National College of Ireland

Groundwater Quality Predictive Analysis Using Machine Learning Techniques: Ireland

MSc Research Project
MSc in Data Analytics

Leslie Rebeca Monroy Ochoa
Student ID: x23169761

School of Computing
National College of Ireland

Supervisor: Jaswinder Singh

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Leslie Rebeca Monroy Ochoa
Student ID:	x23169761
Programme:	MSc in Data Analytics **Year:** 2023-2024
Module:	MSc Research Project
Supervisor:	Jaswinder Singh
Submission Due Date:	Monday 16th September 2024
Project Title:	Groundwater Quality Predictive Analysis Using Machine Learning Techniques: Ireland
Word Count:	7,557 words **Page Count** 21 pages

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: Leslie Rebeca Monroy Ochoa

Date: 13th September 2024

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):	

Groundwater Quality Predictive Analysis Using Machine Learning Techniques: Ireland

Leslie Rebeca Monroy Ochoa
x23169761

Abstract

One of the most important water resources is groundwater, essential for drinking water, agriculture, industry, and environmental sustainability. Ensuring its quality is very crucial for public and ecosystem health. This study applied four supervised machine learning models—Decision Tree, Random Forest, Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM)—to predict key groundwater quality parameters: Alkalinity, Dissolved Oxygen, Conductivity, and Nitrate, using data from monitoring stations across Ireland. Among the models, Random Forest and XGBoost demonstrated superior performance, with Random Forest achieving the highest Accuracy (0.9599), closely followed by XGBoost (0.9562). These results highlighted the potential of machine learning to enhance groundwater monitoring, offering a more efficient, cost-effective, and accurate approach for the analysis of environmental data compared to conventional methods.

Keywords: Groundwater Quality, Ireland, Machine Learning, Predictive Models

1 Introduction

Water is by far, if not the most important, resource on earth. Water is not only used as drinking water[1], but also in different industries for energy production as well as manufacturing processes that are necessary in agriculture, transportation, tourism, entertainment, and recreational sports (Narasaiah, 2005). The subject of our research is groundwater, which is one of the most important types of water bodies on the planet. About 30% of the freshwater used worldwide comes from groundwater, which is especially important for providing access to water during dry seasons in arid regions (Lo et al. 2016).

Groundwater, the water found beneath the surface in spaces of rock, sand, and soil, is the main supply of drinking water for many countries and communities (Younger, 2009). It keeps rivers and wetlands flowing, helping to preserve healthy water ecosystems[2]. So, whether we are talking about surface water or groundwater, the quality of the water is crucial for life, public health, and other daily activities (Parkes et al., 2010). In the past, the Republic of Ireland's groundwater monitoring program prioritised protecting drinking water sources and examining the effects of point source pollution. In order to evaluate the effects of both human activity and climate change on groundwater systems, long-term monitoring is

[1] https://www.gsi.ie/en-ie/programmes-and-projects/groundwater/Pages/default.aspx
[2] https://www.epa.ie/our-services/monitoring--assessment/freshwater--marine/groundwater/

necessary to determine the levels and quality of groundwater (Williams, 2008). Traditionally, this monitoring requires gathering water samples over the seasons and through the years and involves comprehensive analyses to discover patterns and contrast the different properties of the water, and often these procedures need a significant investment of time and money (Wood, 1976).

However, as technology develops, the application of machine learning techniques in the analysis of this data has represented a major advancement in the field of water quality. These techniques help in the knowledge regarding water quality facilitating its comprehension of differences and more intelligent decision-making (Parnika et al., 2024). By identifying and analysing trends, researchers have gained important insights that help them create more precise projections and provide the means to take a proactive approach to environmental and water management initiatives. Relevant literature contains a variety of methods, including both classification and regression approaches to assess and forecast water quality; for instance, Krushna et al. (2024) and Parnika et al. (2024) used a variety of models, including Decision Tree, Support Vector Machine (SVM), Random Forest Classifier, Logistic Regression, XGBoost Classifier, among others.

In this context, this project aims to understand and tackle problems related to water quality, specifically focuSsing on the Republic of Ireland's groundwater resources and on several techniques for assessing and forecasting groundwater quality parameters. Saying that this research problem leads to the formulation of the following research question:

Research Question: To what extent can supervised machine learning models predict geological changes and detect groundwater contaminants, compared to traditional monitoring techniques?

To answer this question, the following objectives were developed:
Objective 1. Critical analysis of relevant water and groundwater quality prediction literature.
Objective 2. Implementation of Groundwater Quality Classification to analyse groundwater parameters.
Objective 3. Implementation and assessment of four supervised machine learning models.
Objective 4. Evaluation, comparison, and contrast of the performance of the implemented models.
Objective 5. Identification of the model with the best performance and results evaluation.

The rest of this document is organised into distinct sections to ensure clarity and facilitate understanding. The document begins with a Related Work section, providing a critical review of previous studies, and highlighting the impacts of related technologies and methodologies on the current research domain. It also compares existing research in the field of machine learning and its application to projecting water and groundwater quality parameters, analysing relevant studies, methodologies, findings, and gaps. Next, the Research Methodology section outlines the defined methodology for the study, detailing the tools and techniques for data collection and analysis, along with the performance metrics defined for

evaluating the results. The Design Specification section describes the framework and design principles adopted for the study. The Implementation section presents the development and integration processes, explaining the steps taken to bring the design to completion. The Evaluation section examines the results, assessing the effectiveness of the implemented methods against the performance metrics, and discusses the implications of the findings. The Conclusion and Future Work section summarises the key contributions of the study, reflects on the limitations, and points out potential directions for future research. Finally, the References section lists all the sources cited throughout the document.

