# Table of contents

|  |  |
| --- | --- |
| Content | Page |
| Table of tables | 1 |
| Table of figures | 1 |
| Problem statement | 2 |
| Literature review | 2 |
| Introduction to Neural Network | x |
| Introduction to XGBoost |  |
| Introduction to Decision Trees |  |
| Introduction to Random Forests |  |
| Introduction to Naïve Bayes |  |
| Methodology |  |
| Data Preparation Phase |  |
| Data Collection |  |
| Data Exploration |  |
| Data Visualization |  |
| Data Preprocessing Phase |  |
| Data Cleansing |  |
| Analysis Phase |  |
| Neural Network |  |
| XGBoost |  |
| Decision Trees |  |
| Decision Trees and Random Forests |  |
| Naïve Bayes |  |
| Results and Discussions |  |  |  |
| Neural Network |  |  |  |
| XGBoost |  |  |  |
| Decision Trees |  |  |  |
| Decision Trees and Random Forests |  |  |  |
| Naïve Bayes |  |  |  |
| Model Selection |  |  |  |
| Model Prediction |  |  |  |
| Summary |  |  |  |
| References |  |  |  |
|  |  |

# Table of tables

|  |  |
| --- | --- |
| Content | Page |
| Table 1 :xxx |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Table of figures

|  |  |
| --- | --- |
| Content | Page |
| Figure 1 xxx |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Problem statement

Most disasters are water-related. Floods, landslides, storms, heat waves, wildfires, extreme cold, droughts and waterborne disease outbreaks are all becoming more frequent and more intense, mainly due to climate change. The impacts of disasters include loss of life and damage to water and sanitation infrastructure, such as waterpoints, wells, toilets and wastewater treatment facilities[unwater]. Consequently, water quality monitoring, analysis, and prediction have emerged as important challenges in several uses of water in our life [Torky]. In this research considers the importance of the safety water by making prediction using machine learning with selected algorithms for saving lives and livelihoods. Machine learning is one of the most important and famous decision support tools nowadays. Recent progress in machine learning has been driven both by the development of new learning algorithms and theory and by the ongoing explosion in the availability of online data and low-cost computation. The adoption of data-intensive machine-learning methods can be found throughout science, technology and commerce, leading to more evidence-based decision-making across many walks of life, including health care, manufacturing, education, financial modeling, policing, and marketing[science]. All of these supports in machine learning are crucial for forecasting and motivating in the future applications. This research aims to develop the machine learning models for prediction water safe with given dataset and obtain the most accurate model. The prediction models conclude neural networks (NN), XGBoost, Decision Tree, Decision Trees and Random Forests and Naïve Bayes for binary classification problem.

# Literature review

The literature review concludes the effectiveness of the article, paper and journal with commitment to use artificial intelligence, machine learning and deep learning trends for solving water safe issues.

