

**PREDICTING WATER POLLUTION CAUSED BY GLASS PARTICLES USING
MACHINE LEARNING TECHNIQUES
BY**

**Md. Monir Hossain Reyad
ID:191-15-12183**

**Sabbir Siddiki
ID:191-15-12379**

**Md.Kawshik Ahmed
ID:191-15-12125**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Ms. Israt Jahan
Lecturer
Department of CSE
Daffodil International University
Designation



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH


JANUARY 23RD, 2023

APPROVAL

This Project/internship titled “Predicting Water Pollution Caused by Glass Pollution using Machine Learning Techniques.”, submitted by Monir Hossain, Sabbir Siddiki & Kawshik Ahmed”, ID No: Student ID 191-15-12183,191-15-12379 & 191-15-12125 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 23rd January, 2023.


BOARD OF EXAMINERS

Chairman


Dr. Touhid Bhuiyan
Professor and Head

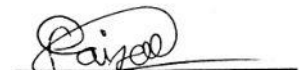
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner


Dr. Md. Zahid Hasan
Associate Professor


Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner


Fahad Faisal
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner


Dr. Ahmed Wasif Reza
Associate Professor

Department of Computer Science and Engineering
East West University

DECLARATION

I hereby declare that this project has been done by us under the supervision of **Ms. Israt Jahan**, **Lecturer Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for an award of any degree or diploma.

Supervised by:

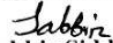


Ms. Israt Jahan
Lecturer
Department of CSE
Daffodil International University

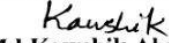
Submitted by:



Md. Monir Hossain Reyad
ID: 191-15-12183
Department of CSE
Daffodil International University



Sabbir Siddiki
ID: 191-15-12379
Department of CSE
Daffodil International University



Md. Kawshik Ahmed
ID: 191-15-12125
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty Allah for his divine blessing made me that possible to complete the final thesis successfully.

I am really grateful and wish my profound indebtedness to **Ms. Israt Jahan, Lecturer, Department of CSE**, Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of “Machine Learning” to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading and many inferior drafts, and correcting them at all stages have made it possible to complete this thesis.

I would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan, Head, Department of CSE**, for his kind help to finish my project and also to other faculty members and the staff of the CSE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents and brother.

ABSTRACT

Water is polluting with different glass particles in Bangladesh. Most of the glass manufacturing industries through their wastage in the river. Because of the indiscriminate release of some industrial, domestic, and mining effluents into the environment and also on living species, water pollution is now a global issue. Numerous strategies for the control, remediation, cleaning, and purification of the water system from the source and at the end of the delivery line should have been developed as a result of the negative impacts of water pollution on humans, animals, and the ecosystem. Among all the methods used are membrane separation, biological precipitation, adsorption, and photo catalysis. Besides, the development of water purification methods that are less expensive, more cost-effective, and simpler to operate are crucial. Because they have the potential to lessen the surface tension that exists between two immiscible liquids, surfactants and bio surfactants are used in the processes of water treatment. Bio surfactants derived from natural sources have gained attention due to their low cost, low impact on the environment, and unique properties that make them useful in conjunction with nano materials to boost their activity and performance. The use and performance of bio surfactant nanomaterial systems in water purification processes are the subject of this review. Water samples from the water source near some glass manufacturing companies and tested the water. With the report of these tested water, a dataset of different glass particles is created. Then after this, using machine learning techniques like Naive Bayes, KNN, and Random Forest algorithm different models are created and compared their accuracy of predicting the water pollution caused by the glass particles. Among them Naïve Bayes model performs with 92% accuracy whereas the KNN model performs with 93.6% accuracy and Random Forest stands out with 96% accuracy which is higher than the other models.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgments	iii
Abstract	iv
CHAPTERS	
CHAPTER 1: INTRODUCTION	1-6
1.1 Introduction	1
1.2 Objective	2
1.3 Motivation	3-4
1.4 Rationale of the Study	4-5
1.5 Research Questions	5
1.6 Expected Outcome	5
1.7 Report Layout	6
CHAPTER 2: BACKGROUND	7-10
2.1 Introduction	7
2.2 Overview	8
2.3 Related Work	9
2.4 Research Summary	9
2.5 Scope of the Problem	10
2.6 Challenges	10

CHAPTER 3: RESEARCH METHODOLOGY	11-26
3.1 Introduction	11
3.2 Working Procedure	12
3.2.1 Data Collection	13
3.2.2 Preprocessing	14
3.2.3 Exploratory Data Analysis	15-18
3.2.4 Algorithms	19-22
3.2.5 Feature Generation	22-23
3.2.6 Model Creation	26
CHAPTER 4: EVALUATION AND COMPARISON	22-25
4.1 Introduction	22
4.2 Result and Analysis	22
4.3 Comparison	23-25
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	26-28
5.1 Impact on Society	26
5.2 Impact on Environment	26-27
5.3 Ethical Aspects	27
5.4 Sustainability Plan	28
CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND FUTURE WORK	29-31
6.1 Summary of the Study	29
6.2 Conclusion	30
6.3 Recommendation	30
6.4 Future Work	31

LIST OF FIGURES

FIGURES		PAGE
Figure3.2.1	Working Procedure of the Research	13
Figure3.2.1.1	Sample of Collected Dataset	14
Figure3.2.3.1	Describing the Dataset	15
Figure3.2.3.2	Heatmap of the Dataset	16
Figure 3.2.3.3	Histogram (bar graphical representation)	17
Figure 3.2.3.4	Pair Plot (pairwise relationship)	18
Figure3.2.4.1	Procedure of Naïve Bayes Algorithms	20
Figure3.2.4.2	Procedure of KNN Algorithm	21
Figure3.2.4.3	Procedure of Random Forest Algorithm	22
Figure3.2.6.1	Model Creation for different Algorithms	23
Figure4.2.1	Accuracy of Naïve Bayes Algorithm	25
Figure4.2.2	Accuracy of KNN Algorithm	25
Figure4.2.3	Accuracy of Random Forest Algorithm	26
Figure 6.4.1	Future Work Diagram	34

LIST OF TABLES

TABLES

PAGE

Table4.3.1	Comparison of the Accuracy between the Algorithms	26
------------	---	----

LIST OF GRAPHS

GRAPHS

PAGE

Graph4.3.1	Comparison of the Accuracy between the Algorithms	27
------------	---	----

LIST OF ABBREVIATION

<u>Short Form</u>	<u>Full Form</u>
DIU	Daffodil International University
CSE	Computer Science & Engineering
ML	Machine Learning
RQ	Research Question

CHAPTER 1

INTRODUCTION

1.1 Introduction

Water is crucial to human health. In actuality, the human body needs mineral water to function. But a number of factors contribute to water pollution in our nation. Fresh water is in scarce supply as a result. Out of Bangladesh's four million wells, 1.12 million still contain arsenic. Poor water quality has a significant influence on public health. Arsenic poisoning currently accounts for one out of every five fatalities in Bangladesh. In addition, there are several ways that water can be contaminated, including by dangerous glass particles. These dangerous glass particles enter the body of humans who drink water containing them, leading to a number of ailments.

Different types of glass pollution contaminate the water in Bangladesh. The bulk of glass-producing industries dump their waste into rivers. As a result, glass particles are combined with water and contaminate it. Glass particles are particularly dangerous to human health. Bangladesh is a riverine country, thus if the rivers' water is tainted with dangerous glass particles, we're in danger. Due to the danger that the glass particles pose to human health. To get around this, water testing was done near various glass manufacturing companies and experiments were conducted. The results of the water test are used to construct a dataset of different glass particles. Following that, machine learning methods like the Naive Bayes, KNN, and Random Forest algorithms are used to examine the precision with which various models forecast how glass particles will affect water pollution. While the KNN model performs with 94% accuracy and the Random Forest model shines out with the accuracy of 96%, the Naive Bayes model performs with 92% exactness. To forecast the water pollution brought on by glass particles such as silicon dioxide (SiO_2), sodium oxide (Na_2O), aluminum oxide (Al_2O_3), magnesium oxide (MgO), and calcium oxide (CaO), we generated various datasets to forecast whether or not there is pollution.

