National College of Ireland

# Predicting River Water Quality Parameters using Supervised Machine Learning Techniques: UK

MSc Research Project
Data Analytics

Stephanie Whelan
Student ID: 19140649

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | |
|---|---|
| **Student Name:** | Stephanie Whelan |
| **Student ID:** | 19140649 |
| **Programme:** | Data Analytics |
| **Year:** | 2022 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Dr. Catherine Mulwa |
| **Submission Due Date:** | 15/08/2022 |
| **Project Title:** | Predicting River Water Quality Parameters using Supervised Machine Learning Techniques: UK |
| **Word Count:** | 10552 |
| **Page Count:** | 28 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | |
| **Date:** | 14th August 2022 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Predicting River Water Quality Parameters using Supervised Machine Learning Techniques: UK

Stephanie Whelan
19140649

**Abstract**

The quality of our water is essential to human health and to our ecosystem. Pollution in water can cause humans to become ill and wildlife to die. Rivers have become one of the most used natural water sources globally, yet in the last decade river pollution has grown due to human activities and climate change increasing the importance of a reliable, fast and affordable way to monitor river water quality. In this study, five supervised machine learning models were applied to a river water quality dataset that were collected from a river and its tributaries located in South East England. They include Decision Trees, Random Forest, Extreme Gradient Boosting, Support Vector Machines and Multiple Linear Regression. Four popular river quality parameters were predicted, they are Dissolved Sodium, Dissolved Nitrate, Gran Alkalinity and Electrical Conductivity. The best performing algorithm was found to be Random Forest when predicting all parameters with an R-Squared value of between 87% and 98%. The results found in this study can help to support the monitoring of river water quality in a fast and inexpensive way and improve the existing testing system in place.

# 1 Introduction

The quality of the water in our rivers, lakes and oceans is essential to our ecosystem, wildlife and human health. Many different factors can lead to water becoming polluted, including environmental, agricultural, and human activities. More specifically, climate change, increased heavy rainfall leading to flooding, agricultural run-off and sewage are just a few of the factors that can lead to water becoming unsafe (Maloo et al., 2018). In the last decade, river pollution has grown due to human activities and climate change (Kurniawan, et al., 2021).

Estimates show that 485 thousand people die from diarrhoea caused by polluted drinking water (Lopez et al., 2021). However, water is not only harmful when consumed, bathing and swimming in contaminated water can also lead to serious illness. For instance, over 120 million cases of gastrointestinal disease and over 50 million cases of respiratory disease cases a year are caused by entering polluted coastal waters (Maloo et al., 2018).

Rivers have become one of the most used natural water sources globally, and this is due to the accessibility and the location of cities being built close to riverbanks (Kurniawan, et al., 2021). Water has also become increasingly popular for recreational activities since the Covid-19 ‘lockdown’ with more and more people entering the coastal waters, rivers and lakes that surround us. This means that more people are being exposed to waters that may be polluted and unsafe for human use.

Although many resources are available that alert people to unsafe waters, including websites such as beaches.ie in Ireland and Surfers Against Sewage in the UK, these methods mainly focus on coastal waters and not our rivers. Traditional methods for testing water for pollution can also be slow, sometimes taking up to 24 hours for results. The quality of water can also dramatically change in a short space of time, meaning water that was proven to be safe can change to being unsafe before the public can be notified which can lead to the public being

at a greater risk of illness (Thoe et al., 2014). It is not only human life that suffers from polluted water, the ecosystem of rivers can also suffer as many essential plants and wildlife require a certain purity level to be able to survive (Khullar and Singh, 2021).

This research looks at river water quality in South East England and how water quality can be predicted in a timely and accurate manner. Currently among the state of art there are limited studies that look at the United Kingdom and the water quality found in their rivers. The research question that this research was based on is stated below:

## 1.1 Research Question and Objectives

***RQ: "How successfully can supervised machine learning techniques be used to predict river water quality parameters to assist in monitoring water quality in a timely manner?"***

To solve the research question, the following objectives were implemented and the results are discussed.

***Obj 1.*** The investigation, critical analysis and review of current literature on the topic of water quality prediction.
***Obj 2.*** Implement and evaluate water quality supervised machine learning algorithms.

> ***Sub Obj 2.1.*** Implemented and evaluated the results of the Decision Trees algorithm.
> ***Sub Obj 2.2.*** Implemented and evaluated the results of the Random Forest algorithm.
> ***Sub Obj 2.3.*** Implemented and evaluated the results of the Extreme Gradient Boosting algorithm.
> ***Sub Obj 2.4.*** Implemented and evaluated the results of the Support Vector Machine algorithm.
> ***Sub Obj 2.5.*** Implemented and evaluated the results of the Multiple linear regression algorithm.

***Obj 3.*** Evaluated additional parameters such as air temperature, sunshine hours and rainfall levels and their effect on the accuracy of the chosen models.
***Obj 4.*** Compare and contrast the implemented supervised machine learning models performance.
***Obj 5.*** Comparison of developed water quality algorithms versus existing algorithms identified in the literature.
***Obj 6.*** Identify the best performing model in this research so that a recommendation can be made and the final results can enhance the current field of research.

**Contributions and Limitations:** This research will contribute to the current state of art by focusing on river water quality prediction in the South East of England, more specifically the River Thames and its tributaries. Five supervised machine learning regression models were used to reliably predict four important river water quality parameters with the addition of weather parameters such as temperature, rainfall and sunshine. The final models were evaluated and compared to other popular models for predicting water quality. The final ICT solution performed very well and could be commercialised by organisations to reliably predict river water quality. It is important to note that although the weather dataset is collected from a weather station in the same region (South East England) as the River Chemistry Data sample points, it is not at exactly the same coordinates which may affect the importance that the weather variables have in the final models.

The remainder of this report is structured as follows. Chapter two will consist of a literature review that will critically analyse peer-reviewed literature in the area of water quality prediction and machine learning. Chapter three will outline the methodology of this research and how the selected machine learning models will be implemented. Chapter 4 will outline the two-tiered design specification. Chapter Five will describe the implementation process and evaluation of the chosen models. Chapter 6 will discuss the results and the last chapter in this research will discuss and conclude the work completed and suggest future work that could be completed to build on this research.

# 2 Related Work on Water Quality (2000 – 2022)

The quality of our water is extremely important to human health and the ecosystem. Pollution entering our coastal water, rivers and lakes can cause humans to become ill, wildlife to die and threaten our ecosystems. Typically, water quality is tested in labs through samples that are taken from the water, this can be expensive and take time. Machine Learning is becoming increasingly more popular for water quality prediction, not only for its cost-effectiveness but also for its speed and accuracy. It also allows for alerts to be put in place at beaches, rivers and lakes to notify the public that the water is polluted. This section will review and critically analyse the research conducted on water quality, its effect on human health and the environment, and the machine learning methods being for prediction.

## 2.1 Water Quality and its Impact on the Environment and Human Health

Polluted water can cause risks to the health of humans and transmit diseases causing serious illness or worse, causing death (Dawood et al., 2021). A problem faced all over the world is the challenge of keeping water supplies clean and free of pollutants. Poor drinking and bathing water quality can lead to humans becoming ill, the outbreak of diseases and can even result in deaths in urban communities (Dawood et al., 2021). Water pollution can be caused by a number of issues including physical, environmental and operational. Issues such as damaged pipelines, chemicals leaching into soils and the by-products of industry entering water supplies can all make water unsafe (Dawood et al.,2021). Water pollution can be defined by its quality, which can be determined by the presence and amount of different features such as turbidity, pH, electrical conductivity, dissolved oxygen (DO), nitrate, temperature, biochemical oxygen demand (BOD) and Sodium (Radhakrishnan & Pillai, 2022). In Britain, it has been found that over 300 million gallons of sewage enters their waters each year (Parker and Frost., 2000). This raw sewage can contain pathogens that are linked to many illnesses such as diarrhoea, fever Mild or influenzal, typhoidal illness and respiratory diseases (Parker and Frost., 2000). Systematic surveillance in the USA has detected multiple outbreaks of waterborne diseases. These outbreaks have been associated with untreated recreational water and recreational swimming pools (Schets et al., 2011). Epidemiological research over the last 50 years has linked disease with the quality of our recreational water all over the world (Parker and Frost., 2000). The most common diseases associated with these outbreaks have been illnesses such as gastroenteritis, skin conditions and neurological diseases for example meningitis and meningoencephalitis. The cause of which is typically from the presence of naegleria fowleri in the water (Schets et al., 2011). Naegleria fowleri is a disease often caused by recreational warm water activities in the summer months such as water skiing, surfing and rafting, which lead to water entering the nasal cavity and can lead to death (Hamaty Jr et al., 2020).

Changing weather conditions can also lead to water becoming polluted and dangerous to our ecosystems. High rainfall and flooding can wash chemicals and byproducts into our oceans, rivers and lakes. The temperature of the water we bathe in can also influence the quality, with warmer water allowing bacteria such as P. aeruginosa and Vibrio spp to thrive (Schets et al., 2011). As the temperature of water rises these bacteria begin to cause problems and can play a huge role in causing waterborne disease outbreaks. This threat of dangerous bacteria being present in our water in high concentrations is only increasing as temperatures continue to rise due to climate change and global warming.

Weather and its impact on bathing water quality in Scotland were looked at by Eze et al., (2014). They modelled monthly counts of viral and non-viral gastrointestinal infections, temperature and the monthly counts of faecal indicator organisms. It was shown that non-viral gastrointestinal infections increased as temperature and humidity increased. Therefore, climate change in the future, including rising temperatures and high levels of rain can result in higher levels of faecal organisms, which will lead to more cases of infectious intestinal disease (Eze et al., 2014).

The quality of our water depends on biological, chemical and physical variables which can originate from both natural sources and humans (Banda et al., 2020). Although these parameters can be vital to water, if found in excessive amounts it can cause a risk to the ecosystem. In order to maintain a quality of water that is safe, a Water Quality Index (WQI) was created to measure safe levels of different water quality parameters (Banda et al., 2020).

## 2.2 Existing Methods for Predicting Water Quality Using Machine Learning Algorithms

Machine learning (ML) is a branch of artificial intelligence and a rapidly growing technical field that lies between computer science and statistics. ML works by mimicking human behaviour and uses data and algorithms to improve its accuracy (Nair and S., 2021). ML is widely used within environmental studies and with the continuous improvement of machine learning methods, even more researchers are using ML for the prediction of water quality (Li et al., 2021).