## Introduction to Neural Networks

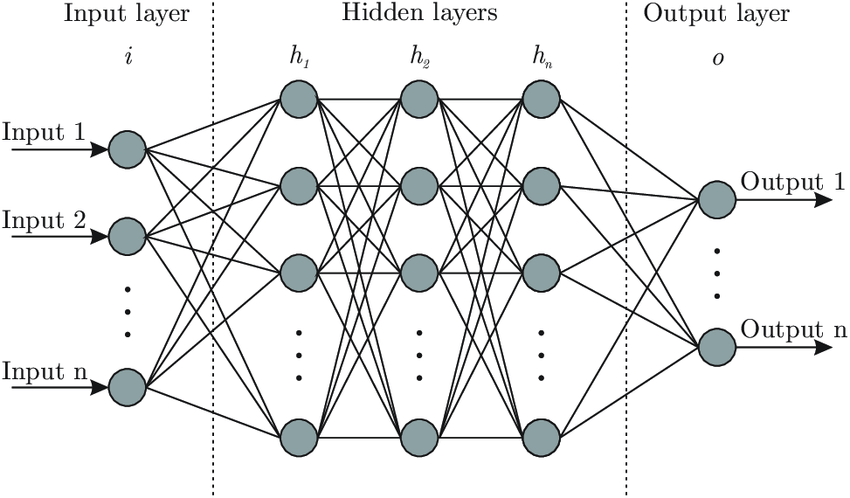


Figure \* Neural Network architecture

<https://medium.com/data-science/designing-your-neural-networks-a5e4617027ed>

A neural network is a machine learning program, or model, that makes decisions in a manner similar to the human brain, by using processes that mimic the way biological neurons work together to identify phenomena, weigh options and arrive at conclusions.

Every neural network consists of layers of nodes, or artificial neurons an input layer, one or more hidden layers, and an output layer. Each node connects to others, and has its own associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

Neural networks rely on training data to learn and improve their accuracy over time. Once they are fine-tuned for accuracy, they are powerful tools in computer science and artificial intelligence, allowing us to classify and cluster data at a high velocity. Tasks in speech recognition or image recognition can take minutes versus hours when compared to the manual identification by human experts. One of the best-known examples of a neural network is Google’s search algorithm.

Neural networks are sometimes called artificial neural networks (ANNs) or simulated neural networks (SNNs). They are a subset of machine learning, and at the heart of deep learning models. [IBM]

(NN papers)

## Introduction to XGBoost

A diagram of a diagram of a product

AI-generated content may be incorrect.

Figure \* XGBoost architecture

https://www.researchgate.net/publication/377778210\_Establishing\_a\_soil\_carbon\_flux\_monitoring\_system\_based\_on\_support\_vector\_machine\_and\_XGBoost/figures

Decision tree (include paper)

Table of compare accuracy between this 3 algorithms or others

(put the paper of relevance)

By researching of Torky, they applied machine learning to classify drinking water samples (safe/unsafe) and predicting water quality index. The experimental results show that the model using Random Forest (RF), and Light Gradient Boosting Machine (Light GBM) models in recognizing safe drinking water samples with superior of the accuracy of 94.7% [Torky]. Besides to the article of Niyongabo, The water quality classification produced by the Random Forest forecast had the highest accuracy of 99.89% in classifying water quality forecasts and how reliable gated recurrent units were in predicting water quality indices and water demand [Niyongabo].

In this research, I applied these three algorithms for evaluating the models for predicting the problems concerning water safety.

