques, R. C., & Gonçalves, J. (2021). Investment in drinking water and sanitation infrastructure and its impact on waterborne diseases dissemination: The Brazilian case. Science of the Total Environment, 779, 146279.

Hussain, M., Cifci, M. A., Sehar, T., Nabi, S., Cheikhrouhou, O., Maqsood, H., Ibrahim, M., & Mohammad, F. (2023). Machine learning based efficient prediction of positive cases of waterborne diseases. BMC Medical Informatics and Decision Making, 23(1), 11.

Ibeto, I. A., & Obaje, A. (2022). Quality Water As A Preventive Measure Of Water Borne Diseases: A Study Of Ibrahim Badamasi Babangida University Lapai Water Supply. Lapai Journal Of Politics, 8(2), 44–52.

Islam, M. S., Hassan-uz-Zaman, M., Islam, M. S., Clemens, J. D., & Ahmed, N. (2020). Waterborne pathogens: Review of outbreaks in developing nations. Waterborne Pathogens, 43–56.

Landrigan PJ, Stegeman JJ, Fleming LE, Allemand D, Anderson DM, Backer LC, Rampal P (2020) Human health and ocean pollution. Ann Glob Health 86(1):151.

Lantagne D, Yates T (2018) Household water treatment and cholera control. J Infect Dis 218(suppl_3): S147–S153.

Magana-Arachchi, D.N.; Wanigatunge, R.P. Ubiquitous Waterborne Pathogens. In Waterborne Pathogens; Elsevier: Amsterdam, The Netherlands, 2020; pp. 15–42.

Manaseh, B. A., John, G., Simon, K., & Christine, M. (2023). Mapping cholera risk in Nairobi County, Kenya: a comprehensive analysis of environmental, socio-economic, and WASH factors. African Journal of Health Sciences, 36(3), 286–298.

Manetu, W. M., & Karanja, A. M. (2021). Waterborne disease risk factors and intervention practices: a review. Open Access Library Journal, 8(5), 1–11.

Mazuze, H., Almendra, R., & Santana, P. (2023). A systematic literature review on factors of socio-environmental vulnerability associated with water-borne diseases. The Journal of Infection in Developing Countries, 17(12), 1658–1666.

Meki, C. D., Ncube, E. J., & Voyi, K. (2022). Community-level interventions for mitigating the risk of waterborne diarrheal diseases: a systematic review. Systematic Reviews, 11(1), 73.

Mugendi, P., & others. (2023). Caregivers Food-water Handling Practices, Knowledge and Prevalence of Food-water Borne Diseases in Chidren (1-5 Years) of Yatta, Kenya: a Case of Sand Dams.

Mutono, N., Wright, J. A., Mutembei, H., Muema, J., Thomas, M. L. H., Mutunga, M., & Thumbi, S. M. (2021). The nexus between improved water supply and water-borne diseases in urban areas in Africa: a scoping review. AAS Open Research, 4.

Ncube, E. J., Voyi, K. V. v, & others. (2022). Community-level interventions for mitigating the risk of Waterborne diarrheal diseases: a systematic review.

Ngowi, H. A. (2020). Prevalence and pattern of waterborne parasitic infections in eastern Africa: a systematic scoping review. Food and Waterborne Parasitology, 20, e00089.

Ntajal, J.; Falkenberg, T.; Kistemann, T.; Evers, M. Influences of Land-Use Dynamics and Surface Water Systems Interactions on Water-Related Infectious Diseases—A Systematic Review. Water 2020, 12, 631.

Ondieki, J. K., Akunga, D. N., Warutere, P. N., & Kenyanya, O. (2022). Socio-demographic and water handling practices affecting quality of household drinking water in Kisii Town, Kisii County, Kenya. Heliyon, 8(5).

Shayo, G. M., Elimbinzi, E., Shao, G. N., & Fabian, C. (2023). Severity of waterborne diseases in developing countries and the effectiveness of ceramic filters for improving water quality. Bulletin of the National Research Centre, 47(1), 113.

Szálkai, K. (2023). Water-borne diseases. In The Palgrave Encyclopedia of Global Security Studies (pp. 1540–1546). Springer.

WHO (2021) Emerging technologies to combat waterborne diseases: advances in water purification and monitoring. World Health Organization Bulletin, 99(2), 98-105.

World Health Organization (WHO 2022) waterborne diseases and health.

APPENDICES

Budget

#	ITEMS	DESCRIPTION	COST
1.	Equipment	Laptop 8GB RAM Windows 10 or above.	Ksh. 40,000.
2.	Software tools	Anaconda/Jupyter Notebook/AWS/pycharm/VsCode/ Google Colab/Spyder/Github/Git/Render.	Ksh. 0
3.	Internet	Strong internet connection	Ksh. 4500
4.	Miscellaneous	Transport/lunch/snacks	Ksh. 1500
5.	Documentation	Printing project documents	Ksh. 600
TOTAL			**Ksh. 46,600**

Work plan

Activities/ weeks	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12
Data Collection												
Data Preprocessing												
Feature Engineering												
Model Development												
Model Training												
Model Testing												
Model Deployment												