Water quality prediction and the use of ML has been looked at in a number of research papers throughout the years. Radhakrishnan & Pillai (2021) focus their research on classification models for predicting water quality in the Narmada River in India. They compare Support Vector Machines, Decision Trees and Naive Bayes algorithms using the parameters such as turbidity, pH, DO, nitrate, temperature and BOD. They found that Decision Trees perform best with a low number of incorrectly classified data followed by Support Vector Machines. Naive Bayes performs the worst with the highest number of incorrectly classified data, with the authors stating that it's evident that Naive Bayes is not an appropriate algorithm for water quality prediction (Radhakrishnan & Pillai, 2021). However, they state that their research could be improved by using a larger dataset.

Support Vector Machines were also used by Kurniawan et al (2021) in their research that also looked at the prediction of water quality in rivers, more specifically the Kelantan River in Malaysia. They used six quality parameters to predict including biochemical oxygen demand, dissolved oxygen, chemical oxygen demand, ammonia nitrogen, and suspended solids and found that using suspended solids as their predictor performed the best with an R Squared score of 93.6%. They found that Support Vector machines worked well on their dataset, yet similar to Radhakrishnan & Pillai (2021), they used a small dataset consisting of only 148 records.

Classification is not the only method of prediction for water quality. In their research, Wang et al. (2021) use regressions, neural networks and ensemble methods to predict water quality in estuarine water. Estuaries can be sources of major pollution for coasts making water quality prediction extremely important in this area. The data used in this research consists of 824 records which were split into training and testing datasets. They found that XGBoost performed the best in predicting estuarine ammonia nitrogen, however like previous studies the dataset that was used was small and no external parameters such as climate, or human activities were used which may have improved the model and its accuracy (Wang et al, 2021).

The quality of coastal water was researched by Thoe et al. (2014), Oh et al. (2021) and Su et al. (2021) and various machine learning models were used for prediction. Thoe et al. (2014) focused their research on the beaches of California, specifically Santa Monica, and evaluated five different machine learning models for the prediction of faecal coliform and enterococci concentrations in the summer using multiple linear regression model, binary logistic regression, partial least square regression model, artificial neural network, and classification tree. Unlike previous studies, Thoe et al. (2014) mention the importance of external factors on pollution such as rainfall, wave height and storm drain conditions and include these as parameters in their models. They find that Classification Trees perform the best, followed by artificial neural network and binary logistic regression. Oh et al. (2021) focus their research on the quality of sea water in the Masan Bay and Nakdong River Estuary, two polluted areas of South Korea, and apply four classification algorithms to the data. They use support vector machines, Random Forest, multinomial logistic regression, and artificial neural networks to predict the water quality level (Oh et al., 2021). They found that when Support Vector Machines and Random Forest were applied to the dataset consisting of over 15K records they could provide faster and real-time water quality monitoring, allowing a quicker response to pollution. Their Support Vector Machines and Random Forest achieved an F1-score of over 95%.

In contrast to previous studies, Su et al. (2021), use satellite observations and Gradient Boosting Machines to estimate the concentration of Chlorophyll-a (chl-a), an important parameter of water quality, in Fujian's coastal waters. They found that their model performs well with an R-Squared value of 77%, yet similar to previous studies they do point out that the model could be improved with more external parameters such as rainfall and temperature.

Neural Networks are also a widely used model for the prediction of water quality. Neural Networks acquire knowledge through experience and store that knowledge in the weights of the connections between neurons (Chafloque et al., 2021). Water quality researchers that have used Neural Networks in their studies include Ewusi et al. (2021), Chafloque et al. (2021), Lopez et al. (2021) and Rani et al. (2022). Ewusi et al. (2021) use gaussian process regression, principal component regression and backpropagation neural network models to predict the total dissolved solids in groundwater, drinking water and surface water. They found that the gaussian process regression model gave the best prediction with an average R-Squared of 98%. They do, however, point out that the mean concentrations of the parameters used were lower than guidelines suggest, which makes their research less reliable than others. This should be taken into account when referencing their work. Chafloque et al. (2021) use neural networks to predict the quality of water for human consumption. The neural network architecture they proposed consists of 7 layers of neural networks with 9 input variables and 5 hidden layers. The activation function ReLu is used and as the last layer the sigmoid function is used. The accuracy of the model is found to be 69% which suggests that the model could be improved in future work and is not appropriate for real world use. Lopez et al. (2021) also use neural networks in their study that looks at the prediction of water quality in the Muvattupuzha River in Kerala using long short-term memory neural networks.

Parameters such as tubridy, total dissolved solids and pH were used for predicting and a dataset consisting of 1826 rows of data was used. They compared their final long short-term memory neural networks model to an artificial neural network and found it outperformed it on all parameters. It is important to note that the data set used was small and this may have affected the results of the final model. Finally, Rani et al. (2022) also used an artificial neural network for their research that looked at the prediction of drinking water quality. The dataset used consists of 63 data points which include the parameters temperature, dissolved oxygen, pH, conductivity and biological oxygen demand. The data set is split into training and testing datasets before the model is applied, achieving high accuracy.

## 2.3 Critical Review of Existing Methods, Algorithms and Results

A critical review of existing methods and their results is shown in Table 1 below. While most models performed quite well, we can see that some of the highest accuracy results came from supervised learning regression methods such as Multiple Linear Regression, Random Forest Regression, Extreme Gradient Boosting (XGBoost) and Support Vector Machines. We can also see that although the gaussian process regression model and back-propagation neural network model performed well with an R-Square of 94% and above, the Mean Absolute Error (MSE) and Root Mean Square Error (RMSE) are considerably higher than other methods listed. It can also be seen that Ahmed et al. (2019) had a poor R-squared result with their random forest, support vector machine and polynomial regression models; however, it must be taken into account that their sample of data was low at only 699 rows. The below table and the results that were obtained from similar machine learning models in the field of water quality prediction will help guide the researcher on the appropriate models to choose for this study.

**Table 1: Critical Review of Existing Water Quality Prediction Methods**

| **Algorithm** | **Evaluation** | **Results** | **Author** |
|---|---|---|---|
| Multiple linear regression | RMSE | 0.4 | Thoe et al. |
| Multiple linear regression | R-Squared, RMSE | 91%, 1.15 | Wang et al. |
| Random Forest | R-Squared, RMSE | 94%, 1.14 | Wang et al. |
| XGBoost | R-Squared, RMSE | 96%, 1.03 | Wang et al. |
| ANN | R-Squared, RMSE | 92%, 1.35 | Wang et al. |
| Random Forest | MAE, MSE, RMSE, R-Squared | 2.31, 9.57, 3.09, 67% | Ahmed et al. |
| Support Vector Machine | MAE, MSE, RMSE, R-Squared | 2.44, 10.63, 3.26, 35% | Ahmed et al. |
| Polynomial Regression | MAE, MSE, RMSE, R-Squared | 2.73, 12.73, 3.57, 49% | Ahmed et al. |
| Support Vector Machine (Classification) | Accuracy | 87% | Radhakrishnan & Pillai |
| Naïve Bayes (Classification) | Accuracy | 75% | Radhakrishnan & Pillai |

| | | | |
|---|---|---|---|
| Decision Trees (Classification) | Accuracy | 87% | Radhakrishnan & Pillai |
| Satellite Observations and Gradient Boosting Machines | R-Squared | 77% | Su et al. |
| Gaussian process regression model | R-Squared, MAE, RMSE | 98%, 7.41, 14.73 | Ewusi et al. |
| Back-propagation Neural Network model | R-Squared, MAE, RMSE | 94%, 9.95, 19.49 | Ewusi et al. |
| Neural Network | Accuracy | 69% | Chafloque et al. |
| Long Short-Term Memory Neural Networks | RMSE | 0.08 | Lopez et al. |
| Artificial Neural Network | MSE, RMSE | 1.12, 1.09 | Rani et al. |

## 2.4 Conclusion and Identified Gaps

It is clear from the research reviewed that there are many types of ML models that can be used for the prediction of water quality including classification, regression, neural networks and ensemble methods. Supervised machine learning methods such as Multiple Linear Regression, Random Forest Regression, Extreme Gradient Boosting and Support Vector Machines are shown to have a high accuracy when used for prediction within the field of water quality while Neural Networks perform inconsistently across many studies with some results being below average. It is also clear that while many studies achieve high evaluation scores, the datasets used are limited in size which may affect results. Research has also shown that most water quality studies focus on areas outside of Europe suggesting there is scope to focus this research on an area located in Europe, specifically the UK. Another gap that is identified from reviewing the literature is the lack of weather variables used for prediction such as rainfall, temperature and sunshine. As is evident from the literature reviewed, weather variables play an important role in determining the quality of water so may have a benefit in the accuracy of a water quality prediction model. The investigation, critical analysis and review of current literature on the topic of water quality prediction have now been completed and ***Obj 1*** has been achieved. Chapter three of this research will explain the methodology and design specification that was carried out for this research.

# 3 Water Quality Prediction Methodology and Design Specification

The following chapter will outline the methodological approach, design specification and data pre-processing methods for the prediction of river water quality in the UK.

## 3.1 River Water Quality Prediction Methodology

The Water Quality Prediction methodology was adapted from the Cross Industry Standard Process for Data Mining (CRISP-DM). This methodology approach was chosen because it is a cross industry standard which means that it is suitable for all data science research regardless of the domain. CRISP-DM has six stages which includes Business understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. Figure 1 shows the adapted water quality prediction methodology approach for this research.