## Introduction to Decision Trees

Vvvvv

## Introduction to Random Forests

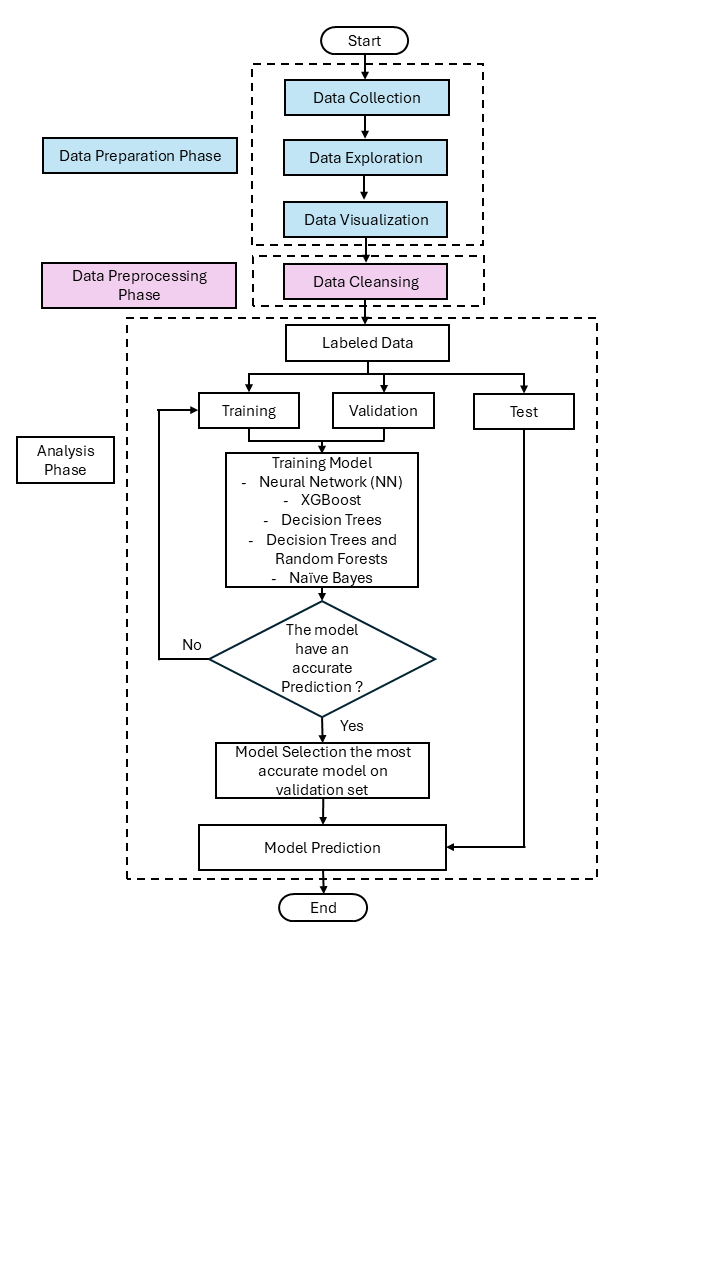
Vvv

## Introduction to Naïve Bayes

vvvvv

# Methodology

The methodology of this work is divided into 3 phases and sub-methods that conclude Data Preparation Phase, Data Preprocessing Phase and Analysis Phase followed by this flow chart.



**Figure \*** Flow chart of this research

## Data Preparation Phase

As far as I write down on the flow chart. The Data Preparation Phase consists of 3 sub-methods that are Data Collection, Data Exploration and Data Visualization.

### *Data Collection*

Data collection and acquisition belong to given assignment data.

### *Data Exploration*

This step can help to find initial patterns, characteristics and interesting points, especially to the data analytics roles. By the way picking the appropriate analysis method will be more advanced when using parallel to data visualization steps. The basic information of this data by reading xlsx file has 4,796 numbers with 21 columns. The 20 colums are aluminium, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, bacteria, viruses, lead, nitrates, nitrites, mercury, perchlorate, radium, selenium, silver and uranium represented to be x value to predict the 1 colum of is\_safe that is represented to be y value in the binary classification problem. The characteristic of x variable is numerical data contrast to y variable that is categorical data. The total dataset contains 7,996 examples and is randomly divided into three parts: training data (4,796 samples – 60%), cross-validation data (1,600 samples – 20%), and test data (1,600 samples – 20%). Lastly, the dataset is well labelled so in the analytics phase I used supervised learning to train the model rather than use unsupervised learning.

A screenshot of a computer

AI-generated content may be incorrect.

**Figure \*** Data Exploration summary from the data frame in ipynb file

### *Data Visualization*

The visualization of the data is necessary tools for diving into the insights of data. About this work, I used popular libraries such as matplotlib and seaborn for making visualization tools to detect the variability of the data, analyze, and plot the graph. To measure the data variability by making boxplot and histogram for showing patterns from Figure \*\*\* to Figure \*\*\* consequently. The correlation analysis will be used to identifying how strong relationships between x and y and so on with x and x.

A group of graphs showing different colored squares

AI-generated content may be incorrect.

**Figure \*** The box plot of aluminum, ammonia, arsenic and barium

Figure \* show that the box plot of *aluminum* unsafe(0) has distributed close to 0, IQR most values are concentrated near the lower end, Numerous high outliers suggesting a long right tail. Highly right skewed with most values near 0 but a few extremely high. And safe (1) IQR Wide the values are more spread out, Outliers: Few to none most values fall within the whiskers, Roughly symmetric or slightly right skewed.

that the box plot of *ammonia* unsafe (0)

A group of boxes with different colored squares

AI-generated content may be incorrect.