1.2 Objective

The two types of objectives we have are narrative and numerical. Narrative objectives are generally accepted definitions of the required levels of water quality for industry, watershed management, and pollution control. They also serve as the foundation for the creation of specific numerical goals. However, the development of numerical goals came first in order to control the diverse effects of pollutants in the water column. Environmental science studies and two decades of regulatory experience have shown that protecting beneficial uses requires monitoring and controlling pollution levels in all areas of the aquatic system. The Regional Board is actively working toward a comprehensive set of goals, which will ensure the protection of all existing and potential beneficial uses. These goals include a numerical declaration of goals. Numerical objectives, on the other hand, typically characterize pollutant concentrations, the physical and chemical characteristics of the water itself, and the water's toxicity to aquatic species. These goals are intended to indicate the maximum amount of pollutants that can be present in the water column without negatively affecting people who consume the water, aquatic organisms that utilize it as a habitat, or any current or potential beneficial uses. The extensive biological, chemical, and physical partitioning data reported in the scientific literature, national water quality standards, studies carried out by other agencies, and data gathered from local environmental and discharge monitoring are the technical foundations of the region's water quality objectives. Although the Regional Board acknowledges that in some circumstances there is insufficient evidence to define precise numerical objectives, the Regional Board considers that a conservative approach to creating objectives has been appropriate in these circumstances. At many various levels of objective generation and implementation of the water quality control plan, in addition to the technical evaluation, the overall viability of achieving targets in terms of technological, institutional, economic, and administrative elements is taken into consideration.

Beside of these we will also have these arguments over this paper by-

- To predict the water pollution caused by glass particles.
- To build a model which can predict the glass pollution in water.
- To aware people about the harmful effect of these glass particles contaminated with water.
- To decrease the pollution of water caused by awaking the glass manufacturing industries not to through the pernicious glass particles.

1.3 Motivation

Researchers have identified a number of causes for these issues, including excessive water use and water pollution. Water contamination will therefore be covered in this study. The two main sources of water contamination, which are natural and human, are well detailed in this study. First, there are natural resources like lava flows, animal excrement, and storm-generated sand. Human resources, such as factory waste dumping in the water, increase nitrate content in water sources, which in turn raises pH and other water structures that have an impact on water utilization. In the past, this issue has been caused by the overuse of pesticides and the disposal of sewage into the water. It was determined that this issue was critical and needed to be resolved. Numerous strategies have been developed to address water pollution using water filtering techniques like coagulation and flocculation, as well as to identify it using various tools like the Nephelo meter, which measures turbidity. This issue has a number of negative side consequences, including serious health issues for both humans and animals, the demise of aquatic life, and water that is unfit for drinking or watering crops. In this study, I will demonstrate every method that has been used to determine if water may be used or not in accordance with certain factors like pH, dissolved oxygen, electrical conductivity, and turbidity.

People now have higher standards for living quality due to the increase of the social economy and population. As a result, freshwater resources are now being used much more frequently. Human water demand has increased eight times in the last century. The primary influence on the earth's natural water cycle now comes from human behavior, which has caused irreparable harm and has an impact that even goes beyond what the earth's ecology can sustain. Urban regions with concentrated populations and

industries have severe water supply and demand conflicts. In the twenty-first century, the availability of water has emerged as a significant barrier to social and economic progress. In order to combat water shortages, alternate water sources must be used in addition to freshwater

Desalinated water, rainwater collecting, reuse of storm water, and recycled water are currently the resources that can be utilized to replace water. Recycled water, which is collected from wastewater treatment, is a stable resource and is not impacted by seasons like storm water and rainwater are. Compared to desalinated water, its manufacture uses less energy. Additionally, during the production phase, recycled water might reduce local water environment contamination. Recycled water's source is wastewater, which disgusts the general public. As a result, it is less accepted by the public than other water sources due to worries about indoor air pollution and health. Therefore, the degree of local citizens' acceptance of recycled water reuse must be taken carefully when using recycled water as a substitute water resource.

As glass particles are very hazardous to human body so, we decided to collect different places water sample and tested them in the laboratory. With the report of this test, we created a dataset of glass particles which are responsible for water pollution. Using this dataset, we created different machine learning models and fit the dataset into these models and then check the accuracy of all the models. All these efforts are only to aware mass people who have known knowledge about the water pollution caused by these various glass particles. So, the main motivation of this study is to help people know about the harmful effects of these glass particles. Besides, we can also reduce the water pollution, if our model can predict whether there is any pollution or not.

1.4 Rationale of the Study

The natural formation of nano materials from complex materials and their increasing global application are considered major concerns. In terms of its micro and nano size, glass contamination has not received much attention until now. As of now, similar to all products, glass is abused and creates squander in measures of millions of tons each year. However, the primary concern is that glass recycling is significantly lower than that of manufactured glass; only one-fourth of all glass produced worldwide is that size. As a result, research on the possibility of glass particles breaking down into nano- and micro-sized glass materials and accumulating in the environment is

urgently required. The purpose of this review is to provide a comprehensive summary of the current state of knowledge regarding the possible degradation, recycling, formation, and environmental fate of micro/nano glass. It likewise covers the conceivable hypothetical issues related with glass defilement in the sea-going environmental elements with conversation of momentum research holes. Aside from that, the likely effects of micro/nano glass on the ecosystem of the soil, its interaction with plants, and the possibility of its movement in the food chain. Our mechanistic outline directs current and upcoming research on this pressing topic[1]

1.5 Research Question

We have selected some basic question on our research work which is being answered stepwise if anyone is interested to ask.

- What are the glass particles that causes pollution of water?
- How we can aware people about the harmful effects of these glass particles?
- Which are the facts that causes the water pollution?
- What are the processes of overcome this problem?

1.6 Expected Outcome

We had to learn how the glass particles were causing water contamination via this research. Thus, in order to determine whether or not water is polluted, we produced a dataset of different glass particles. The overall results might also be informed as follows:

- Can determine the pollution level of the water cause by different glass particles.
- Can reduce the use of the hazardous glass particles finding alternatives for that.
- By the prediction we can also aware people about the harmful effect of it.
- Can find a way to decrease the pollution of the water by predicting it and aware the people who are related with glass industries.

1.7 Report Layout

The report structure includes a synopsis of each chapter that is used in the research study. Below is a quick synopsis of each chapter:

Chapter 1: Discussion of the study's purpose, research question, and anticipated results is covered in.

Chapter 2: Discussion on the background and extent of the difficulties and challenges, as well as an outline of the thesis and associated work.

Chapter 3: Exploratory data analysis, data collection, preparation, and introduction to research methodology.

Chapter 4: Discussion of the model development process, algorithms, outcomes, and algorithm comparisons in

Chapter 5: Discussion of the research, including potential future applications, and thesis conclusion.

CHAPTER 2

BACKGROUND

2.1 Introduction

Water resources are plentiful in Bangladesh, one of the world's most densely populated nations, but they are constantly being polluted. Both sources of surface water and groundwater are contaminated with various pollutants such as coli forms, hazardous trace metals, and other organic and inorganic pollutants. Since the majority of people rely on these water sources, particularly the country's groundwater supplies, which have elevated levels of arsenic, drinking water carries a very high risk of health problems. Bangladesh has a high rate of water-borne illness deaths, especially among youngsters. Water contamination is mostly caused by anthropogenic sources, including untreated industrial effluents, inappropriate home waste disposal, and agricultural runoff. An assessment of the nation's overall water contamination levels and the causes of this serious condition is crucial to evaluate public health risk. For this purpose, we reviewed hundreds of well recognized international and national journals, conference proceedings and other related documents to draw a complete picture of recent water pollution status and its impact on public health; also, the sources of water pollution are identified.