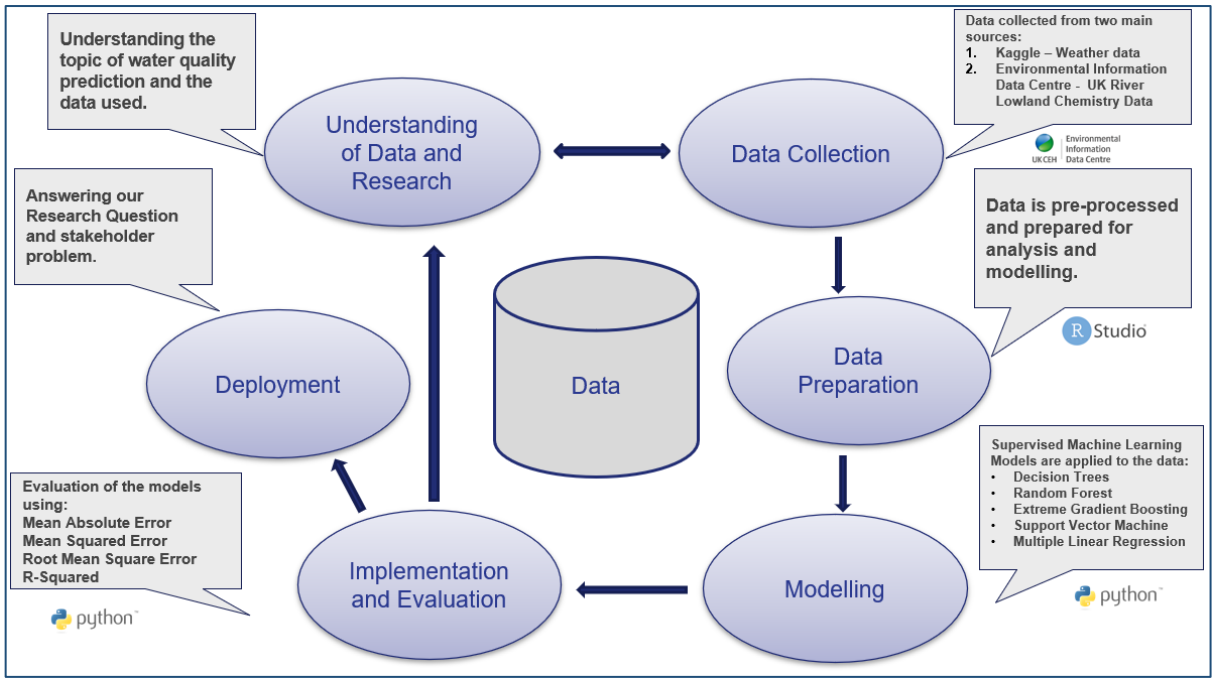


Figure 1: River Water Quality Prediction Methodology

### 3.1.1 Understanding of Research and Data

This step was completed in the previous literature review section of this research. Relevant literature on the importance of water quality and the prediction of water quality with ML was analysed and gaps in the research were identified and discussed. It was evident from the literature that water quality is extremely important to human health and the wider ecosystem. Current methods of water quality prediction involve the collection of water samples and analysis in a lab which can be both expensive and slow. This may lead to results being shared with the public late and humans being unknowingly exposed to polluted and harmful waters or wildlife and plant dying from polluted waters.

### 3.1.2 Data Used

The UK was chosen as the main area for this research. The datasets used for this research includes UK Lowland River Chemistry Data collected from the Environmental Information Data Centre. This dataset will be referred to as the water quality dataset for the remainder of this report. The water quality dataset consists of 9,523 rows of data and 90 variables, 83 of these being water quality parameters. The dataset also shows the date and time the sample was collected. The main collection point of samples was from the Thames catchment area in South

East England and samples were collected between 1993 and 2009.The second dataset that was used for this research was collected from Kaggle. It consists of daily precipitation measurements, air temperature measures and sunshine measurements between 1979 and 2020. The measurements were collected from the Heathrow weather station, which is also located in South East England.

### 3.1.3 Data Preparation

The data was loaded into R-studio and analysed, NA values were removed and water quality dataset and weather dataset were merged. Redundant variables were removed and the data sets were cleaned and transformed. Visualisations were created and correlations were looked at. Finally, the data was loaded into Python to begin implementing the chosen algorithms and evaluate the results. Data preparation is discussed in more detail in Section 4 of this research.

### 3.1.4 Data Modelling

Five machine learning models were chosen for this research, each model was chosen because of its suitability to the problem and because of its wide use and accuracy in water quality research. They are Decision Trees, Random Forest, Extreme Gradient Boosting, Support Vector Machine and Multiple Linear Regression.

### 3.1.5 Evaluation

As the models used for this research are regression models, the evaluation metrics selected for this research are Mean Squared Error, Root Mean Squared Error, Mean Absolute Error and R-Squared. These evaluation metrics were chosen because of their suitability for regression algorithms and wide use within the wider research area and the literature reviewed in Section 2 of this research paper. Each evaluation method is discussed in more detail below:

- Mean Squared Error (MSE) - The MSE calculates how close a regression line is to the data points. It is calculated by subtracting the predicted value from the observed value and squaring that difference. A small MSE is preferred as it shows that the error in the model is small (Gupta, 2022).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

- Mean Absolute Error (MAE) is calculated by looking at the mean over the absolute differences between observed and predicted values. Both directions are treated the same which means that outliers do not play a role (MAE, 2022).

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

- Root Mean Squared Error (RMSE) shows how far the prediction values are from the true values using the Euclidean distance. As a rule of thumb, the lower the RMSE the better the model fits the data (Grace-Martin, 2022)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- R-Squared (R2) shows how well the model fits the data. If the R2 is close to 1, it shows that the model has predicted well. If the model is close to 0, the larger the distance between the actual and predicted values (R2, 2022).

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

# 4 Design Specification

The design specification for this research is shown below in Figure 2. It shows the process taken to complete this project and is broken up into a two-tiered architecture that consists of a Business Logic Layer and a Presentation Layer. The Business Logic Layer consists of the extraction, cleaning and pre-processing of the data as well as the implementation of the chosen machine learning models and the evaluation of the model performance. The second tier consists of the presentation layer, which involves presenting the results to our stakeholders. The stakeholders consist of the County Council, the UK Environment Agency and the public that access the water for recreational use.

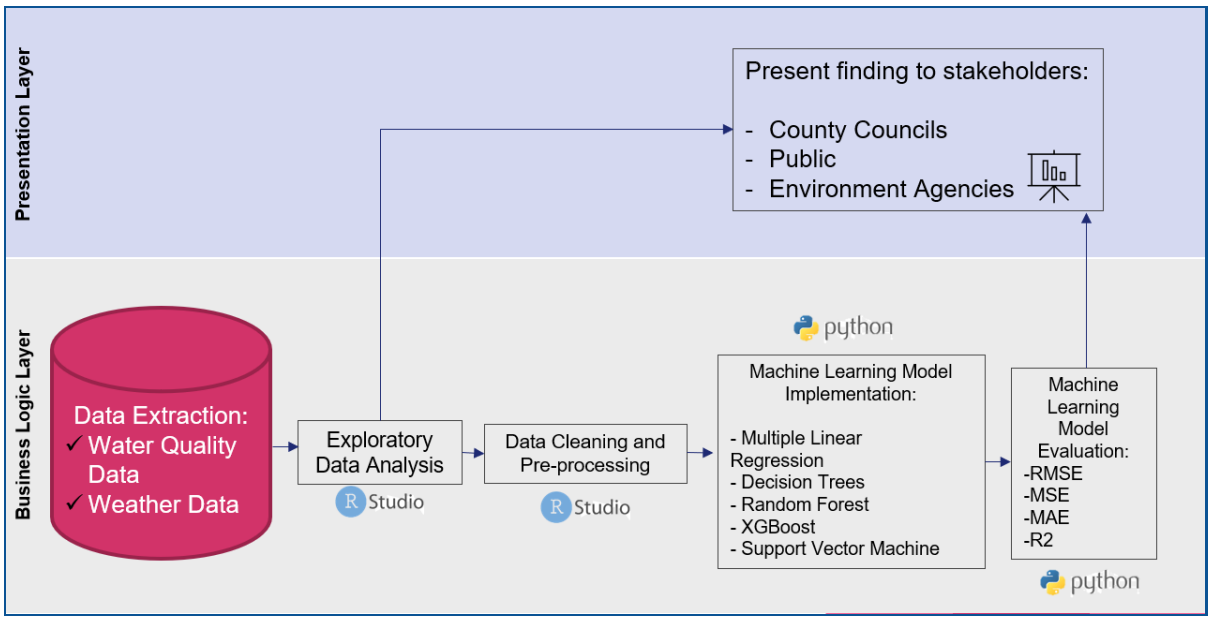

Figure 2: River Water Quality Prediction Design Specification

# 5 Implementation, Evaluation and Results of Water Quality Prediction Models

**Implementation:** Five supervised machine learning techniques were applied to the data, they include Decision Trees, Random Forest, Extreme Gradient Boosting, Multiple Linear Regression and Support Vector Machines. All models except Extreme Gradient Boosting were built using the Scikit-learn library in Python. Extreme Gradient Boosting was built using the XGBoost library in Python.

**Evaluation:** The models were evaluated using a selection of appropriate evaluation metrics from the sklearn.metrics Module in Python, Mean Squared Error, Root Mean Squared Error, Mean Absolute Error and R-Squared.

**Environmental setup:** This research was carried out using R Studio and Python Programming Language. R Studio was used for data cleaning and pre-processing while Jupyter Notebook was used for exploratory analysis, the implementation of the chosen machine learning models and the evaluation of their performance.

## 5.1 Data Exploration

Once the data is cleaned, the dataset was written to the directory and uploaded to Python for exploration. A number of visualisations were created. Figure 3 looks at the correlation between the variables in the dataset. It is clear that some negative and positive correlation exists between variables, yet it is evident that there is a weak correlation between our water quality parameter variables and our weather variables as shown in more detail in Figure 4.