2 Related Work

The quality of the different bodies of water that exist is of utmost importance for humans and ecosystems, since we depend largely on them for various reasons, the main ones being drinking water, sanitation and agriculture. This section explores known studies and approaches for assessing and forecasting the variables and circumstances that affect water quality, with a focus on groundwater water bodies in the Republic of Ireland.

2.1 Water Quality and Health Implications

It is well known that there is a direct association between bad water quality and serious health problems. When water gets contaminated or sanitation fails, it spreads diseases such as cholera, typhoid, polio, diarrhoea, dysentery, and hepatitis A. Health hazards can be avoided by addressing these issues in time when there are poor or even no systems of water supply and sanitation. It is estimated that poor hand hygiene, deficient sanitation, and contaminated drinking water cause one million deaths annually due to diarrhoea. In 2021, more than 251.4 million individuals needed preventive therapy for schistosomiasis, an acute and chronic illness brought on by parasitic worms that are acquired by contact with contaminated water[3].

As more advanced epidemiological techniques have been developed recently, more data about the effects of water on health has been gathered. More research must be done to increase the understanding of the significance of transmission and the link between population exposure and diseases, especially considering the significance of the disease burden associated with water supply, sanitation, and hygiene (Fewtrell and Bartram, 2001). Many organisations globally are raising awareness and coming up with several initiatives to address this crisis, one of them is from the United Nations, and it belongs to their Sustainable Development Goals (SDGs) Goal 6: Clean water and Sanitation. This goal aims to ensure that everyone has access to affordable and safe drinking water and adequate sanitation facilities. Additionally, it targets reducing pollution, ending dumping, minimising hazardous releases minimisation; cutting down on untreated wastewater quantity; and increasing global recycling rates and safe reuse that would enhance water quality around the world[4].

[3] https://www.who.int/news-room/fact-sheets/detail/drinking-water
[4] https://sdgs.un.org/goals/goal6

When considering the global health implications of water quality, it is important not to overlook local circumstances. For example, understanding the quality of groundwater in certain areas, such as Ireland, will give visibility on issues and practices related to such areas. According to Bates et al. (2008), increased water temperatures and variations in extreme weather, such as droughts and floods, are expected to have an impact on water quality and aggravate a variety of water pollution issues. The following section focused on Irish groundwater, discussing its importance, actual status, main pollutants, and the organisations involved in monitoring and protecting these water bodies.

2.2 Groundwater Quality in Ireland

In Ireland, groundwater provides about 25% of the country's water supply and is crucial for agriculture, as drinking water, and for maintaining ecosystem health, as per the Groundwater Waterbody WFD Status 2016-2021, the overall chemical quality of 91% of groundwater bodies is in good condition (Craig and Daly, 2010). The preservation of groundwater ecosystems is crucial, since they contribute to the flows of rivers and lake levels when the aquifers release groundwater through springs and seeps into the sea as well as on lands (Williams and Lee, 2008). According to the type of aquifer beneath the surface water body and the time of year, each component's contribution changes (Craig and Daly, 2010). In certain instances, a significant amount (50–100%) of surface water may consist of groundwater discharge; therefore, the quality of these waters is impacted by the quality and quantity of groundwater discharged (EPA, 2010).

However, despite its importance and ongoing efforts to maintain its quality, Ireland's groundwater quality is threatened by various challenges such as human activities, pollution, and environmental factors (Robins and Misstear, 2000). Furthermore, it is anticipated that sea level rise will increase the salinisation of groundwater and estuaries, reducing the amount of freshwater that is available to people and ecosystems in coastal areas such as Ireland (Bates et al., 2008). This is particularly concerning for regions in the south and southeast, where the highest percentage of sites with elevated and rising nitrate concentrations are being found[5], requiring continuous monitoring and innovative approaches. The next section investigated and compared existing approaches for analysing and forecasting water quality using machine learning techniques and considered how these technologies can improve the monitoring and address the difficulties related to water quality.

2.3 Methods for Forecasting Water Quality Using Machine Learning Techniques

Machine learning (ML) is a developing field of computational algorithms that use environmental learning to simulate human intelligence (El Naqa and Murphy, 2015). With the fast growth of data on the aquatic environment, machine learning has emerged as a

[5] https://www.epa.ie/our-services/monitoring--assessment/freshwater--marine/groundwater/#:~:text=Groundwater%20status%20in%20Ireland

critical tool for data analysis, classification, and prediction, providing solutions to manage water pollution, enhance water quality, and protect water ecosystems (Zhu et al., 2022).

The study and analysis of the different water bodies and the use of ML have been investigated in several research projects published over the years. Zhu et al. (2022) present a comprehensive introduction of 45 ML techniques used to estimate water quality in different water bodies, water treatment and management systems, discussing the application of Decision Trees (DT), Support Vector Machines (SVM), and Random Forests (RF), among others, highlighting advancements in handling large datasets, and reviewing the performance metrics used to evaluate these models, enabling readers to understand the strengths and limitations of every model in different water contexts. However, the research lacks a comparison of the results and performance metrics of the various models discussed and also how these techniques can be customised to add depth to the analysis.