**Figure \*** The box plot of cadmium, chloramine, chromium and copper

A group of boxes with different colored squares

AI-generated content may be incorrect.

**Figure \*** The box plot of fluoride, bacteria, viruses and lead

A group of blue and purple boxes

AI-generated content may be incorrect.

**Figure \*** The box plot of nitrates, nitrites, mercury and perchlorate

A group of graphs with different colored squares

AI-generated content may be incorrect.

**Figure \*** The box plot of radium, selenium, silver and uranium

A graph with lines and dots

AI-generated content may be incorrect.

Figure \* The boxplot of overall dataset

A screenshot of a graph

AI-generated content may be incorrect.

**Figure \*** Histogram plot of the dataset

A chart with numbers and symbols

AI-generated content may be incorrect.

**Figure \*** The correlation analysis

From the figure \*, graph plot shows that x variables have a negligible to moderate relationship with y variable from the strongest [aluminum – is\_safe] about 0.35 (35%) to the lowest [lead – is\_safe] about -0.007 (0.7%) consequently. The scoring criteria defines in table \*.

**Table \*** Interpretation of correlation strength

|  |  |
| --- | --- |
| **Correlation Coefficient** | **Strength** |
| 0.00 – 0.10 | Negligible |
| 0.11 – 0.30 | Weak |
| 0.31 – 0.50 | Moderate |
| 0.50 – 1.00 | Strong |

**Table \*** Identify relationship of x and y

|  |  |  |  |
| --- | --- | --- | --- |
| **Pair** | **Correlation Coefficient** | **Strong/Moderate /Weak** | **Positive/Negative** |
| aluminum – is\_safe | 0.35 | Moderate | Positive |
| ammonia – is\_safe | -0.018 | Negligible | Negative |
| arsenic – is\_safe | -0.12 | Weak | Negative |
| barium – is\_safe | 0.1 | Negligible | Positive |
| cadmium – is\_safe | -0.26 | Weak - Moderate | Negative |
| chloramine – is\_safe | 0.21 | Weak | Positive |
| chromium – is\_safe | 0.19 | Weak | Positive |
| copper – is\_safe | 0.049 | Negligible | Positive |
| fluoride – is\_safe | 0.017 | Negligible | Positive |
| bacteria – is\_safe | -0.023 | Negligible | Negative |
| viruses – is\_safe | -0.096 | Negligible | Negative |
| lead – is\_safe | -0.007 | Negligible | Negative |
| nitrates – is\_safe | -0.067 | Negligible | Negative |
| nitrites – is\_safe | 0.05 | Negligible | Positive |
| mercury – is\_safe | -0.03 | Negligible | Negative |
| perchlorate – is\_safe | 0.083 | Negligible | Positive |
| radium – is\_safe | 0.08 | Negligible | Positive |
| selenium – is\_safe | -0.045 | Negligible | Negative |
| silver – is\_safe | 0.11 | Weak | Positive |
| uranium – is\_safe | -0.072 | Negligible | Negative |

The x variables have many columns for complying with machine learning models so that can be used many computation times too. However, if I use the cut off (pruning) to reduce the dimension by dropping the low-correlation features (Negligible) might help with simple regression but not work for criteria for feature importance. In evaluated models with more complex such as neural networks, XGBoost, Random forests and etc. may capture the necessary or non-linear relationships with target and also with multicollinearity which can cause redundancy in linear models. In those cases, cutting off based on linear correlation could remove valuable. The future applies aim to improve model interpretability is cut off features with weak correlation to reduce overfitting and use more advanced feature selection method or model-based importance metrics.

## Data Preprocessing Phase

As far as I write on the flow chart. The Data Preprocessing Phase consists of 1 sub-method that is Data Cleansing.