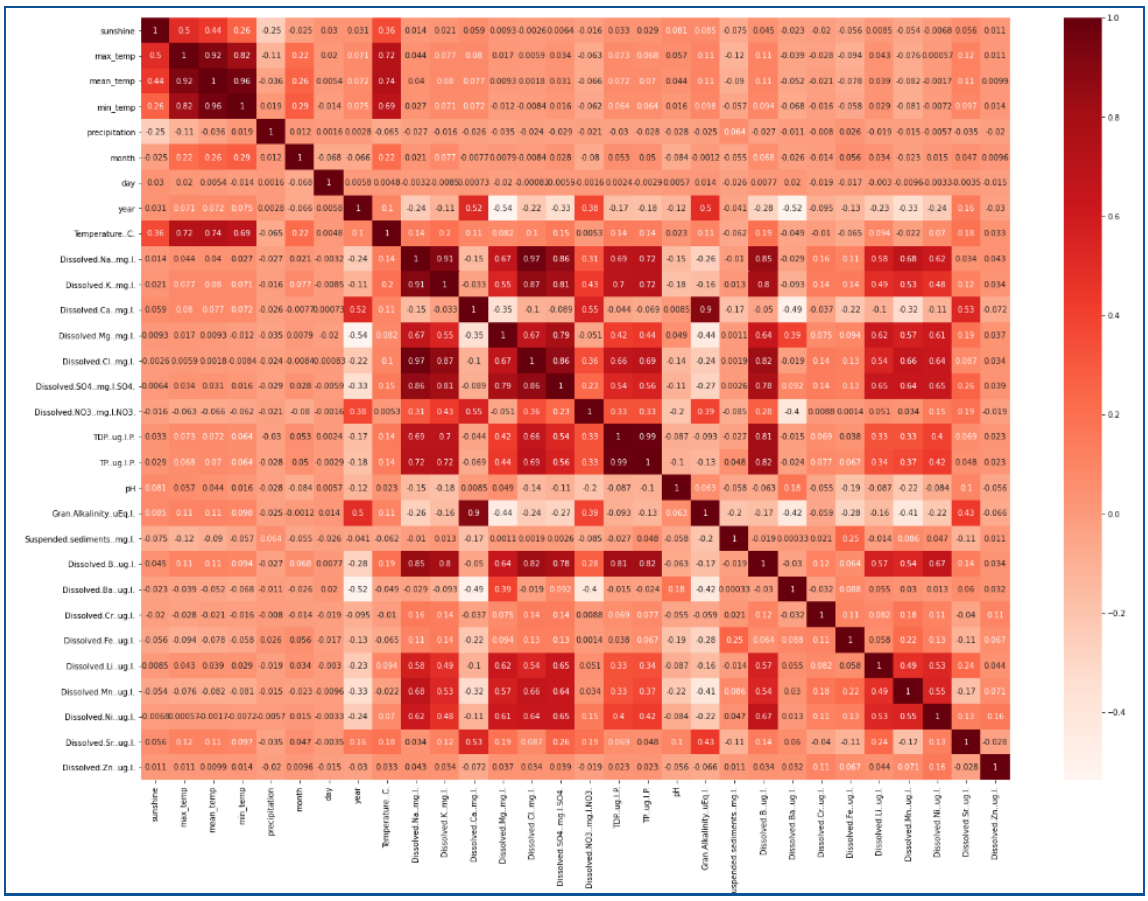


Figure 3: Correlation between River Water Quality variables.

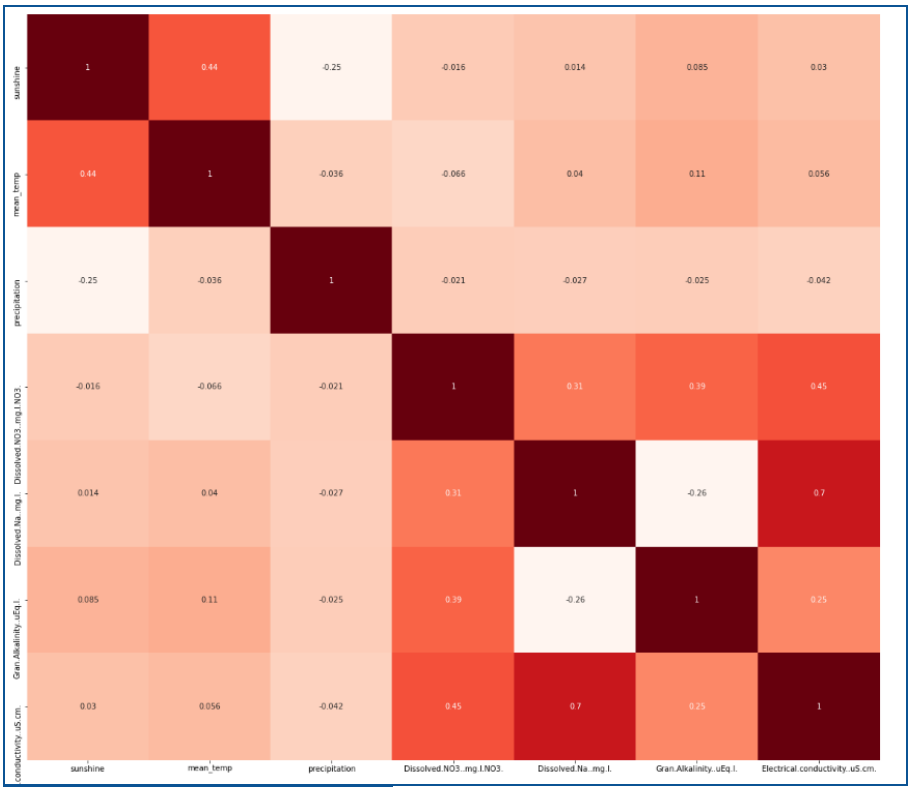

Figure 4: Correlation between River Water Quality variables and Weather variables.

## 5.2 Data Pre-processing

The datasets that were used for this research include the Water Quality Data collected from the Environmental Information Data Centre and the weather dataset that was collected from Kaggle and is originally sourced from the European Climate Assessment and measures weather information from a station located near Heathrow airport, in London. The River Water Quality dataset spans from 1993 to 2009 and contains over 9 thousand rows of daily water quality parameter measurements and 90 variables. The weather dataset spans from 1979 to 2020 and consists of over 15 thousand rows of daily weather observations and 10 variables. Both datasets were uploaded to R Studio using the read.csv library.

### 5.2.1 Data Transformation, Feature Engineering and Selection

Once loaded into R Studio, the Water Quality dataset was inspected for NA values. It can be seen that 46% of the dataset consisted of NA values, which is very high. As removing all NA values would decrease the dataset by close to 50% the decision was made to remove unwanted columns with a high presence of NA values and then to replace all remaining NA values left in the dataset with the median or mean of each column. The dataset originally consisted of 90 variables, though after removing unwanted columns and columns with greater than six thousand NA values, the dataset had 25 variables.

A histogram was then created for all remaining columns to check if they were evenly distributed or skewed. For columns that were evenly distributed the NA values were replaced with the mean, while columns that were skewed the NA values were replaced with the median (Kumar, 2022). The date format was then changed into a date data type. The Weather dataset consists of 10 variables, which contain less than 1% of NA values. As the amount of missing data was low, the decision was made to remove the NA values from the dataset. One column,

Snow Depth, was removed from the dataset as it was deemed unnecessary for the analysis. The date column was then separated in separate Year, Month and Day columns. Finally, the existing date column was changed into the data format date.

To create the final dataset, the Water Quality and Weather datasets were merged using an Inner Join on the date column. The final dataset consists of 9,490 rows and 38 weather and water quality parameter variables. These variables are shown below in Table 2.

**Table 2: Variables in the Dataset**

| **Variable Name** | **Description** | **Variable Name** | **Description** |
|---|---|---|---|
| *Mean Temperature* | Average daily temperature | *pH* | A measure of how acidic/basic water is |
| *Precipitation* | Daily Rainfall | *Gran Alkalinity* | An accurate test for low level samples |
| *Sunshine* | Daily sunshine hours | *Suspended Sediments* | Fine inorganic particles of clay and silt |
| *Dissolved Na* | Dissolved Sodium | *Dissolved B* | Dissolved Boron |
| *Dissolved K* | Dissolved Potassium | *Dissolved Ba* | Dissolved Barium |
| *Dissolved Ca* | Dissolved Calcium | *Dissolved Cr* | Dissolved Chromium |
| *Dissolved Mg* | Dissolved Magnesium | *Dissolved Fe* | Dissolved Iron |
| *Dissolved Cl* | Dissolved Chlorine | *Dissolved Li* | Dissolved Lithium |
| *Dissolved SO4* | Dissolved Sulphate | *Dissolved Mn* | Dissolved Manganese |
| *Dissolved NO3* | Dissolved Nitrate | *Dissolved Ni* | Dissolved Nickel |
| *TDP* | Total Dissolved Potassium | *Dissolved Sr* | Dissolved Strontium |
| *Electrical Conductivity* | Ability of water to conduct an electrical current | *Dissolved Zr* | Dissolved Zirconium |

The dataset was loaded into Python for exploration and modelling using the Pandas library. Before the models were applied to the dataset the data were scaled using the StandardScaler library from Scikit-learn in Python. As the data consists of a number of different measurements it was important to scale it so that the models could perform their best on data are of a similar scale and are close to being normally distributed. StandardScaler standardises a feature by subtracting the mean and then scaling it to unit variance. This scaling method was chosen as it is widely used in the industry (Hale, 2022). The dataset was then split into training and testing datasets. Four Water Quality parameters were chosen for prediction, they are Dissolved Sodium, Dissolved Nitrate, Gran Alkalinity and Electrical Conductivity. Each parameter's importance on the quality of water is discussed below.

***Electrical Conductivity:*** Measures the ability of water to conduct an electrical current. A higher concentration of dissolved charged chemicals in a river means that there is a greater electrical current which can indicate that the river water has become polluted and can harm inhabitants of the water (US EPA, 2022)

***Dissolved Sodium:*** Road salting, agriculture, sewage and new infrastructure can all lead to elevated levels of Sodium in our rivers. These high Sodium levels can damage our ecosystem and cause harm to animals (Lockwood, 2022).

***Dissolved Nitrate:*** Nitrogen can be harmful to people and nature and cause water to become polluted. High levels of nitrogen are typically caused by agricultural runoff and is one of the leading causes of water pollution in Europe (European Commission, 2022).

***Gran Alkalinity:*** Higher levels of Alkalinity can prevent harmful pH changes that can have a negative effect on water quality and be harmful to aquatic life. Natural and human factors can influence alkalinity levels in rivers, including weather factors such as rainfall and urbanisation which can cause construction particles to wash into rivers (Alkalinity and Hardness, 2022).

## 5.3 Data Modelling

Five supervised ML models were applied to the dataset, they are Decision Trees, Random Forest, Extreme Gradient Boosting, Multiple Linear Regression and Support Vector Machine. Before the models were applied, the data was first scaled and then split into 80% training and 20% testing datasets.

### 5.3.1 Implementation and Evaluation of Decision Tree Regression

Decision Tree Regression was the first algorithm applied to the dataset using the Scikit-learn library in Python. Decision Trees (DT) are a tree-based algorithm that uses a number of decisions to reach its final outcome. They can be used for both classification and regression problems. Although DT are relatively easy to use and interpret, they can be prone to overfitting and contain biases which should be kept in mind when using (Gupta, 2022).

The dataset was first split into training and testing datasets with an 80:20 split. As the dataset contains a number of different measurements it was scaled using the StandardScaler library from Scikit-learn. The DT algorithm was first applied to the training data to train the algorithm and then to the test data for prediction. Hyper parameter tuning using GridSearchCV from the Scikit-learn library was then applied to the model to see what parameters are optimal in our model; they are shown in Figure 5 below.