Because glass particles pose a serious risk to human health, we chose to gather water samples from several sites and test them in a lab. Using the test report, we produced a dataset of glass particles that contribute to water contamination. With this dataset, we built a variety of machine learning models, added the data to these models, and then assessed the correctness of each model. All of these initiatives are merely meant to inform the public about the water contamination caused by these different glass particles. The main objective of this study is to inform the public about the dangers of these glass particles. Additionally, if our model can determine whether there is or not pollution.

2.2 Overview

The purpose of this study is to provide such a model or algorithm which can detect the amount of glass particles whose are responsible for the pollution of water. By which we can reduce the water being polluted due to rottenness. Moreover, this will help the researchers by providing the detection much more easily. This study's secondary sources of data include government publications from various agencies, academic journals, books, and websites, as well as the Bangladesh Water Control Board's annual report. The location of water testing laboratories (Arsenic, Fluoride, Iron, and ph value), a map of the district's central PWS scheme, and a map of an ongoing PWS scheme allow us to view the true situation in some districts. Research has shown that water pollution has an impact on not just human sickness and mortality, but also the entire ecosystem. In the research area, soil erosion, industry, urbanization, and overpopulation are the main contributors to water contamination. We can go without food for a short while, but not water. The survival of all organisms depends on it. All living things, including humans, depend on it for food production, economic growth, and survival. Water covers two thirds of the earth's surface. A whopping 98% of the water is sea water, making it unsafe to drink due to the high salt content. Fresh water makes up about 2% of the world, but glaciers and polar icecaps contain 1.6% of it. Aquifers and wells contain an additional 0.36%. Therefore, only 0.036% of the water on the earth may be found in lakes and rivers. The quality of surface and ground water, as well as its regional and seasonal availability, have a significant impact on the environment, economic growth, and development. Human activities have an impact on water quality, which is deteriorating as a result of causes like population expansion, urbanization, agricultural development, and others. Because its effects last for a long time, polluted water affects future generations as well as the lives of those living today. If a body of water is contaminated, all living things and people are forced to drink it since they have no other choice. It causes tumors, birth defects, and other disorders, and it has an impact on their skin, lungs, brain, liver, and kidneys.

2.3 Related Work

In November 2013, Deepak Pant et al., proposed research on Pollution because of hazard glass pollution. Discussion on hazard glass pollution is pervasive on a global scale, both with regard to quantity or quality and related health issues. Mercury and lead are the main contaminants in fluorescent lamp and cathode ray tube (CRT) are the types of waste glass. Nanoparticles such as Fe, Zn, Se, Cu, Ni alone or with the formulation can deactivate the heavy metals presence is also discussed the paper[2].

In 2019, S. Mohurli et al. proposed an index on A KNN Classifier Study to Predict the Index of Water Pollution. This study is discussed about the necessity of determining the proportion arises because drinking water may contain a variety of parameters in varying proportions. The fundamentals of the KNN classifier and the current state of drinking water are examined in this paper. It likewise concentrates on the utilization of k-closest neighbor classifier to anticipate and gauge the precision of the extent of boundaries accessible with regards to the quality record[3].

In August 2021, Hemalatha Nambisan et al., published a book on Prediction of Plastic Degrading Microbes. In this book, Plastic is declared as one of the major polluted elements of the environment. . Using several machine learning methods such decision trees, random forests, support vector machines, and k-nearest neighbors, they have developed some prediction models of bacteria that degrade plastic. Among these, the prediction was made with 99.1% accuracy using the Random Forest model[4].

2.4 Research Summary

In Bangladesh, various glass particles pollute the water. The majority of industries that produce glass dispose of their waste in the river. Glass particles, which are extremely harmful to the human body, are therefore mixed with water and contaminate it. Water samples were taken from a water source close to some glass manufacturing facilities, and the water was tested. A dataset of various glass particles is created using the water test report. Then after this, utilizing machine learning methods like Naive Bayes, KNN, and Random Forest, various models are made and observed their exactness of anticipating the water contamination caused by the various glass particles.

2.5 Scope of the Problem

As it is a proposed model, so there must be some scope of problems. This model provides its predictions based on the images. Firstly, model needs to be trained and then it will be able to generate the output if a given fruit is fresh or rotten. So, it is clearly seen that this process is so lengthy. Secondly, it takes such less time to take a fresh fruit to become rotten, so this can be also a scope of problem in this study.

2.6 Challenges

A lot of challenges are being face during this research. Collection of the image data was the big challenge. After collecting the images from different sources, then we need to preprocess the data to be ready for fit into the model which was also challenging.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

To predict the water pollution caused by the glass particles, firstly we have collected some sample water and tested them. With the help of the report of this tested report we have created a dataset which is called data collection. Then after this, we need to preprocess the data to become ready for the prediction. After this, exploratory data analysis is applied. In this section, we need find the correlation between the predictor class and targeted class using heatmap.

3.2 Working Procedure

Step 1: Determine the issue.

Step 2: Review the Literature.

Step 3: Identify the Issue.

Step 4: Explanation of Terms and Concepts.

Step 5: Define the Population.

Step 6: Create the instrumentation plan

Step 7: Gather Information.

Step 8: Analyze the data

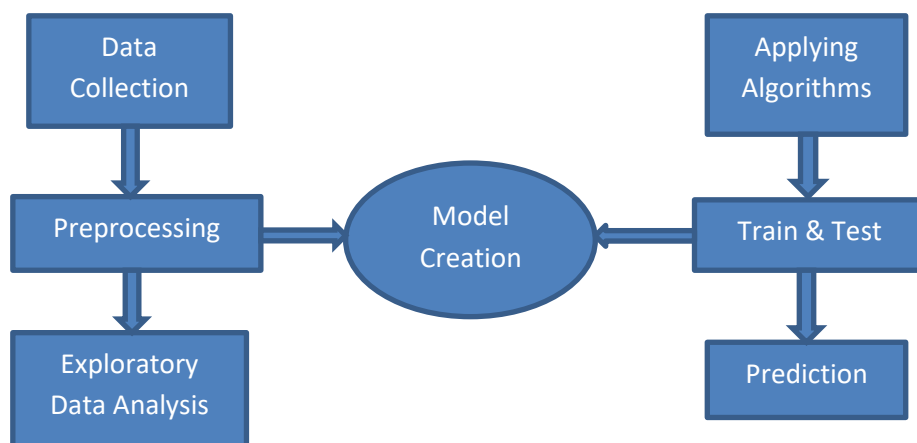


Fig.3.2.1: Working Procedure of the Research

3.2.1 Data Collection

In order to answer specific research-based questions, test hypotheses, and assess results, data collecting is the act of gathering and measuring information on variables of datasets depending on interest. Here in our research, we wanted to predict the pollution of water by the elements which are produced by the glass products. We have researched over 1100+ water and we gave the outcome as percentages evaluating them as polluted or not being polluted by those elements (SiO₂, Na₂O, Al₂O₃, MgO, CaO)

```
df = pd.read_csv('/content/gdrive/MyDrive/Glass Pollution/Glass_Pollution.csv')
df.head(10)
```

	SiO ₂	Na ₂ O	Al ₂ O ₃	MgO	CaO	Polluted
0	65.7	13.4	6.8	3.3	4.1	No
1	73.2	11.5	5.3	2.9	6.3	No
2	68.1	3.2	8.1	3.7	2.9	No
3	83.6	19.2	5.5	4.3	9.2	Yes
4	78.3	15.4	7.2	4.7	3.4	No
5	63.8	14.3	4.7	3.3	6.9	No
6	88.7	9.6	6.7	4.6	8.1	Yes
7	91.4	11.5	7.3	4.1	4.5	Yes
8	68.1	3.2	8.1	3.7	2.9	No
9	82.6	19.8	11.5	3.3	5.2	Yes

Fig.3.2.1.1: Sample of Collected Dataset

3.2.2 Preprocessing

Data preprocessing is a machine learning or data mining approach used to convert the raw data into a format that is both practical and effective. We will preprocess our dataset in this part. To start, we must determine whether the dataset contains any default null values. Then mapping of the targeted class will be performed. As we have used categorical data in our dataset so we must have convert the categorical targeted class into numerical values using mapping function.