### *Data Cleansing*

This method is to check the missing data and operate to cope with it by coding.

A list of different types of bacteria

AI-generated content may be incorrect.

Figure \* The result of checking the missing data

The results show that in this data there are no missing data occurring. So, in this research do not need to be handling.

## Analysis Phase

The data that used in this research is water safety that splitting into 3 sets already with ratio are Training set 60%, Validation set 20% and Testing 10% before applying to evaluate the machine learning models. In this part I will contain the training method of each model with potential setting up parameters to get high accuracy of validation set, feature selection to concentrate on fixing the overfitting or underfitting. The results of the models show in the results and discussions.

### *Neural Network*

The parameters of a learning algorithm to set up for training method.

Input layer = 4796

* Nodes in input layer

Hidden layer (n\_h) = 100

* Nodes in Hidden layer

Output layer = 4796

* Nodes in Output layer

Alpha (Learning rate) = 0.01

Number of iterations = 500

* Activation Functions is a sigmoid function, used to apply non-linear transformation on input to map it to output.

Error Computation

* Loss = Actual\_Value - Predicted\_Value
* Cost = Summation (Loss)

Log loss function

* -Sum ( Log (Pred) Actual + Log (1 - Pred ) Actual ) / m

### *XGBoost*

The parameters of a learning algorithm

Number of iterations = 500

Alpha (Learning rate) = 0.01

Depth of each tree = 4

Fraction of samples to use for each tree = 0.8

Fraction of features to use for each tree = 0.8

### *Decision Trees*

criterion{“gini”, “entropy”, “log\_loss”}, default=”gini”

The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “log\_loss” and “entropy” both for the Shannon information gain.

max\_depthint, default=None

The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min\_samples\_split samples.

### *Decision Trees and Random Forests*

Vvvvvvvv

### *Naïve Bayes*

vvvv

# Results and discussion

In these sections, I’ll explore the results after evaluating models with highlighted algorithms. The results have been processed by accuracy from the validation set and discuss the performance of the model by using the learning curve.

## Neural Network

After training the model of water safety, I will go to the next step of validation. The validation process belongs to the validation dataset.



**Figure \*** The accuracy of the neural network model

A blue and white chart with numbers and a blue square

AI-generated content may be incorrect.

**Figure \*** Confusion matrix of the neural network model

From the confusion matrix. We focused on actual versus prediction of True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) as shown in the figure \*

The accuracy of the model = 88.4375%

The interpretation of the confusion matrix is that:

True Positives (TP) = 0 The model did not correctly predict any positive instances.

False Positives (FP) = 185 The model incorrectly predicted 185 instances as positive when they were actually negative.

True Negatives (TN) = 1415 The model correctly predicted 1415 instances as negative.

False Negatives (FN) = 0 The model did not miss any positive instances (because it never predicted any positives).

The model is heavily biased toward predicting only the negative class (0) or always predicting unsafety water and it’s never predicted the positive class (1) or safety water, which is a sign of extreme class imbalance or a miscalibrated model. Since all actual positives were misclassified as negatives, the model might be overfitting to the negative class.

A graph of a graph

AI-generated content may be incorrect.

Figure \* Learning curve on a NN model

Figure \* tell the model performance. The training is shown as a blue line which shows the accuracy slowly drop over time of iterations (number of iterations = 500) because of the training fit while validation as an orange is stable over time of iterations because the model uses a single validation data set or not generalized with more data. The training and validation have a close gap between the graphs that indicate low variance but high bias because the model has hit a performance ceiling. In additional the stable validation graph may have a cause of class imbalance that the dataset has many more negative samples than positive ones so that the model might be learning to always predict 0(unsafety) or feature issue so that the features might not be informative enough to distinguish between positive and negative cases.

## XGBoost

After test on the validation set. The model got the point that contains information of these.



**Figure \*** The accuracy of the XGBoost model

A blue and white diagram

AI-generated content may be incorrect.