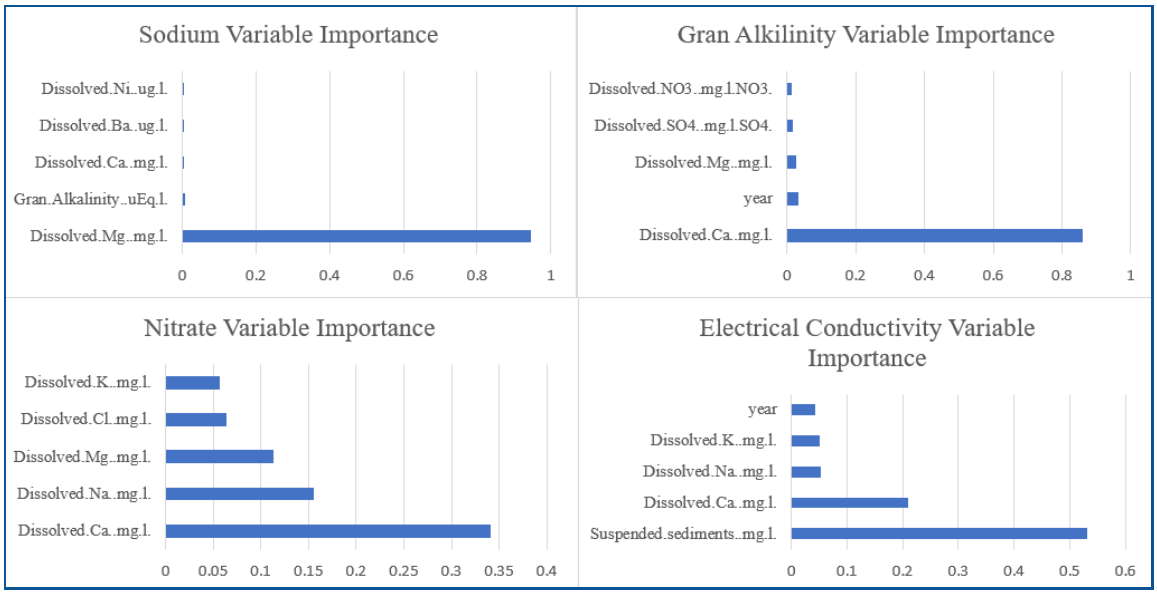

Figure 5: Decision Tree Feature importance for each parameter predicted.

The final model was evaluated using the Metrics library from Scikit-learn and the MAE, MSE, RMSE and R2 values were calculated and are shown in Table 3. It is shown that the parameter with the highest R-Squared is Sodium, followed closely by Gran Alkalinity at 91%. All parameters perform well with the lowest performance being Nitrate with an R-Squared of 78%.

**Table 3: Comparison of Decision Tree Model Performance**

| **Parameter** | **MAE** | **MSE** | **RMSE** | **R2** |
|---|---|---|---|---|
| Dissolved Nitrate | 0.283 | 0.228 | 0.478 | 78% |
| Dissolved Sodium | 0.067 | 0.026 | 0.161 | 97% |
| Gran Alkalinity | 0.170 | 0.085 | 0.292 | 91% |
| Electrical Conductivity | 0.197 | 0.186 | 0.431 | 82% |

The Expected and Predicted values of the best performing parameter Dissolved Sodium are shown in Figure 6 below. It can clearly be seen that the orange and blue lines which represent the expected and predicted values closely overlap each other showing the model is a good fit. The DT model has now been implemented and evaluated, ***Sub Obj 2.1*** has now been completed.

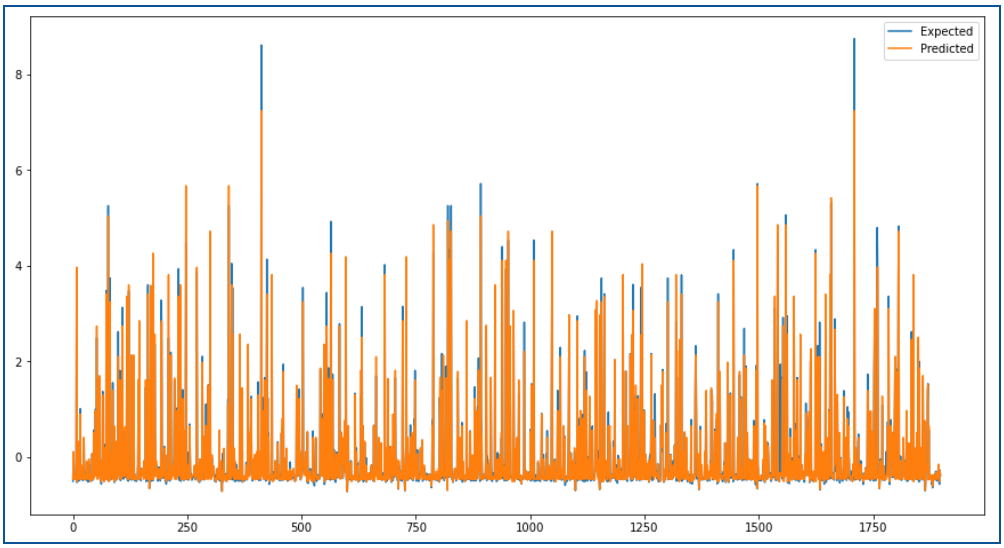

Figure 6: Expected Vs Predicted values for Dissolved Sodium

### 5.3.2 Implementation and Evaluation of Random Forest Regression

The next algorithm that was applied to our dataset is Random Forest. Random Forest (RF) is another tree-based algorithm that is commonly used for both classification and regression problems. It combines the output of multiple DT to come to a result and has been praised for its ease of use and accuracy (IBM Cloud Education, 2022). It has many benefits over traditional DT including the reduced risk of it overfitting and the ease at which it determines feature importance. It does however have its disadvantages with it being more time-consuming than other algorithms and its complexity (IBM Cloud Education, 2022).

The dataset was first split into training and testing datasets with a 80:20 split and the variables were scaled before the model was fit using the RandomForestRegressor library from Scikit-learn. The model was trained on the training data and then to the test data to make a prediction. The feature importance was then looked at to see what features were most important to the RF model. The model was then evaluated and the results are shown in Table 4. It can be seen that all parameters performed well with an R-Squared of 87% or above with Sodium having the highest R-Squared value at 98%. Sodium also has the lowest RMSE, MSE and MAE out of all parameters.

**Table 4: Comparison of Random Forest Model Performance**

| **Parameter** | **MAE** | **MSE** | **RMSE** | **R2** |
|---|---|---|---|---|
| Dissolved Nitrate | 0.171 | 0.087 | 0.296 | 92% |
| Dissolved Sodium | 0.046 | 0.018 | 0.133 | 98% |
| Gran Alkalinity | 0.130 | 0.049 | 0.222 | 95% |
| Electrical Conductivity | 0.152 | 0.128 | 0.358 | 87% |

A graph was created to show the expected values versus the predicted values for the best performing parameter Sodium. The results are shown in Figure 7 below with predicted values in orange and expected values in blue. The graph shows how close the predicted values are to the expected values with most of the blue line being overlapped by orange, indicating that the model is a good predictor of Sodium. The RF model has now been implemented and evaluated, ***Sub Obj 2.2*** has now been completed.

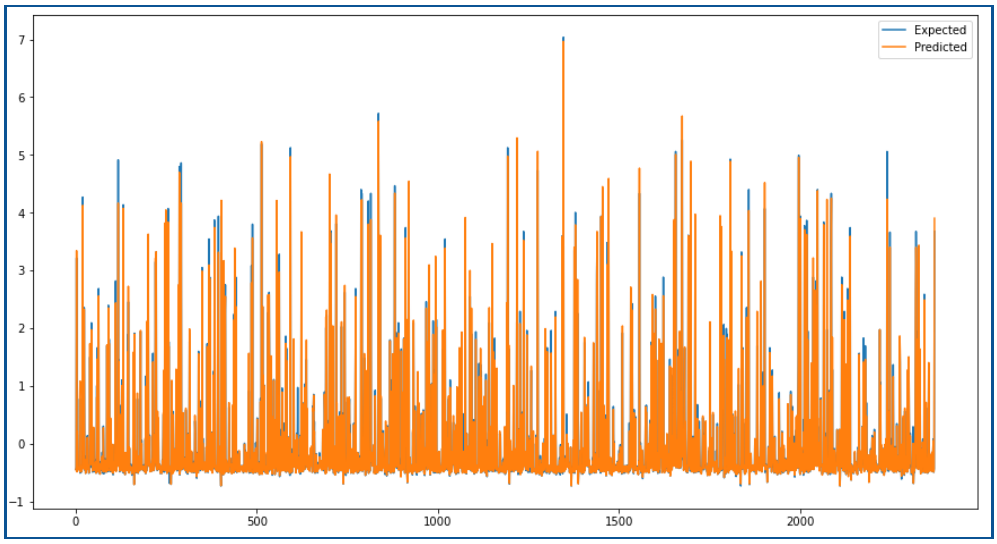

Figure 7: Expected Vs Predicted values for Dissolved Sodium

### 5.3.3 Implementation and Evaluation of Extreme Gradient Boosting

Extreme Gradient Boosting was applied to the dataset next. Extreme gradient boosting (XGBoost) is a ML algorithm that is tree-based. XGBoost can be used for both regression and classification and has the capacity to do parallel computations making it ten times faster than other gradient boosting models (Srivastava, 2022). Using the XGBRegressor function from the xgboost library in python, the model was first scaled before it was fit to the training data and then used on the test data for predicting. The model was then evaluated and the results are shown in Table 5. As shown below, the Gran Alkalinity and Electrical Conductivity parameters perform best with an R-Squared of 87%. Nitrate performed the lowest with an R-Squared of 79%.

**Table 5: Comparison of Extreme Gradient Boosting Model Performance**

| Parameter | MAE | MSE | RMSE | R2 |
|---|---|---|---|---|
| Nitrate | 0.312 | 0.185 | 0.430 | 79% |
| Sodium | 0.359 | 0.155 | 0.394 | 85% |
| Gran Alkalinity | 0.306 | 0.137 | 0.370 | 87% |
| Electrical Conductivity | 0.211 | 0.129 | 0.359 | 87% |

The expected values versus the predicted values for parameter Gran Alkalinity are plotted and shown in Figure 8 below. The Extreme Gradient Boosting model has now been implemented and evaluated, ***Sub Obj 2.3*** has now been completed.