3.2.3 Exploratory Data Analysis

Exploratory data analysis (EDA) is primarily a method for examining data sets and summarizing their key features, frequently utilizing statistical graphics and other approaches for data visualization. EDA differs from traditional hypothesis testing of particular elements or objects in that it is primarily used to explore what the data can tell us beyond the formal modeling. A statistical model can be utilized or not.

```
[ ] #Describe  
df.describe()
```

	SiO2	Na2O	Al2O3	MgO	CaO	Polluted
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
mean	73.890000	12.751000	6.915000	3.918000	5.202000	0.260000
std	7.431104	5.317372	1.849454	0.543163	2.046627	0.440844
min	63.800000	3.200000	4.300000	2.900000	2.200000	0.000000
25%	68.100000	9.550000	5.700000	3.300000	3.400000	0.000000
50%	70.900000	13.400000	6.300000	3.900000	5.200000	0.000000
75%	78.300000	15.400000	8.200000	4.300000	6.900000	1.000000
max	91.400000	23.200000	11.500000	4.800000	9.200000	1.000000

Fig.3.2.3.1: Describing the Dataset

Heatmap of our Dataset

A heatmap uses a grid of colored squares to display values for the main variable of interest across two general axis variables. A bar chart or histogram is used to divide the dataset's axis variables into ranges, and the color of each cell represents the value of the principal variable in the corresponding cell range. In the given chart we declared 1 as pollution of the water is Yes and others of less than 1 is less polluted or not polluted. After evaluating the dataset we had found that the number of elements of SiO₂ and Na₂O is higher than other components of Glass.

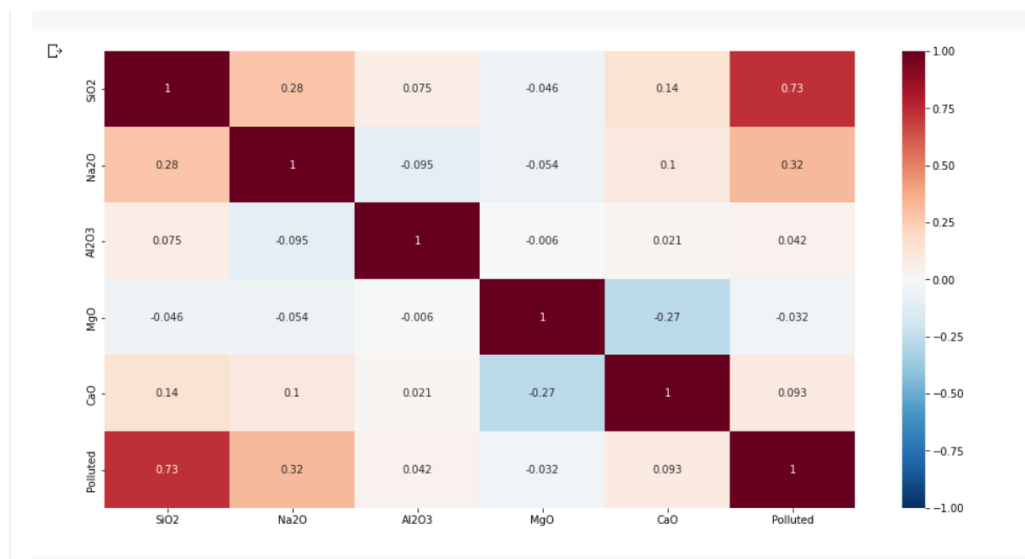


Fig.3.2.3.2: Heatmap of the Dataset

Histogram (bar graphical representation)

A histogram essentially shows how frequently continuous variables in a dataset occur. A bar graph, on the other hand, is a diagrammatic comparison of discrete variables. The bar graph shows categorical data of the elements of the glass components, whereas the histogram below shows numerical data. These are mainly used in statistics to demonstrate the number of occurrences of specific variables within a predetermined range and the effect of pollution in water.

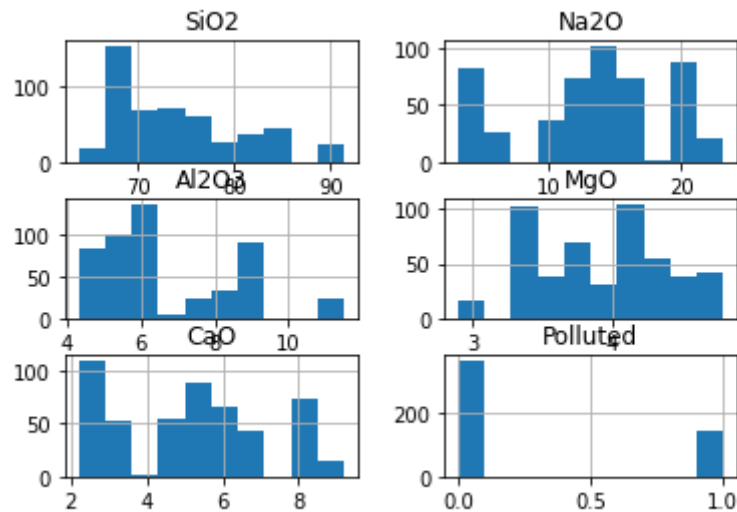


Fig. 3.2.3.3: Histogram (bar graphical representation)

Pair Plot (pair wise relationship)

With the same data set we collected earlier for our research, we attempted to establish a grid of axes in this function so that each numerical variable in the data will be shared over the y-axes across a single row and the x-axes across a single column