RF, SVM, Logistic Regression (LR), DT and, XGBoost and other classifier algorithms were utilised by Nasir et al. (2022) to forecast water quality accurately. The data used was collected between 2005 and 2014 from various states in India, They utilised accuracy, precision, recall and, F1 Score as performance measurements, finding that CATBoost and RF outperformed the other classifiers with 0.94 and 0.93 of accuracy, respectively. Despite the fact that the paper discussed the practical implications of the findings, the dataset was limited to only 1679 samples, and the authors did not discuss the data restrictions or possible findings with different datasets. On the other hand, Parnika et al. (2023) also utilised the same metrics and implemented RF, DT, and SVM to forecast the water potability using nine features and a target class potability, obtaining a 88.75% accuracy from RF; however, they highlighted the need of creating a robust model as future work since the results were not adequate for use in real world situations given that the quality of the water should not be ignored. Another limitation observed is that the authors simply removed observation with missing values instead of exploring other alternatives such as imputing missing data. Following classification problems, Krushna et al. (2024) applied Logistic Regression, DT, Gaussian Naïve Bayes, SVM, and XGBoost to analyse and forecast water quality, achieving a training accuracy of approximately 78% and 75% for SVM and XGBoost, respectively yet they pointed out the differences between the train and test results may indicate a possible overfitting. In this paper, the use of Synthetic Minority Oversampling Technique (SMOTE)to address class imbalance was a significant strength; nevertheless, the results suggested there is room for improvement, considering that the accuracy levels may not be enough when evaluating water quality. The authors focused future work on adding additional features, optimising model hyperparameters, and exploring ensemble methods. Another example of classification application in water quality potability forecasting is the work done by Kaddoura (2022), where RF, Gradient Boosting Trees, LR, XGBoost, DT, k-nearest neighbour (KNN), SVM among others, were implemented utilising Precision, Recall, F1 Score, and ROC AUC as performance metrics. The results showed that SVM and ANN outperformed the other algorithms, achieving F1-score values of 63.8% and 63.9%, respectively. Other models, including RF and KNN, showed acceptable F1-score values. While they use real-world data, such as Nasir et al. (2022), the dataset size may limit its

application in broader circumstances. Kaddoura also mentioned that the proposed approach would be refined in future work to improve the performance of these algorithms by tuning the hyperparameters to discover the ideal model configuration to obtain the most optimised result. Like previous authors, Radhakrishnan and Pillai (2020) implemented SVM, Decision Tree, and Naïve Bayes to classify and predict water quality. With an accuracy of 98.50%, the decision tree algorithm was determined to be the best performer classification model, demonstrating its suitability for this classification task, yet the accuracy difference between the two datasets may indicate overfitting and raise questions of the model's performance since there is a lack of discussion of the hyperparameter tuning carried out.

Table 1. summarises the algorithms, evaluation metrics, and results obtained in each of the reviewed papers. All the studies contributed to valuable insights into water quality assessment using ML techniques and provided different techniques to address data pre-processing, modelling, and evaluation processes, highlighting the importance of data understanding, model and feature selection, and hyperparameter tuning.

Algorithm	Evaluation Metrics	Results	Author
LR	Accuracy, Precision, Recall, F1 Score	0.7291, 0.7247, 0.7292, 0.7249	Nasir et al.
SVM	Accuracy, Precision, Recall, F1 Score	0.8068, 0.81302, 0.8068, 0.80601	Nasir et al.
DT	Accuracy, Precision, Recall, F1 Score	0.81623, 0.8169, 0.8163, 0.156	Nasir et al.
XGB	Accuracy, Precision, Recall, F1 Score	0.8807, 0.8836, 0.8807, 0.8804	Nasir et al.
MLP	Accuracy, Precision, Recall, F1 Score	0.8863, 0.8890 0.8863, 0.8864	Nasir et al.
RF	Accuracy, Precision, Recall, F1 Score	0.9393, 0.9397, 0.9393, 0.9394	Nasir et al.
CATBoost	Accuracy, Precision, Recall, F1 Score	0.9451, 0.9458, 0.9451, 0.9449	Nasir et al.
RF	Accuracy, Precision, Recall, F1 Score	0.8875, 0.89, 0.89, 0.89	Parnika et al.
DT	Accuracy, Precision, Recall, F1 Score	0.78, 0.78, 0.77, 0.78	Parnika et al.
SVC	Accuracy, Precision, Recall, F1 Score	0.63, 0.64, 0.62, 0.63	Parnika et al.
LR	Train and Test Accuracy	70.18%, 67.00%	Krushna et al.
DT	Train and Test Accuracy	71.3%, 68.15%	Krushna et al.
GNB	Train and Test Accuracy	71.06%, 70.69%	Krushna et al.
SVM	Train and Test Accuracy	78.48%, 57.50%	Krushna et al.
XGB	Train and Test Accuracy	74.73%, 68.75%	Krushna et al.
RF	Precision, Recall, F1-Score, ROC AUC	45.9, 92.8 61.4, 0.702	Kaddoura
XGB	Precision, Recall, F1-Score, ROC AUC	45.7, 89.8 60.6, 0.667	Kaddoura
DT	Precision, Recall,	45.1, 88.6,	Kaddoura

	F1-Score, ROC AUC	59.8, 0.654	
SVM	Precision, Recall, F1-Score, ROC AUC	50.0, 88.0, 63.8, 0.731	Kaddoura
SVM	Accuracy dataset 1, dataset 2	87.10, 95.63	Radhakrishnan et al.
Decision Tree	Accuracy dataset 1, dataset 2	87.10, 98.50	Radhakrishnan et al.
Naïve Bayes	Accuracy dataset 1, dataset 2	74.60, 95.17	Radhakrishnan et al.