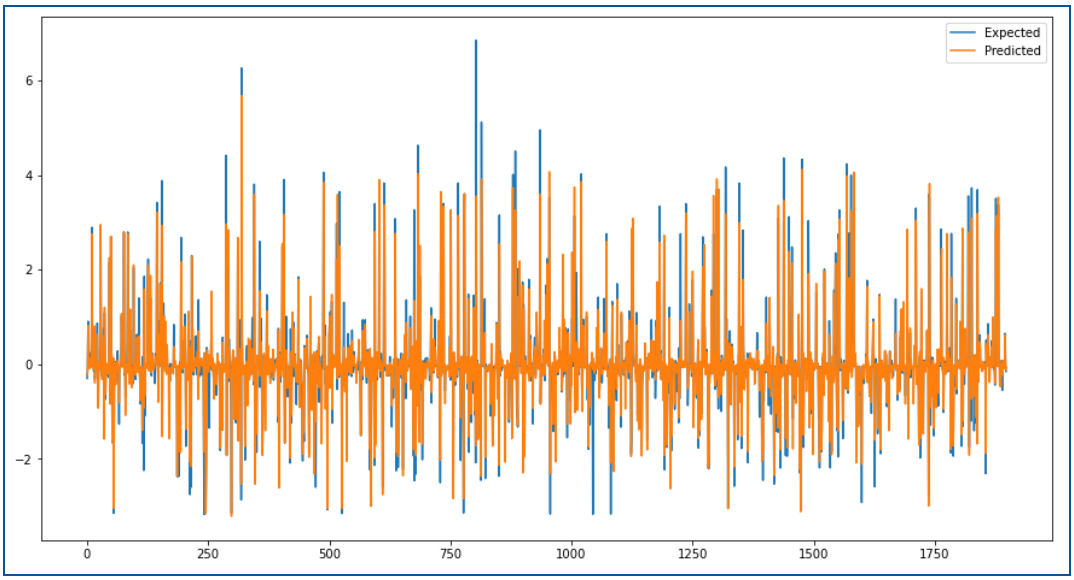

Figure 8: Expected Vs Predicted values for Gran Alkalinity

### 5.3.4 Implementation and Evaluation of Multiple Linear Regression

Multiple linear regression (MLR) is a statistical technique that uses two or more independent variables to predict the outcome of the dependent variable. Before applying the model, the dataset was first split into training and testing datasets with a 80:20 split and the variables were scaled before the model was fit using the Linear Regression library from Scikit-learn. Four parameters were tested and the model was then evaluated for each dependent variable. The results are shown in Table 6 below. Sodium was the best performing parameter with an R2 of 97% followed by Gran Alkalinity with an R2 of 89%. Nitrate performs the worst in all evaluation metrics with an R2 of only 60%.

**Table 6: Comparison of Multiple Regression Model Performance**

| Parameter | MAE | MSE | RMSE | R2 |
|---|---|---|---|---|
| Nitrate | 0.396 | 0.356 | 0.597 | 60% |
| Sodium | 0.095 | 0.028 | 0.167 | 97% |
| Gran Alkalinity | 0.202 | 0.110 | 0.332 | 89% |
| Electrical Conductivity | 0.280 | 0.210 | 0.458 | 79% |

A linear plot and a line chart were created for the best performing parameter Dissolved Sodium and is shown in Figure 9 below. It can be seen that the data is close to the line and follows a linear shape.

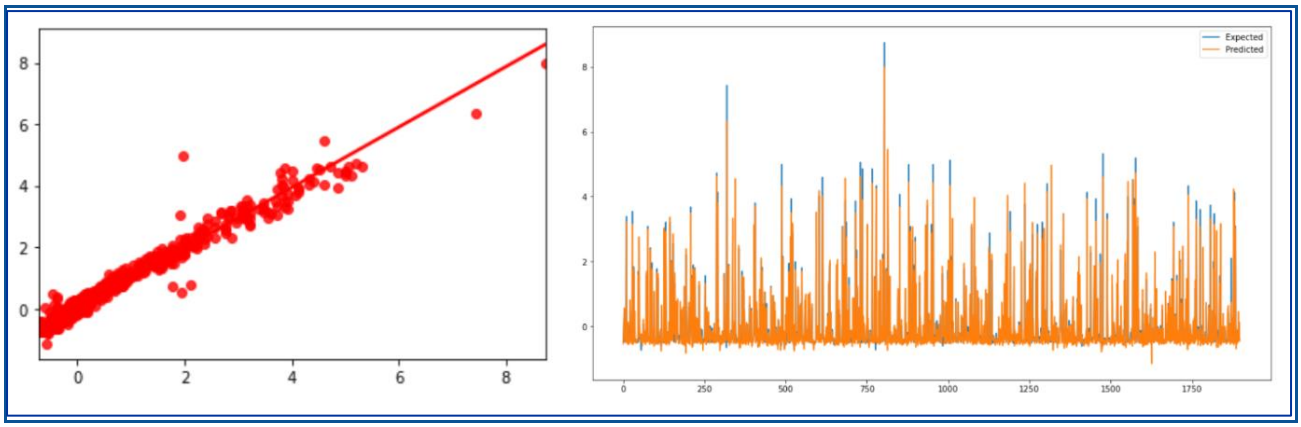

Figure 9: Expected Vs Predicted values for Dissolved Sodium.

It is important to note that MLR has a number of assumptions that must be met, these assumptions are discussed in more detail below and each assumption is tested on the data so that the model can be improved.

***Assumption 1:*** There is a linear relationship between the dependent and independent variables.
***Assumption 2:*** The data should not show multicollinearity which is when the independent variable is correlated with another independent variable.
***Assumption 3:*** Homoscedasticity which means that the residuals have a constant variance.
***Assumption 4:*** Multivariate Normality which occurs when the distribution of the residuals are normal.
**Assumption 5:** Observations should be independent of each other.

To improve the performance of the model at predicting Dissolved Nitrate, each assumption was tested on the data. The linear relationship between each dependent variable and the independent variables were assessed using scatter plots. Results Found that there were quite a few variables that did not have a linear relationship with the dependent variable, which were removed from the dataset. Next, multicollinearity was tested by calculating the VIF for each variable and removing those that had a value above 10. Homoscedasticity and Multivariate Normality were tested by creating a scatterplot between the residuals and the dependent variable and a histogram of the residuals with a normal curve; the results are shown in Figure 10. It can be seen that Homoscedasticity may exist within the dataset as the data points are not evenly distributed randomly in the plot. Multivariate Normality appears to exist as the graph shows an even distribution.

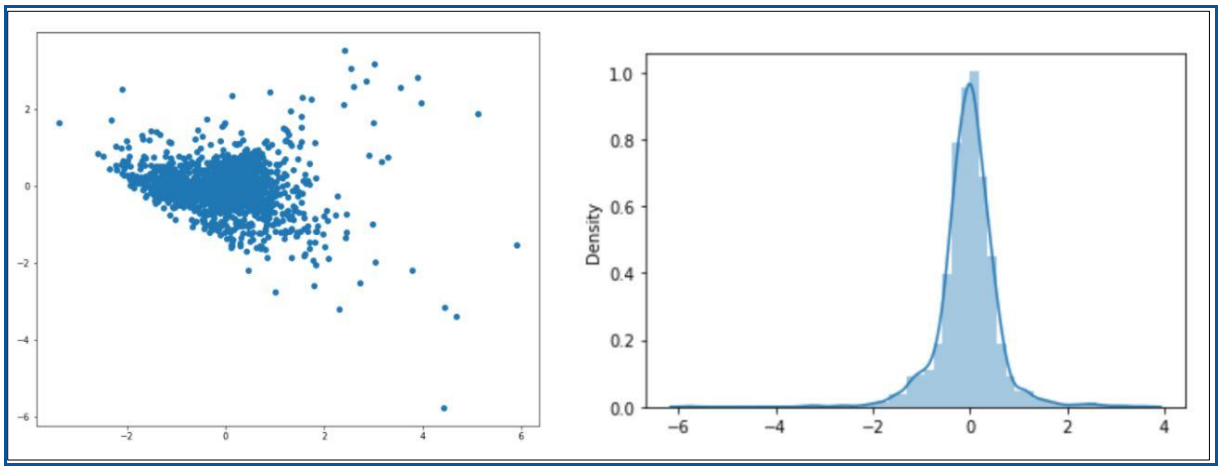


Figure 10: Scatter plot and Histogram with Normal Curve of the residuals.

Finally assumption 5, the observations should be independent of each other, was tested by using the statsmodels.stats package in Python to calculate the Durban Watson statistic. It was found that the Durban Watson statistic had a value of 1.4 indicating that there is no autocorrelation and the observations are independent of each other. After completing assumption testing, the MLR model was fit to the new dataset and produced an R2 of 58%. As this value is lower than the original R2 value, this model may not be suitable for predicting River Water Quality and the parameter Dissolved Nitrate. The MLR model has now been implemented and evaluated, ***Sub Obj 2.4*** has now been completed.

### 5.3.5 Implementation and Evaluation of Support Vector Machines

The final algorithm that was applied to the dataset is Support Vector Machine using the SVR library from Scikit-learn. Support Vector Machine algorithm is a powerful supervised ML algorithm that can be used for both classification and regression problems and is known for its ability to produce high accuracy scores with low computational power compared to other algorithms. Support Vector Machines work well in smaller datasets that do not have a high level of noise (Ray, 2022). The dataset was split into training and testing datasets with a 80:20 split. Variable values were scaled before the model was fit. The model was trained on the training data and then fit to the test data to make a prediction. The model was then evaluated and the results are shown in Table 8.

It can be seen that all parameters perform well with R-Squared values of 85% and above. The parameter Sodium once again performs the best for its prediction accuracy with Gran Alkalinity having the second highest R-Squared.

**Table 8: Comparison of Support Vector Machine Model Performance**

| Parameter | MAE | MSE | RMSE | R2 |
|---|---|---|---|---|
| Nitrate | 0.227 | 0.149 | 0.386 | 85% |
| Sodium | 0.078 | 0.038 | 0.196 | 96% |
| Gran Alkalinity | 0.146 | 0.070 | 0.264 | 93% |
| Electrical Conductivity | 0.176 | 0.137 | 0.370 | 86% |

The expected values versus the predicted values for the parameter Dissolved Sodium are shown in Figure 11 below. The graph shows the prediction accuracy of the model and shows that it performed well. The Support Vector Machine model has now been implemented and evaluated, ***Sub Obj 2.5*** has now been completed.