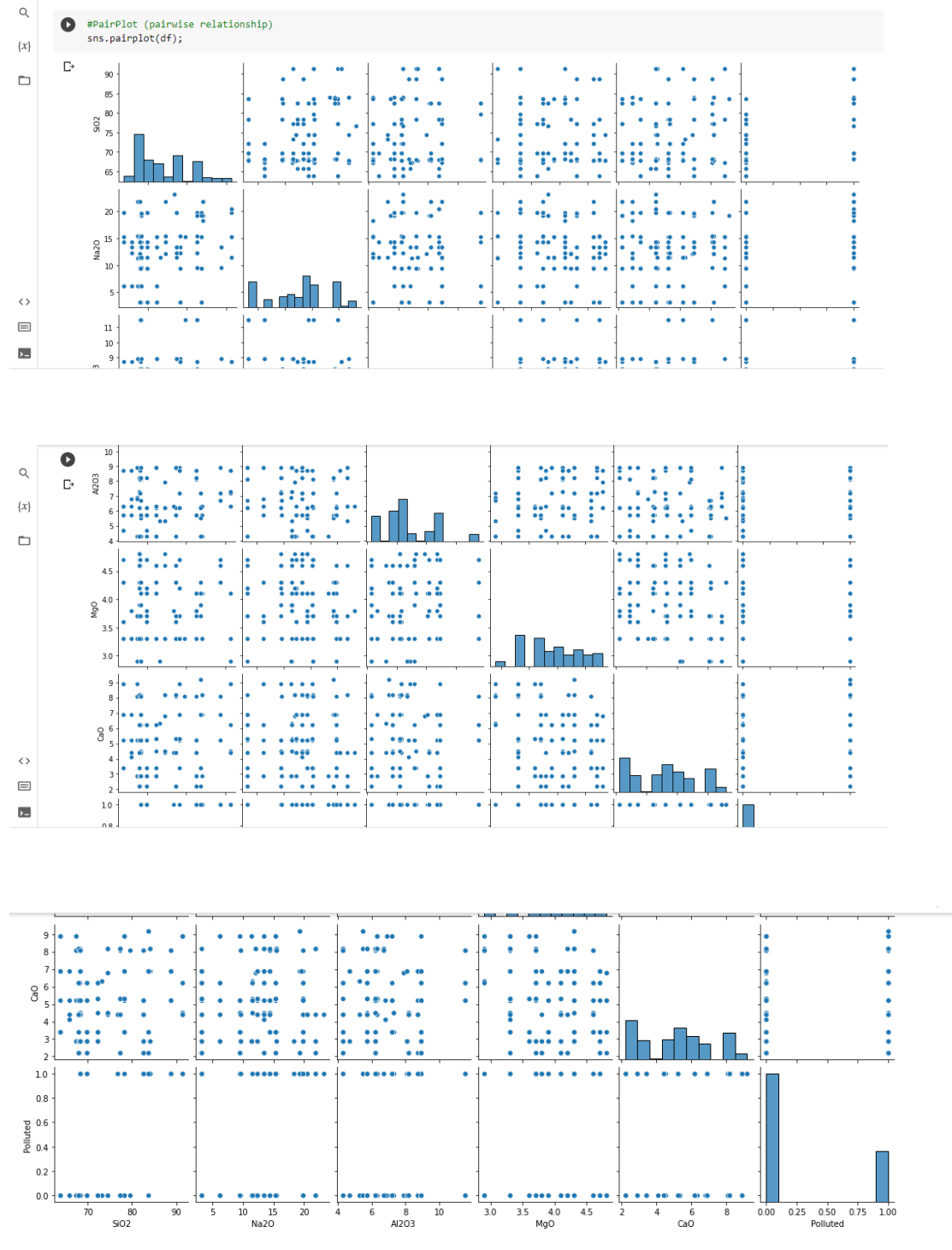


Fig.3.2.3.4 Pair Plot (pairwise relationship)

3.2.4 Algorithms

As we have use categorical data, so most popular algorithms for predicting categorical data have been used here. Those are:

- i. **Naïve Bayes Algorithm**
- ii. **K-Nearest Neighbor Algorithm**
- iii. **Random Forest Algorithm**

i. **Naïve Bayes Algorithm:**

Based on the "Bayes Theorem," the Naive Bayes algorithm is a supervised learning technique that is frequently used to tackle classification problems. It is mostly employed to classify text using a large training dataset. Being able to quickly create machine learning models that are capable of producing quick predictions, Naive Bayes Classifier is one of the simplest and most effective classification algorithms[5].

Being a probabilistic classifier, it bases its predictions on the likelihood of each object. The Bayes Theorem is used by the Naive Bayes algorithm to forecast object probabilities. The equation below represents the Bayes theorem as follows:

$$P(X|Y) = \frac{P(Y|X) * P(X)}{P(Y)} \dots\dots\dots (1)$$

Here,

- **P(X|Y) is Posterior probability:** Probability of assumptionX on the observed event Y.
- **P(Y|X) is Likelihood probability:** Probability of the authentication given that the probability of an assumption is true.
- **P(X) is Major Probability:** Probability of assumption before observing the authentication.
- **P(Y) is Minor Probability:** Probability of Authentication.

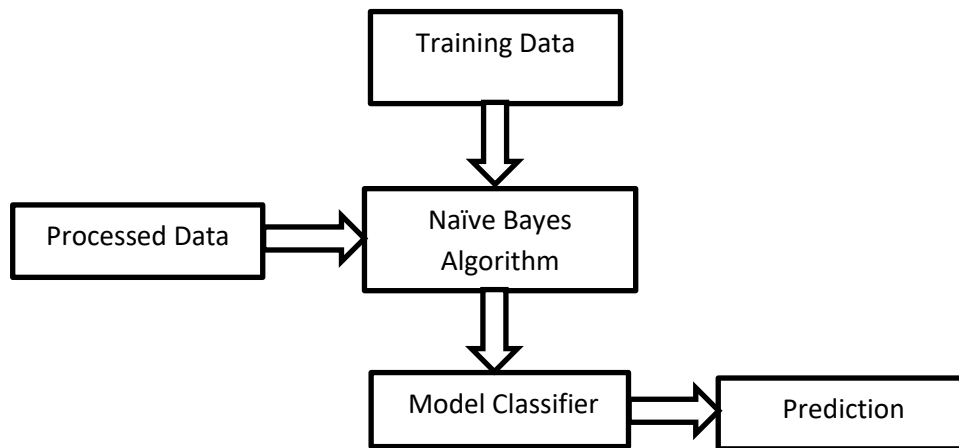


Fig.3.2.4.1: Procedure of Naïve Bayes Algorithm

ii. K-Nearest Neighbor Algorithm:

The K-Nearest Neighbor algorithm is the most basic Supervised Learning-based on Machine Learning technique. Based on how similar or unlike they are to one another, the K-NN algorithm assigned the new case or data to the category that is closest to the existing categories. The K-NN method uses similarity to categorize new data points and stores all of the existing data. This demonstrates that when new data appears in a dataset, the K-NN algorithm can categorize it quickly and accurately. In addition to this ,The K-NN algorithm can be used for both classification and regression, but it is mostly used for classification problems[6].

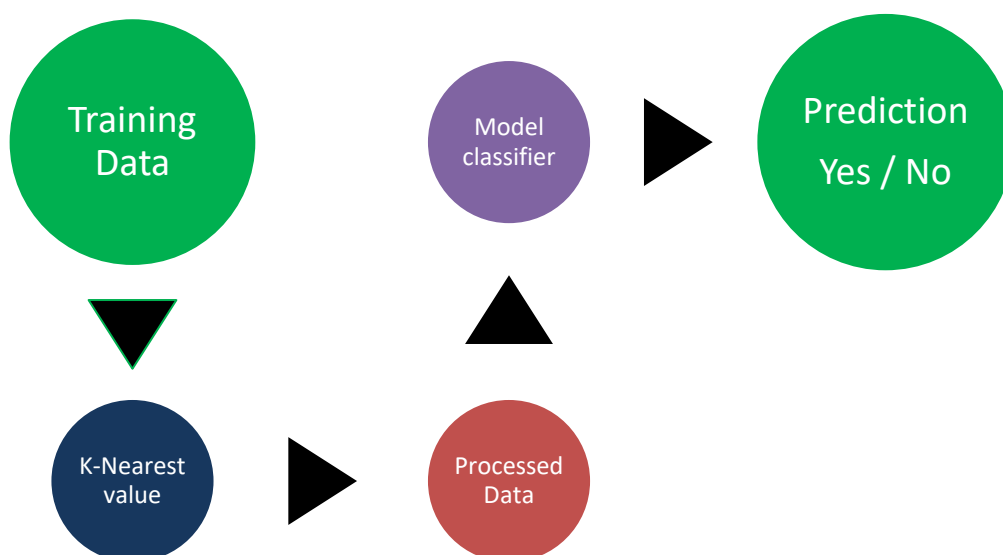


Fig.3.2.4.2: Procedure of KNN Algorithm

iii. Random Forest Algorithm:

One of the well-known supervised machine learning algorithms or methods is the Random Forest algorithm. It can be applied to regression and classification problems in ML. It is primarily based on the concept of ensemble learning, which combines several classifiers to solve complicated problems and improve the performance of the model. Random Forest is a particular kind of classifier that uses a number of decision trees on various subsets of the provided dataset and averages them to improve the dataset's predicted accuracy. Random Forest classifier predicts the final result based on the majority of predictions from each tree rather than relying on a single decision tree. That is how accuracy increases and the issue of overfitting is avoided when there are more trees in the forest[7].