Table 1. Evaluation of Existing Water Quality Models

3 Research Methodology

This section outlines the research methodology applied as well as the factors considered during the research, analysis, and forecasting of groundwater quality indicators in the context of Ireland.

3.1 Groundwater Prediction Methodology

The Groundwater Prediction Methodology used in this research project was specifically tailored based on CRISP-DM with the clear objective of answering our research question and objectives. CRISP-DM was selected to its well-defined structure that offer guidance and clarity throughout its six phases, being frequently utilised for machine learning projects. This decision ensured that the study followed a clear and defined course, starting with the understanding of the problem and the data collected, continuing with the preparation and modelling to finally conclude the interpretation and evaluation of the implemented models. The groundwater quality parameters in Ireland were analysed and forecasted using the adapted CRISP-DM approach, as illustrated in Figure 1.

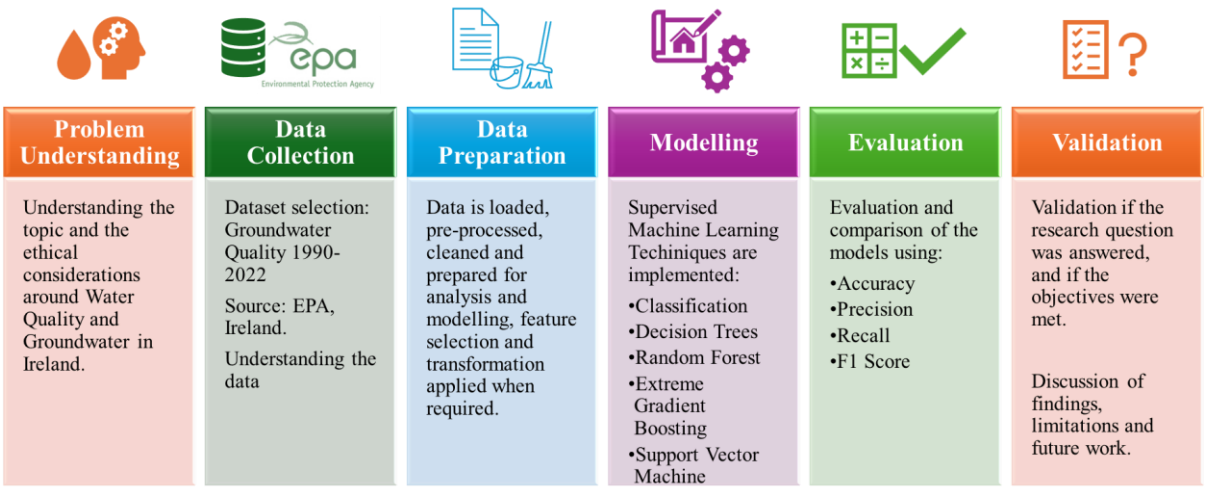

Figure 1. Groundwater Prediction Methodology

3.1.1 Problem Understanding

This phase was completed in Section 2 of this document, where Objective 1 was completed. Related Work section explored, analysed, compared and contrasted relevant water quality literature and its implications on health and ecosystems, it was also highlighted the importance and status of groundwater in Ireland, as well as the main factors influencing the

quality of groundwater water bodies and finally we concluded with the most currently used methods for the analysis and prediction of water quality parameters and how the incorporation of machine learning techniques have streamlined and improved the processes applied to analyse and prognosticate water quality nowadays.

3.1.2 Data Collection

After a thorough investigation and motivated by one of the United Nations Sustainable Development Goals (SDGs) and being the country in which I currently reside, the principal study region chosen for this research was Ireland, specifically Ireland's groundwater water bodies. The Groundwater Quality dataset used for this research was obtained from the Environmental Protection Agency Geo Portal. The dataset includes the samples obtained through monitoring groundwater stations in Ireland between 1990 and 2022. Physical factors such as temperature and turbidity, and chemical factors like pH, dissolved oxygen, conductivity and hardness were included in the dataset, along with identification information such as site name, county, code, and the date of the sample.

3.1.3 Data Preparation

The dataset was loaded into R Studio where we performed the data cleaning, transformation, feature selection, analysis and visualization. R Studio was chosen due to its ability to handle huge datasets efficiently, as we needed to work with a dataset that included 16,231 records and 304 attributes. Once the dataset was cleaned, the data was loaded into Jupyter through Anaconda navigator, where the chosen algorithms were implemented, analysed, and evaluated. Python programming language was selected for this step due to its powerful libraries specifically useful for machine learning and data analysis.