**Figure \*** Confusion matrix of a XGBoost model

The accuracy of the model = 96.50%

The interpretation of the confusion matrix is that:

A graph of a number of trees

AI-generated content may be incorrect.

**Figure \*** Learning curve on a XGBoost model

## Decision Trees

From the documentation of Scikitlearn website, the decision Tree has a criterion parameter function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “log\_loss” and “entropy” both for the Shannon information gain. In this model I compared 2 criteria functions to measure the differentiation of accuracy.

**Decision tree with entropy as a criteria**

  
**Figure \*** The accuracy of the Decision Trees model with entropy

A blue and white diagram

AI-generated content may be incorrect.

**Figure\*** Confusion matrix of the Decision Trees model with entropy

**Decision tree with Gini as a criteria**



**Figure \*** The accuracy of the Decision Trees model with Gini

A blue and white diagram

AI-generated content may be incorrect.

**Figure\*** Confusion matrix of the Decision Trees model with Gini

The Gini criteria gave a more accuracy than entropy so in the decision trees model I likely to use Gini index as a based of the model and build the decision tree for the algorithm decision making.

A graph with a line and a line graph

AI-generated content may be incorrect.

Figure \* Learning curve on a Decision Tree model

A diagram of a diagram

AI-generated content may be incorrect.

Figure \* Example of a Decision Trees architecture by using Gini index

## Decision Trees and Random Forests



**Figure \*** The accuracy of the Decision Trees and Random Forests model

A blue and white diagram

AI-generated content may be incorrect.

Figure\* Confusion matrix of a Decision Trees and Random Forests model

A graph with a line graph

AI-generated content may be incorrect.

Figure \* Learning curve on a Decision Trees and Random Forests model

A diagram of a network

AI-generated content may be incorrect.

Figure \* Example of a Decision Trees and Random Forests architecture

## Naïve Bayes



**Figure \*** The accuracy of the Naïve Bayes model

A blue and white diagram

AI-generated content may be incorrect.

Figure\* Confusion matrix of a Naïve Bayes model

A graph of a training course

AI-generated content may be incorrect.

Figure \* Learning curve on a Naïve Bayes model

## Model Selection

Select the most accurate model with the validation set. The result shows that the XGBoost model gave the highest accuracy which is 96.50%. However, The second highest accuracy is Decision Trees model with Gini which is 96.19% and

## Model Prediction

Vvvvvv

## Summary

Vvvvvv

Table \* Summary of the models

|  |  |  |
| --- | --- | --- |
| Models | Validation accuracy | Confusion Matrix |
| Neural Network | 88.4375% |  |
| XGBoost | 96.50% |  |
| Decision Trees with entropy | 95.50% |  |
| Decision Trees with Gini | 95.75% |  |
| Decision Trees and Random Forests | 95.38% |  |
| Naïve Bayes | 84.12% |  |

# References

สำหรับเนื้อหา

<https://www.unwater.org/water-facts/water-and-disasters>

https://www.science.org/doi/abs/10.1126/science.aaa8415

[Mohamed Torky]Recognizing Safe Drinking Water and Predicting Water Quality Index using Machine Learning Framework

[Alain Niyongabo] Predicting Urban Water Consumption and Health Using Artificial Intelligence Techniques in Tanganyika Lake, East Africa

<https://www.ibm.com/think/topics/neural-networks>

<https://www.matillion.com/learn/blog/data-exploration>

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

สำหรับ coding

https://www.kaggle.com/code/shivamb/a-very-comprehensive-tutorial-nn-cnn

https://adityagoel123.medium.com/binary-classification-using-decision-tree-model-c4e20c8a5afb

https://github.com/mGalarnyk/Python\_Tutorials/blob/master/Sklearn/CART/Visualization/DecisionTreesVisualization.ipynb

https://www.kaggle.com/code/wrecked22/basic-binary-classification-using-xgboost

https://towardsdatascience.com/the-naive-bayes-classifier-how-it-works-e229e7970b84/