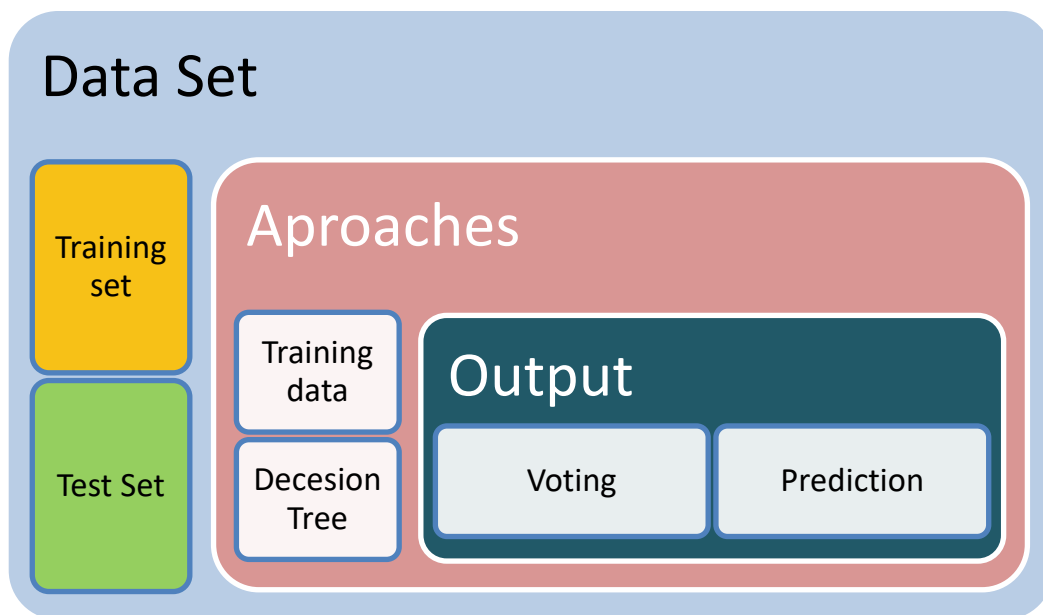


Fig.3.2.4.3: Procedure of Random Forest Algorithm

3.2.5 Feature Generation

The water quality index (WQI) is a type of numerical expression used to assess the condition of a specific body of water and is intended to be easily understood by writers from various nations. Analyzing the heat map, we got the feature classes to predict the targeted class. In that case, SiO_2 and Na_2O are the feature classes, because those two classes have the highest relation with the targeted class. Then after this we need to train our model using 75% data from our dataset and leaving 25% data to test the prediction.

Value ranges (%) Status of water quality 90 to 100 fantastic 70-90 Excellent 50–70 Medium 25–50 Bad 0–25 very poor Horton first proposed the formula for calculating the water quality index in 1965 [12, 13]. This formula reflects the composite influence of various parameters significant for the assessment and management of water quality and takes into account all requirements for determining the quality of surface waters. This indicator was first used to highlight the physical-chemical changes that could take place throughout the year on the quality of flowing water [14, 15]. The water quality index is most frequently used to assess the quality of surface water. This index uses data from several characteristics to rate the caliber of water bodies using a mathematical procedure. Value ranges (%) Status of water quality 90 to 100 fantastic 70-90 Excellent 50–70 Medium 25–50 Bad 0–25 very poor Horton first proposed the formula for calculating the water quality index in 1965 [12, 13]. This formula reflects the composite influence of various parameters significant for the assessment and management of water quality and takes into account all requirements for determining the quality of surface waters. This indicator was first used to highlight the physical-chemical changes that could take place throughout the year on the quality of flowing water [14, 15]. The water quality index is most frequently used to assess the quality of surface water. This index uses data from several characteristics to rate the caliber of water bodies using a mathematical procedure. Value ranges (%) Status of water quality 90 to 100 fantastic 70-90 Excellent 50–70 Medium 25–50 Bad 0–25 very poor Horton first proposed the formula for calculating the water quality index in 1965 [12, 13]. This formula reflects the composite influence of various parameters significant for the assessment and management of water quality and takes into account all requirements for determining the quality of surface waters. This indicator was first used to highlight the physical-chemical changes that could take place throughout the

year on the quality of flowing water [14, 15]. The water quality index is most frequently used to assess the quality of surface water. This index uses data from several characteristics to rate the caliber of water bodies using a mathematical procedure. This index is a mathematical formula that combines data from many criteria to score the quality of water bodies using values from 1 to 100 that may be divided into five classes, each class having a particular quality state and usage domain [13, 16].

3.2.6 Model Creation

In this section we will generate models for each algorithm separately. For Naïve Bayes algorithm we have to import NaïveBayes classifier from sklearn, for K-Nearest Neighbor we need to import K-NearestNeighbor classifier from sklearn and for Random Forest algorithm we need to import RandomForest classifier from sklearn. After creating the model, the trained dataset needs to be fitted in the models individually and then predict each model's accuracy. Fig.4.3.1 describes all the classifiers which we have used to predict the water pollution.

Naive Bayes Algorithm	K-Nearest Neighbor Algorithm	Random Forest Algorithm
<pre>#Creating a model using Naive Bayes algorithm model_NB = GaussianNB() model_NB.fit(x_train,y_train) GaussianNB()</pre>	<pre>from sklearn.neighbors import KNeighborsClassifier #Creating a model using KNN algorithm model_KNN = KNeighborsClassifier() model_KNN.fit(x_train, y_train) KNeighborsClassifier()</pre>	<pre>from sklearn.ensemble import RandomForestClassifier #Creating a model using Random Forest algorithm model_RF = RandomForestClassifier(n_estimators= 10, criterion="entropy") model_RF.fit(x_train, y_train) RandomForestClassifier(criterion='entropy', n_estimators=10)</pre>

Fig.3.2.6.1: Model Creation for different Algorithms

CHAPTER 4

MODEL CREATION AND EVALUATION

4.1 Introduction

After getting the featured classes to predict the targeted class we build up two different models. Then using our dataset, we have trained these models and then using the test data we predicted the accuracy of both models. For a better solution and greater accuracy in this prediction research, we gathered a lot of data from various water samples that were contaminated with different types of glass particles. We then used those data in various models using machine learning techniques to identify the best model and algorithm that gave us a better solution and greater accuracy than the others. It is possible to assert that having a larger dataset and a superior model that yields superior results with high accuracy is crucial to improving outcomes.

4.2 Result and Analysis

The water quality index has been increased over a long period of time (2004-2014) and applied to approximately 10 sampling sections in order to evaluate the water quality of the Buriganga and other rivers in our nation. This index takes into account the maximum annual, the minimum annual, and the mean annual values of the seven physical, chemical, and biological parameters listed below: DO (oxygen saturation in percent), pH (in pH units), BOD5 (biochemical oxygen demand in mg O₂/L), and others, (IUPAC).

```

Classification Report:
              precision    recall  f1-score   support

     0           0.95       0.95      0.95         19
     1           0.83       0.83      0.83          6

   accuracy          0.92         25
  macro avg          0.89         25
 weighted avg          0.92         25

Confusion Matrix:
[[18  1]
 [ 1  5]]
Accuracy Score: 0.92
Precision: 0.8333333333333334
Recall: 0.8333333333333334

```

Fig.4.2.1: Accuracy of Naïve Bayes Algorithm

```

Classification Report:
              precision    recall  f1-score   support

     0           1.00       0.91      0.95         89
     1           0.82       1.00      0.90         36

   accuracy          0.94        125
  macro avg          0.91        125
 weighted avg          0.95        125

Confusion Matrix:
[[81  8]
 [ 0 36]]
Accuracy Score: 0.936
Precision: 0.8181818181818182
Recall: 1.0

```

Fig.4.2.2: Accuracy of KNN Algorithm

```

Classification Report:
              precision    recall  f1-score   support

     0           1.00      0.96      0.98         89
     1           0.90      1.00      0.95         36

 accuracy              0.97         125
 macro avg           0.95      0.98      0.96         125
 weighted avg       0.97      0.97      0.97         125

Confusion Matrix:
[[85  4]
 [ 0 36]]
Accuracy Score: 0.968
Precision: 0.9
Recall: 1.0