As mentioned above, the groundwater quality dataset was loaded in R Studio, where the data was examined, pre-processed and cleaned. The first step taken in the data pre-processing phase was to identify the percentage of NA and "--" values of each attribute (columns) it was observed that >200 attributes of the dataset consisted of NA or "--" values. To address this issue, reduction criterion was utilised removing columns from the dataset that had more than 10% missing values. The purpose of this decision was to improve the dataset quality. High percentages of missing values often result in bias from differences between complete data and missing values (Kaiser, 2014). Originally, the dataset consisted of 304 attributes, once the function was executed, the dataset was reduced to 58 variables, this allowed us to reduce the necessary computational resources but above all increase the performance of the models and avoid biased results. Subsequently, the second row containing the units for each of the parameters was deleted. Then, to decide whether to use the mean or median to impute the NA values, we excluded the identification variables (County, Site Name, New Code and Sample Date), and histograms of all numerical variables were generated to validate the distribution of each column, the skewness was then checked and the appropriate imputation method was determined (As mentioned by Donders at al. (2010)mean imputation if the skewness was close to 0 and the distribution was symmetric and median imputation if the skewness was significantly different from 0 and the distribution was asymmetric).

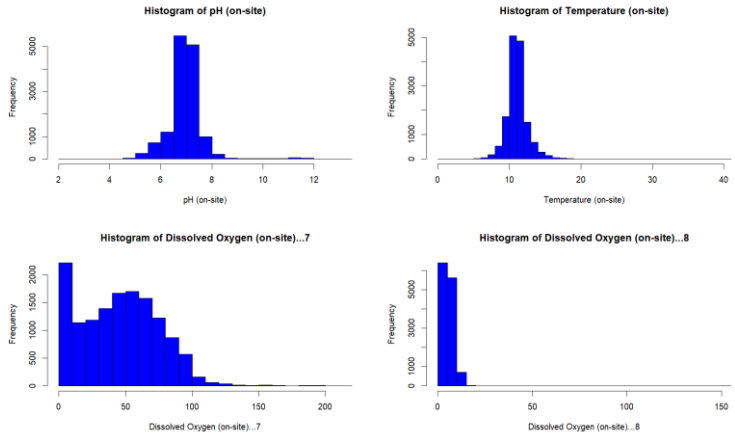

Figure 2. Variables Skewness

After confirming the appropriate imputation method, the NA values were replaced by the mean or median respectively, and then excluded columns and the numerical columns dataset were combined into a single data frame and then the sample date was transformed into a date format. The final dataset consisted of 58 variables and a total of 16,229 samples taken from 1990 to 2022. Once the data was cleaned, several visualizations were created to analyse the patterns and relationships within the dataset. These visualizations included time series plots, box plots, histograms, and bar charts. To analyse interdependencies between variables, a detailed heatmap representing a correlation matrix was generated, showing strong, subtle and no correlations between our variables. The heatmap showed that mostly all our variables were not strongly correlated but still there were few candidates for predictors.

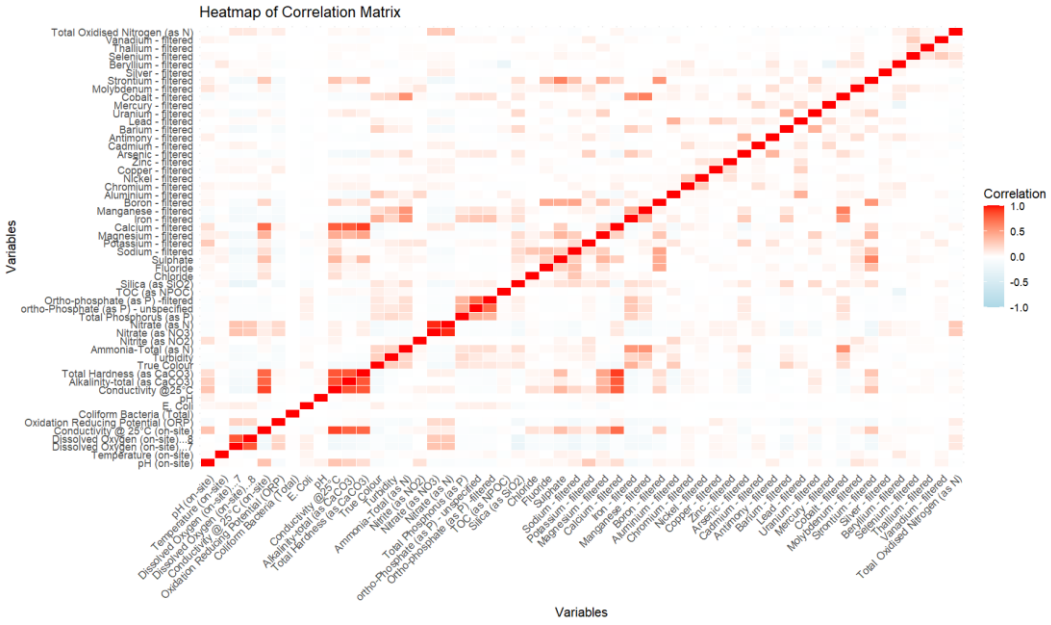

Figure 3. Correlation Matrix Heatmap

Following the data preparation phase, the remaining data attributes underwent a cleaning process that included transformation and encoding where needed. The dataset was

cleared of outliers, redundant data, and inconsistencies to make sure the data was ready to use to obtain the most accurate results. An 80:20 ratio was used to split the dataset into training and test sets. NaN values, and rows with missing values were eliminated to ensure clean data. The cleaned and transformed dataset was then used for analysis and implementation. Sections that were addressed later in this document.