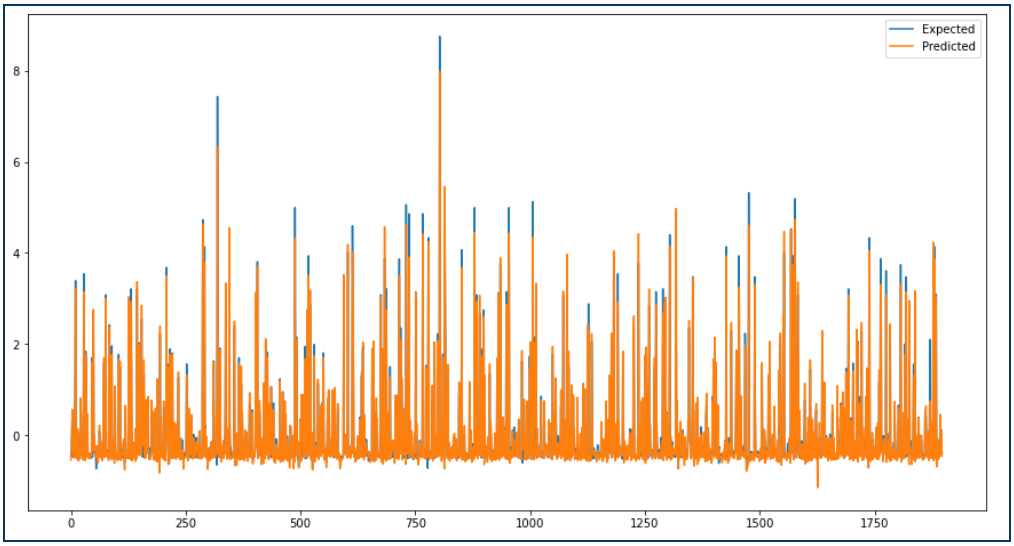

Figure 11: Expected Vs Predicted values for Dissolved Sodium.

All models have now been implemented and evaluated, ***Obj 2*** is now complete.

# 6 Discussion and Results

This chapter will look at the results of each of the five ML models used for the prediction of river water quality parameters and compare them to identify the optimal model. The five models in this research will also be compared to the models identified in the reviewed literature.

## 6.1 Comparison of Developed River Water Quality Models

Five ML models were successfully applied to the data and four river quality parameters were predicted. The models were then evaluated using R2, MAE, MSE and RMSE. All models performed well when predicting the four chosen river quality parameters, with no model having an R-Squared value of below 60%. Random Forest Regression appears to have performed the best when predicting all four parameters, with the lowest R-Squared value being 87% for the prediction if Electrical Conductivity. Table 9 clearly shows the performance of all five algorithms when predicting the four water quality parameters. It can be seen that RF performs the best in all evaluation metrics for predicting all four parameters, including Nitrate, Sodium and Gran Alkalinity and Electrical Conductivity. Random Forest predicted the parameter Sodium the best with an R-Squared of 97%. Multiple linear regression performs the worst when predicting two of the parameters, such that Nitrate and Electrical Conductivity reported an R-Square of 60% and 79% respectively. This may be due to homoscedasticity that appears to exist within the dataset. Multiple Linear Regression performs best when predicting the parameter Sodium with an R-Squared value of 97%. Extreme Gradient Boosting performs worst out of all models when predicting both Sodium and Gran Alkalinity, nonetheless it performs joint first when predicting Electrical Conductivity with an R-Squared value of 87%. Support Vector Machines also perform well when predicting the four parameters with no R-Squared value below 85%. It performs best when predicting Sodium with an R-Squared value of 96%.

**Table 9: Comparison of Machine Learning Models Performance**

| Algorithm | Parameter | MAE | MSE | RMSE | R2 |
|---|---|---|---|---|---|
| Decision Trees | Dissolved Nitrate | 0.283 | 0.228 | 0.478 | 78% |
| Random Forest | | 0.171 | 0.087 | 0.296 | 92% |
| XGBoost | | 0.312 | 0.185 | 0.430 | 79% |
| Support Vector Machine | | 0.227 | 0.149 | 0.386 | 85% |
| Multiple Linear Regression | | 0.396 | 0.356 | 0.597 | 60% |
| Decision Trees | Dissolved Sodium | 0.067 | 0.026 | 0.161 | 97% |
| Random Forest | | 0.046 | 0.018 | 0.133 | 98% |
| XGBoost | | 0.359 | 0.155 | 0.394 | 85% |
| Support Vector Machine | | 0.078 | 0.038 | 0.196 | 96% |

| | | | | | |
|---|---|---|---|---|---|
| Multiple Linear Regression | | 0.095 | 0.028 | 0.167 | 97% |
| Decision Trees | Gran Alkalinity | 0.170 | 0.085 | 0.292 | 91% |
| Random Forest | | 0.130 | 0.049 | 0.222 | 95% |
| XGBoost | | 0.306 | 0.137 | 0.370 | 87% |
| Support Vector Machine | | 0.146 | 0.070 | 0.264 | 93% |
| Multiple Linear Regression | | 0.202 | 0.110 | 0.332 | 89% |
| Decision Trees | Electrical Conductivity | 0.197 | 0.186 | 0.431 | 82% |
| Random Forest | | 0.152 | 0.128 | 0.358 | 87% |
| XGBoost | | 0.211 | 0.129 | 0.359 | 87% |
| Support Vector Machine | | 0.176 | 0.137 | 0.370 | 86% |
| Multiple Linear Regression | | 0.280 | 0.210 | 0.458 | 79% |

The importance of each feature used was also looked at when implementing the DT Model and RF Model. It can be seen from Figure 5 that Dissolved Magnesium is the most important feature when predicting Dissolved Sodium levels, Dissolved Calcium is the most important feature when predicting Gran Alkalinity and Dissolved Nitrate, and finally suspended sediments is the most important variable for predicting Electrical Conductivity. It can also be seen from this table that the weather variables included in the models do not appear in the top five variables which may suggest that they do not play an important role in the models.

A comparison of machine learning models used in this research and the importance of features has now been completed, ***Obj 3*** and ***Obj 4*** are now complete.

## 6.2 Comparison of Developed River Water Quality Models and Existing Models

Multiple Linear Regression is used by Thoe et al. (2014) and Wang et al. (2021) to predict water quality in coastal waters and in estuaries respectively. Thoe et al. (2014) achieve an RMSE value of 0.4 for the prediction of E-Coli while Wang et al. (2021) achieve an RMSE

of 1.15 and an R-Squared of 91% for the prediction of Ammonium. The multiple regression models in this study achieve a lower RMSE than 0.4 for predicting Sodium and Gran Alkalinity parameters. This shows that the models in this study achieve a better fit than the one used by Thoe et al. (2014). Furthermore, this study achieves a higher R-Square value and a lower RMSE value for the prediction of Sodium compared to the Multiple Regression model used by Wang et al. (2021). The authors also use RF and XGBoost for the prediction of Ammonium in estuaries with results achieving an R-Squared value of 94% and an RMSE of 1.14 for their RF model (Wang et al., 2014). Compared to the RF models used in this study we can see that the models in this study outperform theres with an R-Squared of 98% for the prediction of Sodium and an R-Squared of 95% for the prediction of Gran Alkalinity with an RMSE that is lower than that seen in Wang et al. (2021) study. Ahmed et al. (2019), also apply RF, Support Vector Machines, and Polynomial Regression for the prediction of water quality. The RF and Support Vector Machine models in this study outperform those in Ahmed et al. (2019) study significantly with their RF achieving an R-Squared values of 67% compared to an R-Squared value of 98% in this study and an R-Square value of 35% for their Support Vector Machine compared to 96% in this study.

Classification models were looked at by Radhakrishnan & Pillai (2022). The authors used both Support Vector Machine and DT to classify poor water quality and achieved an R-Squared of 87% for both. Compared to the models used in this study it can be seen that a higher R-Squared was achieved for the prediction of both Sodium and Gran Alkalinity. Finally, the models used in this study are compared to Neural Networks that were used in the literature reviewed. Chafloque et al. (2021) use neural networks to predict the quality of water for human consumption. Results found that the accuracy of their model is 69% which is significantly smaller than the models used in this study. Rani et al. (2022) use an Artificial Neural Network for their research that looks at the prediction of drinking water quality and achieve an MSE and RMSE of 1.12 and 1.09 respectively, which is higher than both the MSE and RMSE achieved by all models used in this study.

As evident from this comparison, supervised ML methods in this study outperform those used in the literature reviewed and outperform both neural networks and classification methods. This further enforces the decision to use the selected supervised machine learning regression models in this research. A comparison of ML models used in literature has now been completed, ***Obj 5*** is now complete.

# 7 Conclusion and Future Work

The research questions for this study is *"How successfully can supervised machine learning techniques be used to predict river water quality parameters to assist in monitoring water quality in a timely manner?"* the following section will aim to answer this question. After reviewing similar literature five suitable ML models were selected, they are DT, RF,

Extreme Gradient Boosting, Support Vector Machine and MLR. Each model was then evaluated with methods suitable for regression, including MAE, MSE, RMSE and R2. It was found that RF has the highest R-Squared value for predicting all four parameters, they include Nitrate, Sodium, Gran Alkalinity and Electrical Conductivity. The MAE, MSE and RMSE, also remain low for all parameters indicating that the models are a good fit for the data. It can also be seen that Sodium is the water quality parameter with the highest prediction accuracy for four out of the five models and therefore could be recommended as a parameter to use for the prediction of river water quality in the future. Multiple Linear Regression appears to perform worse when predicting Nitrate with an R-Squared of only 60% and a relatively higher MSE, MAE and RMSE than other models used in this research. When looked at further, it was

seen that there are some variables that do not have a linear relationship with the Nitrate parameter which could be the cause for the low prediction accuracy. It can also be seen that homoscedasticity may exist within the dataset which could also be contributing to the poor results. This could be looked at in the future and the data set manipulated using the log function to create a linear relationship so the fit of the models could be improved. Additional weather parameters such as rainfall, air temperature and sunshine hours were added to the data to investigate if they would improve the performance of the model. As stated in the literature reviewed these additional parameters do play an important role in water quality. High levels of rainfall cause chemicals from industry and agriculture to enter into rivers and cause pollution and high air temperatures cause pollutants to become more toxic in the water. However, it was seen that in the case of this research, the additional weather parameters did not play an important role in the models and were ranked out of the top ten when looking at feature importance for the DT model, RF and Extreme Gradient Boosting. This is something that can be looked at in the future to improve the model's prediction accuracy. In addition, different weather variables could be used or a different weather dataset that uses measurements taken from a weather station at a more exact location than where the river samples were taken. Future work could also involve the application of the models used in this study to other water types, such as drinking water and coastal water to see if the high levels of accuracy remain. The best performing algorithm has been identified and recommendations have been made, ***Obj 6*** is now complete. The research question *"How successfully can supervised machine learning techniques be used to predict river water quality parameters to assist in monitoring water quality in a timely manner?"* has been answered.