```

Fig.4.2.3: Accuracy of Random Forest Algorithm

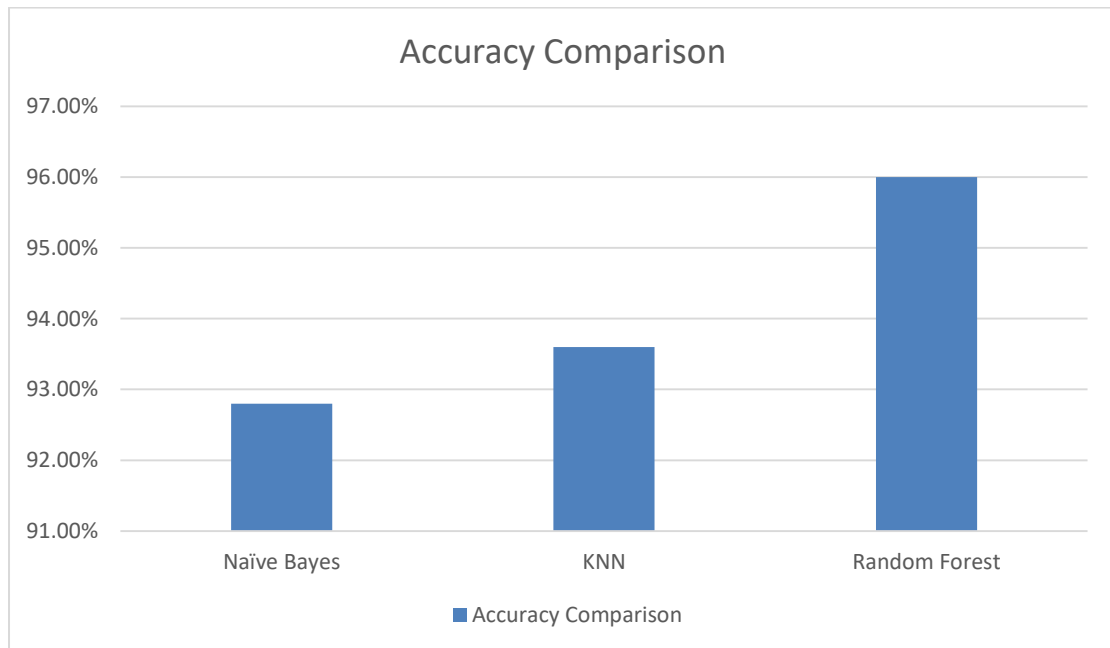
4.3 Comparison between Algorithms

Here, Naïve Bayes algorithm performs with 92.8%, K-Nearest Neighbor algorithm shows the accuracy of 93.6%, and Random Forest algorithm stands out with the accuracy of 96%. So, we can see that among all the algorithms, Random Forest algorithm performs better than all other algorithms. Considering the accuracy of all the algorithm we assure and easily say that, Random Forest algorithm can predict water pollution with highest accuracy of 96%. So, Random Forest algorithm is more efficient than the other algorithms in that research.

Algorithm Name	Accuracy
Naive Bayes	92.8%
KNN	93.6%
Random Forest	96%

Table 4.3.1: Comparison of Accuracy between the Algorithms

Considering for the accuracy of all algorithm we easily come to the decision that, Random Forest algorithm can predict water pollution with highest accuracy of 96%. So, Random Forest algorithm is more efficient than the other algorithms in that research.



Graph.4.3.1: Comparison of Accuracy between the Algorithms

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

According to the United Nations (2016), diarrheal diseases cause more than two million fatalities worldwide each year, with poor sanitation and contaminated water being the main causes of about 90% of deaths and the group that is most affected by them being children. Poor water quality is associated to 80% of diseases and 50% of child mortality worldwide, and it is directly or indirectly to blame for more than 50 different diseases. But illnesses linked to water pollution, such as those that cause diarrhea, skin disorders, malnutrition, and even cancer, are all too common. Investigating how water pollution affects human health, the diversity of diseases, and the importance of drinking clean water became crucial as a result, with important theoretical and practical ramifications for reaching goals for sustainable development. Unfortunately, there are still few research findings that comprehensively investigate the effects of water pollution on human health and the wide range of diseases, despite the fact that much literature and research concentrate on a particular disease and water pollution. In light of the foregoing context and discussion, this study focuses on the effects of water pollution on human health and the heterogeneity of its diseases.

5.2 Impact on Environment

In order for ecosystems to be healthy and productive, a complex network of organisms such as bacteria, fungus, plants, and animals must all interact with one another, either directly or indirectly. The damage caused by even one of these species has the potential to set off a chain reaction that puts entire aquatic environments in danger. Water pollution causes an algal bloom in a large lake or sea because the proliferation of recently added nutrients boosts plant and algae development and lowers the amount of water oxygen levels. Lack of oxygen, or eutrophication, can cause the demise of plants and animals as well as the development of "dead zones," or waterways that are virtually lifeless. These toxic algal blooms could also produce neurotoxins that are detrimental to marine life, including whales and sea turtles.

Chemicals and heavy metals from industrial and municipal waste water also contaminate waterways. These toxins, which reduce an organism's lifetime and reproductive capacity by moving up the food chain when prey is consumed by predators, are unquestionably damaging to all living things. However, giant fish like tuna and others do so by accumulating a lot of poisons like mercury. . Marine debris also suggests a significant threat to marine ecosystems and the sea life, with the potential to smother, strangle, and starve the species. Most of this solid debris, including drink cans and plastic bags, is carried away by storm drains and sewers before being dumped into the ocean. In certain cases, it clings together to form floating rubbish patches or particles, turning the waters into a trashy soup. Fishing gear and other sorts of garbage have destroyed more than 200 different species of marine life. Despite absorbing around 25% of the carbon pollution created annually by burning fossil fuels, oceans are getting increasingly acidic and deadly. The neural systems of sharks, clownfish, and other marine animals may be impacted by this process, which makes it more difficult for shellfish and other species to form shells.

5.3 Ethical Aspects

Pollution of the land, air, and water all have various consequences on living beings. However, one of the most major types of pollution is water contamination, which is bad for human health. Water contamination is one of the main reasons for the loss in freshwater supplies. Pollution of the water causes a huge deal of suffering to living creatures. There are countless factors that contribute to water pollution, some of which are marine dumping, sewage spills, and oil spills. Water pollution is caused by many different things, such as oil, indestructible objects, garbage in lakes and rivers, and flushing. The effects of water pollution on society include poor sanitation, water-borne infections, and a lack of beneficial animals for the eco-system. There is a moral conundrum with water pollution because ship dumping cannot entirely be avoided and may contaminate the land and air. Therefore, every country should make an effort to decrease water pollution. Water pollution is a major challenge in the 21st century.

5.4 Sustainability Plan

An important issue today is water contamination. In a world when freshwater demand is increasing and water resources are scarce, this water pollution increases the already-stressed pressure on water supplies. The term "water pollution" refers to a variety of energy sources or substances that have adverse human effects, including degradation of water quality relative to its use in economic activities like agriculture and industry, loss of amenities, harm to living resources such as plants or animals, risks to human health, and restrictions on insufficient activities. Agricultural components containing chemicals like pesticides and herbicides, the dumping of waste materials into rivers and the ocean, and heavy metals from oil and gas development are just a few of the many sources of water pollution. We are aware that a variety of hazards to natural systems and human health involve water pollution. In dirty water, disease vectors are more likely to exist. water that has been contaminated by industrial and agricultural processes and contains high amounts of several harmful metals (including arsenic, cadmium, and mercury) and synthetic organic compounds (like PCBs). According to estimates, unclean water is to blame for 80% of all infectious diseases in the globe. 3.4 million people each year, mostly children, pass away from drinking water pollution. These pollutants have the potential to poison individuals, contaminate aquifers, and accumulate in groundwater. Nutrient-rich water can lead to eutrophication of the soil and water as well as toxic algal blooms that threaten aquatic biodiversity. Our water supplies are also further strained by water-based contaminants including medications and personal care items. As the world's population rises and the number of causes of water pollution increases, the quality of the water is increasingly likely to deteriorate, causing hazards to human health and the environment as well as social and economic problems. Water quality challenges, which present new risks to water security and sustainable development, are a major challenge for both developed and developing countries, according to the UNESCO.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND FUTURE WORK