The data exploration step started with the cleaned dataset loaded into Python, where we applied domain knowledge for feature selection, allowing us to select the most relevant parameters for analysis and prediction of groundwater quality. The analysis was performed basing the selection on an understanding of the problem, metadata, and general purpose of the investigation. The literature indicated that not only pH, Temperature, Conductivity, Dissolved Oxygen, Ammonia, Phosphate among oThers are important factors that affect groundwater quality but also heavy metals such as Lead, Cadmium, and Arsenic.

In total, four groundwater quality parameters regularly tested in Ireland - Conductivity, Dissolved Oxygen, Alkalinity and Nitrate (as NO3)- were selected for this study. In particular, the Water Framework Directive (2000/60/EC) includes Conductivity, Dissolved Oxygen, and Nitrate are in the list of parameters that must be monitored (EPA, 2003). This emphasises the importance of these parameters in determining and guaranteeing the quality of groundwater resources.

Alkalinity – This parameter measures the capacity of the water to resist changes and neutralise acids, avoiding acidity. Carbonate, phosphates, and hydroxides are common substances that raise the alkalinity of water (Dohare et al., 2014).
Conductivity – Measures the ability of water to conduct electricity, and it depends on the quantity and kind of ions present in the solution (Dohare et al., 2014). Conductivity is correlated with the quantity of charged particles in the water, and it is a crucial metric in evaluations of groundwater quality for irrigation and drinking water (Tutmez et al., 2006).
Dissolved Oxygen – Shows the amount of oxygen that is present in water. By controlling the valence state of trace metals and limiting the bacterial metabolism of dissolved organic species, dissolved oxygen (D.O.) concentration has a major impact on groundwater quality (Rose and Long, 1988).
Nitrate (as NO3) – Shows the amount of nitrate ions present in the sample. Increased Nitrate as NO3 concentrations in groundwater can be concerning because they can indicate a loss of soil fertility above the surface, result in eutrophication when groundwater discharges into surface water, and pose health risks to both people and animals (McLay et al., 2001).

The target variables, "Alkalinity-total (as CaCO3)," "Dissolved Oxygen (on-site)," "Conductivity @25°C," and "Nitrate (as NO3)," were binned into three categories (low, medium, and high) and transformed using quantiles before starting the implementation of all our four models.

3.1.4 Modelling

The modelling phase consisted of the selection and implementation of four machine learning techniques, analysis and forecasts of the groundwater bodies of the Republic of Ireland. Following the investigation objectives, our first approach was to classify the groundwater quality establishing thresholds for key hydrochemistry parameters. After, Decision Trees, Random Forest, Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM) classifiers models were implemented. The classification techniques and the models were specifically chosen based on their suitability to the problem making them ideal for analysing and forecasting groundwater quality parameters, making sure we effectively address the research question and meet our research objectives. The detailed steps and processes followed during data modelling were discussed further in section 5 of this paper.

3.1.5 Evaluation

The metrics chosen for this research were accuracy, precision, recall, and F1 Score. These metrics were chosen based on their suitability and wide application for classification algorithms, as stated in the research examined in Section 2 of this paper. Each evaluation metric is discussed below.

- **Accuracy:** Accuracy is the most used measure for classifier evaluation. It evaluates the algorithm overall efficiency by predicting the probability of the true value of the class label Malek et al. (2022). The following equation is used to determine accuracy:

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Instances}}$$

- **Precision:** According to Nasir et al. (2022) precision refers to the ratio of accurately predicted positive observations to the total number of expected positive observations. Precision is calculated using the equation below:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

- **Recall:** The true positive rate is typically used to describe the model's sensitivity or recall. It determines how frequently the algorithm detects the proper classification from the given data versus the actual accurate classification occurring in the dataset (Udin et al. 2023). Recall is determined using the following equation:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

- **F1 Score:** F1-score is another indicator of a model's accuracy on a dataset. It evaluates multiclass classification. It is a method for balancing the precision and recall of a prediction model (Uddin et al. 2023). The F1-score is obtained as follows:

$$\text{F1} = 2 \text{ x} \frac{\text{Precision x Recall}}{\text{Precision + Recall}}$$

4 Design Specification

The key components, requirements, and steps that guided this research are shown in this section. These elements provided the structure of the research to meet the objectives and answer the research question. The specification design shows at a high level the stages that guided the research process and are shown graphically in Figure 4.