# Acknowledgement

I would like to thank my supervisor Dr. Catherine Mulwa who has supported and guided me through this research. I would also like to thank my family, friends and Arthur who have provided me with continuous support while completing this research.

# References

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A., Irfan, R. and García-Nieto, J., 2019. Efficient Water Quality Prediction Using Supervised Machine Learning. *Water*, 11(11), p.2210.

Banda, T. and Kumarasamy, M., 2020. Development of Water Quality Indices (WQIs):A Review. *Polish Journal of Environmental Studies*, 29(3), pp.2011-2021.

Chafloque, R., Rodriguez, C., Pomachagua, Y. and Hilario, M., 2021. Predictive Neural Networks Model for Detection of Water Quality for Human Consumption. *2021 13th International Conference on Computational Intelligence and Communication Networks (CICN)*, 13, pp.1-5.

Data.gov.uk. 2022. UK lowland river chemistry - data.gov.uk. [online] Available at: <https://www.data.gov.uk/dataset/c1d7e6d9-bb3e-4971-b241-da438a643a0b/uk-lowland-river-chemistry> [Accessed 15 March 2022].

Dawood, T., Elwakil, E., Novoa, H. and Gárate Delgado, J., 2021. Toward urban sustainability and clean potable water: Prediction of water quality via artificial neural networks. *Journal of Cleaner Production*, 291, pp.1-12.

Ec.europa.eu. 2022. *Nitrates - Water pollution - Environment - European Commission*. [online] Available at: <https://ec.europa.eu/environment/water/water-nitrates/index_en.html> [Accessed 1 August 2022].

Ewusi, A., Ahenkorah, I. and Aikins, D., 2021. Modelling of total dissolved solids in water supply systems using regression and supervised machine learning approaches. *Applied Water Science*, 11(2), pp.1-16.

Extension.usu.edu. 2022. *Alkalinity and Hardness*. [online] Available at: <https://extension.usu.edu/waterquality/learnaboutsurfacewater/propertiesofwater/alkalinity> [Accessed 1 August 2022].

EZE, J., SCOTT, E., POLLOCK, K., STIDSON, R., MILLER, C. and LEE, D., 2013. The association of weather and bathing water quality on the incidence of gastrointestinal illness in the west of Scotland. *Epidemiology and Infection*, 142(6), pp.1289-1299.

Grace-Martin, K., 2022. *Measures of Model Fit for Linear Regression Models*. [online] The Analysis Factor. Available at: <https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/> [Accessed 5 August 2022].

Gupta, A., 2022. *Mean Squared Error : Overview, Examples, Concepts and More*. [online] simplilearn. Available at: <https://www.simplilearn.com/tutorials/statistics-tutorial/mean-squared-error#:~:text=The%20Mean%20Squared%20Error%20measures,it%20relates%20to%20a%20function.> [Accessed 5 August 2022].

Hale, J., 2022. *Scale, Standardize, or Normalize with Scikit-Learn*. [online] Medium. Available at: <https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02> [Accessed 4 August 2022].

Hamaty, E., Faiek, S., Nandi, M., Stidd, D., Trivedi, M. and Kandukuri, H., 2020. A Fatal Case of Primary Amoebic Meningoencephalitis from Recreational Waters. *Case Reports in Critical Care*, 2020, pp.1-6.

IBM Cloud Education., 2022. *What is Random Forest?*. [online] Ibm.com. Available at: <https://www.ibm.com/cloud/learn/random-forest> [Accessed 10 July 2022].

Kaggle.com. 2022. London Weather Data. [online] Available at: <https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data> [Accessed 15 May 2022].

Khullar, S. and Singh, N., 2022. River Water Quality Classification using a Hybrid Machine Learning Technique. *2022 9th International Conference on Computing for Sustainable Global Development*, 9, pp.808-813.

Knowledge Center. 2022. *MAE / Mean absolute error*. [online] Available at: <https://peltarion.com/knowledge-center/documentation/evaluation-view/regression-loss-metrics/mae-/-mean-absolute-error> [Accessed 4 August 2022].

Kumar, A., 2022. *Python - Replace Missing Values with Mean, Median & Mode - Data Analytics*. [online] Data Analytics. Available at: <https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/#:~:text=You%20can%20use%20mean%20value,to%20replace%20the%20missing%20values.> [Accessed 14 August 2022].

Kurniawan, I., Hayder, G. and Mustafa, H., 2021. Predicting Water Quality Parameters in a Complex River System. *Journal of Ecological Engineering*, 22(1), pp.250-257.

Li, L., Qiao, J., Yu, G., Wang, L., Li, H., Liao, C. and Zhu, Z., 2022. Interpretable tree-based ensemble model for predicting beach water quality. *Water Research*, 211, pp.1-12.

Li, X., Ding, J. and Ilyas, N., 2020. Machine learning method for quick identification of water quality index (WQI) based on Sentinel-2 MSI data: Ebinur Lake case study. *Water Supply*, 21(3), pp.1291-1312.

Lockwood, D., 2022. *For healthier lakes, rivers, and drinking water, hold the salt*. [online] Cen.acs.org. Available at: <https://cen.acs.org/environment/water/healthier-lakes-rivers-drinking-water/97/i6> [Accessed 1 August 2022].

Lopez, A., Haripriya, N., Raveendran, K., Baby, S. and Priya, C., 2021. Water quality prediction system using LSTM NN and IoT. *2021 IEEE International Power and Renewable Energy Conference (IPRECON)*, pp.1-6.

Maloo, A., Fulke, A., Khade, K., Sharma, A. and Sukumaran, S., 2018. Virulence gene and antibiogram profile as markers of pathogenic Escherichia coli in tropical beaches of North Western India: Implications for water quality and human health. *Estuarine, Coastal and Shelf Science*, 208, pp.118-130.

Mohammed Ati, E., Jazar, Z., Ajmi, R. and Zeki Farooq, H., 2020. Water pollution and its relationship to human health: A review. *EurAsian Journal of BioSciences*, 14, pp.7473-7476.

Nair, J. and Vijaya, M., 2021. Predictive Models for River Water Quality using Machine Learning and Big Data Techniques - A Survey. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pp.1-7.

Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., Ehteram, M. and Elshafie, A., 2019. Machine learning methods for better water quality prediction. *Journof Hydrology*, 578, pp.1-18.

Nazeer, M., Bilal, M., Alsahli, M., Shahzad, M. and Waqas, A., 2017. Evaluation of Empirical and Machine Learning Algorithms for Estimation of Coastal Water Quality Parameters. *ISPRS International Journal of Geo-Information*, 6(11), pp.1-15.

Oh, H., Jeong, M., Jeon, S., Lee, T., Kim, G. and Youm, M., 2021. Sea Water Quality Estimation Using Machine Learning Algorithms. *Journal of Coastal Research*, 114(sp1), pp.425-427.

Parker, J. and Frost, S., 2000. Environmental health aspects of coastal bathing water standards in the UK. *Environmental Management and Health*, 11(5), pp.447-454.

Peltarion. 2022. *R2 / R-squared*. [online] Available at: <https://peltarion.com/knowledge-center/documentation/evaluation-view/regression-loss-metrics/r2-/-r-squared> [Accessed 4 August 2022].

Radhakrishnan, N. and Pillai, A., 2020. Comparison of Water Quality Classification Models using Machine Learning. *Proceedings of the Fifth International Conference on Communication and Electronics Systems*, pp.1-6.

Ragi, N., Holla, R. and G, M., 2019. Predicting Water Quality Parameters Using Machine Learning. *International Conference on Recent Trends on Electronics, Information, Communication & Technology*, 4, pp.1-4.

Rani, J., Sujeethra, R. and Rubavathi, C., 2022. Prediction of Water Quality using Artificial Neural Network. *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 4, pp.1-12.

Ray, S., 2022. *SVM | Support Vector Machine Algorithm in Machine Learning*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/> [Accessed 10 July 2022].

Schets F., De Roda Husman, A. and Havavelaar, A., 2010. Disease outbreaks associated with untreated recreational water use. Epidemiology and Infection, 139(7), pp.1114-1125.

Srivastava, T., 2022. *XGBoost In R | A Complete Tutorial Using XGBoost In R*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2016/01/xgboost-algorithm-easy-steps/#:~:text=Extreme%20Gradient%20Boosting%20(xgboost)%20is,computation%20on%20a%20single%20machine.> [Accessed 9 July 2022].

Su, H., Lu, X., Chen, Z., Zhang, H., Lu, W. and Wu, W., 2021. Estimating Coastal Chlorophyll-A Concentration from Time-Series OLCI Data Based on Machine Learning. *Remote Sensing*, 13(4), p.576.

Thoe, W., Gold, M., Griesbach, A., Grimmer, M., Taggart, M. and Boehm, A., 2014. Predicting water quality at Santa Monica Beach: Evaluation of five different models for public notification of unsafe swimming conditions. Water Research, 67, pp.105-117.

Uddin, M., Nash, S., Rahman, A. and Olbert, A., 2022. A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment. *Water Research*, 219, pp.1-20.

US EPA. 2022. *Indicators: Conductivity | US EPA*. [online] Available at: <https://www.epa.gov/national-aquatic-resource-surveys/indicators-conductivity> [Accessed 2 August 2022].

Wang, S., Peng, H. and Liang, S., 2022. Prediction of estuarine water quality using interpretable machine learning approach. *Journal of Hydrology*, 605, pp.1-12.

Xu, J., Xu, Z., Kuang, J., Lin, C., Xiao, L., Huang, X. and Zhang, Y., 2021. An Alternative to Laboratory Testing: Random Forest-Based Water Quality Prediction Framework for Inland and Nearshore Water Bodies. *Water*, 13(22), p.3262.