6.1 Summary

Bangladesh's water is polluted by many sorts of glass pollution. Most glass-producing businesses discharge their trash into rivers. Glass fragments interact with water as a result, contaminating it. When water is contaminated, substances are released into a body of water where they dissolve, are suspended in the water, or are deposited on the bottom and build up to the point where they interfere with the aquatic ecosystem's ability to function. Substances taken from the air, silt from soil erosion, chemical fertilizers and pesticides, runoff from septic tanks, effluent from livestock feedlots, chemical wastes (some toxic) from industries, plastics, and sewage and other urban wastes from cities and towns are some of the contributors to water pollution. a locality that is far from the source. In a watershed, a community upstream may have access to similarly clean water, whereas one downstream may get a partially diluted mixture of urban, industrial, and agricultural wastes. Additionally, marine systems are impacted by water pollution because they can pick up pollutants from contaminated rivers, streams, or point sources like large ships or oil spills. The overabundance of nutrients in such a phase promotes algal water blooms when the amount of organic objects in the water exceeds the capacity of the microorganisms to break it down and recycle it. When these algae die, their remains combine with the organic waste already present in the water, causing the water to eventually run low on oxygen. When oxygen-free organisms attack organic wastes, they release gases like methane and hydrogen which are harmful to the oxygen-requiring forms of life. The end consequence is a body of water that is stench-filled with rubbish. The health of people is particularly threatened by glass particles. Water samples were taken from a water source close to some glass manufacturing facilities and analyzed. A dataset of various glass particles is created using the water test findings. The accuracy of several models in predicting the water pollution brought on by various glass particles is then assessed. These models were developed using machine learning approaches such as Naive Bayes, KNN, and Random Forest.

6.2 Conclusion

All of the models have an accuracy rate of more than 90% in predicting water contamination. Therefore, it is simple to conclude that any model we created here is more than capable of producing excellent results. These models can aid us in preventing water pollution brought on by various glass particles. If our model can predict, we can also reduce water pollution. All of these initiatives merely serve to increase public awareness of the water damage that these various glass particles cause. The main goal of this study is to educate people about the harmful effects of these glass particles, and our models can accurately forecast whether or not there is water contamination with a 96% accuracy rate. Therefore, we may conclude that our model is significantly more successful to move on to the next stage.

6.3 Recommendation

This section discusses the study's recommendations. The major goal of this study is to forecast the water pollution that glass particles would create, and to make suggestions for how the results of the study might be applied to further this field of study. It also makes suggestions for how to handle and resolve a variety of problems for better results. Additionally, if the problems can be resolved and a better glass pollution prediction model is created, it can be applied to various regions of the nation to determine the overall water pollution brought on by dangerous glass particles nationwide. Eventually, drinking this contaminated water will prevent people from contracting many diseases.

6.4 Future Work

Because the water contamination brought on by the glass particles mixed in can be predicted by our model. However, it takes time to test the water and then determine which glass particles are mixed with how much water. To get around this and improve the model's dependability, we want to create a real-time application that can foretell water pollution right away. This procedure will then be prepared for more applications.

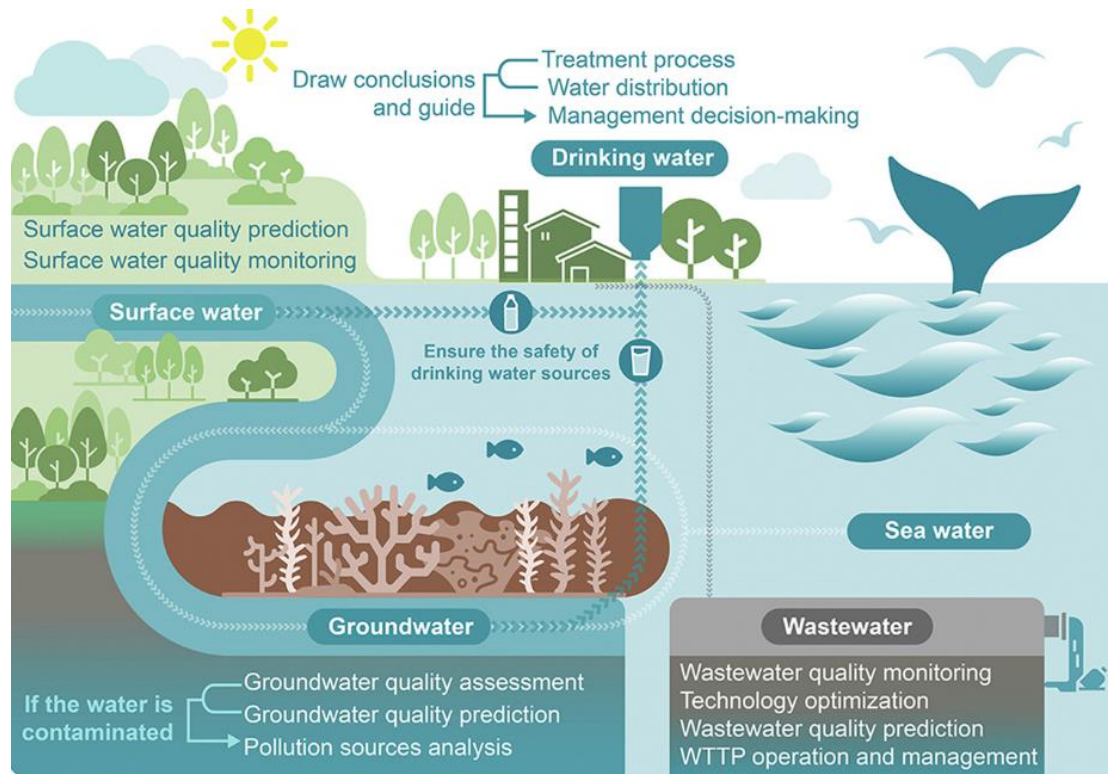


Fig.6.4.1:Future Work Diagram

REFERENCES

- [1] S. A. S. K. Sonu Kumari, "Micro/nano glass pollution as an emerging pollutant in near future," *Journal of Hazardous Materials Advances*, p. 6, 2022.
- [2] P. S. Deepak Pant, "Pollution due to hazardous glass waste," *Environmental Science and Pollution Research*, vol. 21, no. 4, p. 23, 2013.
- [3] M. H. D. Savita Mahurle, "A Study of KNN Classifier to Predict Water Pollution Index," *Advances in Intelligent Systems and Computing*, 2019.
- [4] A. W. T. A. Hemalatha Nambisan, Prediction of Plastic Degrading Microbes, Karnataka, India: St Aloysius College of Management and Information Technology, 2021.
- [5] Javatpoint, "Naïve Bayes Classifier Algorithm," Javatpoint, [Online]. Available: <https://www.javatpoint.com/machine-learning-naive-bayes-classifier>.
- [6] Javatpoint, "K-Nearest Neighbor(KNN) Algorithm for Machine Learning," Javatpoint, [Online]. Available: <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>.
- [7] Javatpoint, "Random Forest Algorithm," Javatpoint, [Online]. Available: <https://www.javatpoint.com/machine-learning-random-forest-algorithm>.
- [8] C. Ma, H. H. Zhang, and X. Wang, "Machine learning for big data analytics in plants," *Trends in Plant Science*, vol. 19, no. 12, pp. 798–808, 2014.

Match Overview



29%



1	dspace.daffodilvarsity... Internet Source	6%	>
2	www.hindawi.com Internet Source	3%	>
3	Submitted to Daffodil I... Student Paper	2%	>
4	Submitted to Jacksonv... Student Paper	2%	>
5	Submitted to National I... Student Paper	1%	>
6	www.britannica.com Internet Source	1%	>
7	www.researchgate.net Internet Source	1%	>