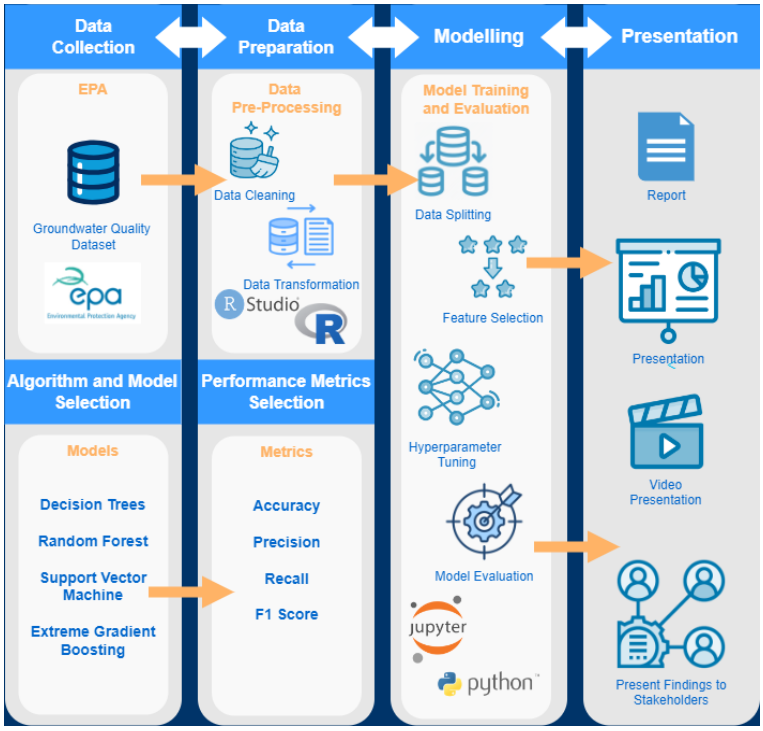

Figure 4. Groundwater Quality Parameter Design Specification

Data collection indicates the source of our dataset, Algorithm and Model Selection presents the selected models for our research, Data Preparation display the steps taken during the data pre-processing, following the Performance Metrics Selection where the metrics were selected based on their suitability to the problem and its wide application to the selected models, proceeding with Modelling and the actions during this phase, and finally the presentation layer where the conclusion, findings and limitations were presented to stakeholders.

5 Implementation and Evaluation

Implementation: Four supervised machine learning models were developed and evaluated. Decision Trees, Random Forest, Extreme Gradient Boosting and Support Vector Machine were implemented employing Python as the programming language, utilising Scikit-learn (Sklearn) and XGBoost libraries.

Evaluation: The implemented models were evaluated using four performance metrics from the sklearn.metrics module from Python.

5.1 Groundwater Quality Classification Implementation and Evaluation

In general, Ireland's water bodies quality is good, in specific, groundwater quality is monitored and evaluated through different monitoring programs for regulated operations, drinking water supplies, and the EPA national groundwater monitoring program[6]. This monitoring aligns with the growing acceptance of the idea of "acceptable" or "tolerable risk", recognising that while some risk cannot be fully accepted but may be accepted to some extent or in consideration of more important or pressing matters (Fewtrell and Bartram, 2001). Strict rules and regulations may work against the good uses of water, making it impossible for society to profit from them.

In this context, the EPA in Ireland employs Guideline Values for specific parameters to assess whether a sample requires further testing, balancing strict regulations with the practical use and benefits of groundwater resources.

The first experiment conducted in our research was Groundwater Quality Classification utilising the Guideline Values for the Protection of Groundwater in Ireland. These values were intended to be used in the process of characterising groundwater bodies and to determine whether further research or action is required in the event the guideline values were exceeded (EPA, 2003). First, a classification algorithm was implemented using interim guideline values for different groundwater quality metrics, shown in the table 2 below. After applying this function to the dataset, a new column was created that showed whether each sample needed further action or if all the parameters were within the guideline values.

Parameters (mg/l)	**Interim Guideline Value**
Ammonia (as ammonium)	0.15
Calcium	200
Chloride	30
Hardness as CaCO3	200
Magnesium	50
Nitrate	25
Potassium	5
Sodium	150
Sulphate	200

Table 2. Interim Guideline Values

Second, a bar chart was used to show the distribution of groundwater quality parameters classifications after summarising the findings of the classification process and histograms were used to illustrate the frequency of different values for each parameter to better understand the data and the distribution of the evaluated parameters. Finally, using 'Sample Date' as the time axis, time series plots were generated to observe the trends of these parameters over time.

[6] https://www.epa.ie/our-services/monitoring--assessment/freshwater--marine/groundwater/

The Groundwater Quality Classification experiment effectively classified the samples based on the interim guideline values and provided useful insights through different types of visualisations. The classification function was successful in identifying samples that fell within acceptable limits and those that needed further investigation as shown in Figure 5. The findings suggested that a significant number of samples were categorised as “Further Action Required” implying that either there is a general problem with the groundwater bodies, or the guidelines interim values are strict and need to be reevaluated with a count of 13978 vs 2251.

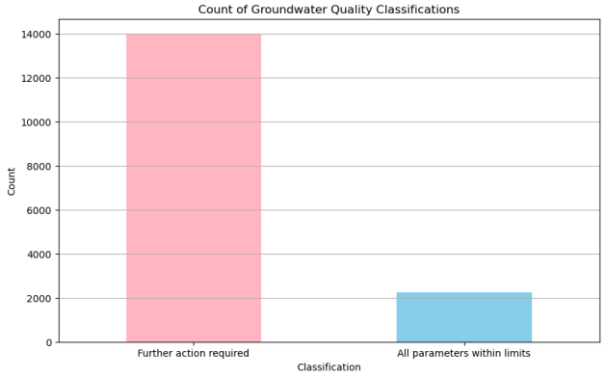

Figure 5. Groundwater Classification

Time series plots provided a temporal perspective on changes in groundwater quality by illustrating trends and fluctuations across time, being particularly helpful when analysing historical data as shown in Figure 6.

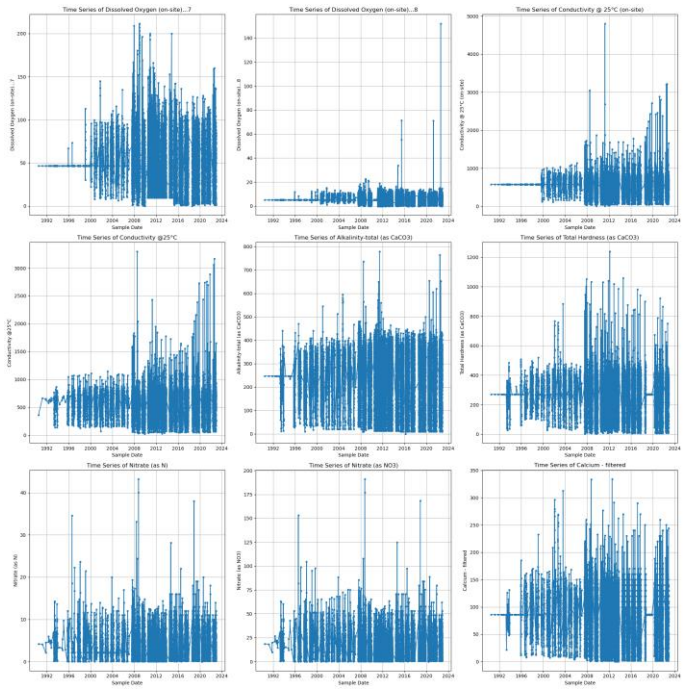

Figure 6. Time Series IGV

In conclusion, the experiment proved effective in offering a review of data related to groundwater quality. Given the large number of samples that need further action, the interim guideline values should be reviewed. With this experiment, Objective 2 is completed.

5.2 Decision Tree Implementation and Evaluation

According to Nasir et al. (2022) definition a decision tree is a recursive top-down division that follows a divide-and-conquer method. Its underlying algorithm is essentially greedy. The creation of a decision tree is divided into two stages: tree building and pruning. Being the tree-building stage the first stage, during which a subset of the training data is chosen, and a decision tree is constructed using the breadth-first recursive technique until each leaf node belongs to the same class. The second is the pruning stage, which uses the remaining data to analyse the built decision tree and correct any errors, prunes, and adds nodes until a good decision tree is formed.

The Decision Tree classifier was the first model applied to our data. The model was trained on the training set with an 80:20 split ratio, GridSearchCV from the Scikit-learn library was used to implement hyperparameter tuning to find the best parameters, max_depth, min_samples_split, min_samples_leaf, were applied. Grid search investigates all possible parameter combinations within the stated parameter ranges, thoroughly examining the parameter space to discover the optimum configuration (Wang et al., 2023). Once we obtained the best parameters, the final model was trained using these parameters. Finally, the model was evaluated using the Scikit-learn library's accuracy, precision, recall, and F1 score metrics, shown in Table 3 below.

Parameter	Accuracy	Precision	Recall	F1 Score
Alkalinity-total (as CaCO3)	0.764017	0.771619	0.764017	0.766490
Dissolved Oxygen	0.882009	0.886731	0.882009	0.882903
Conductivity	0.817930	0.823680	0.817930	0.819675
Nitrate (as NO3)	0.959951	0.963180	0.959951	0.960275

Table 3. Decision Tree Classifier Performance

5.3 Random Forest Implementation and Evaluation

Random Forest is an ensemble learning method that uses decision trees to generate numerous weak learners. The final forecasts were made by combining the predictions of individual trees via average or vote. Random Forest splits just a random subset of features at each decision tree node (Wang et al., 2023).

Random Forest classifier was the next model implemented to our data. Like DT configuration hyperparameter tuning was done using GridSearchCV technique from Scikit-learn library. The tuned parameters were n_estimators, max_depth, min_samples_split, and min_samples_leaf, and criterion. GridSearchCV, in association with 3-fold cross-validation, was used to determine the optimum combination from these parameters. Once the best parameters were identified, the final Random Forest classifier was trained using these values. The trained model was then utilised to forecast the target variables from the test dataset. Finally, the model performance was tested using aCcuracy, precision, recall, and F1 score metrics, shown in Table 5. Based on the best set of parameters, the Random Forest classifier was finally trained with these values. Afterwards, the target variables of the test dataset were forecasted by the trained model. In the final step, the performance of the model is tested using aCcuracy, precision, recall, and F1 score metrics, which are shown in Table 4.

Parameter	Accuracy	Precision	Recall	F1 Score
Alkalinity-total (as CaCO3)	0.812384	0.814800	0.812384	0.813144
Dissolved Oxygen	0.881701	0.889554	0.881701	0.883216
Conductivity	0.837338	0.840674	0.837338	0.838631
Nitrate (as NO3)	0.959951	0.963264	0.959951	0.960285

Table 4. Table 5. Random Forest Classifier Performance

Random Forest feature importance was also checked, showing that Total Hardness (as CaCO3) is the most relevant variable for this model as presented in Figure 7 below.

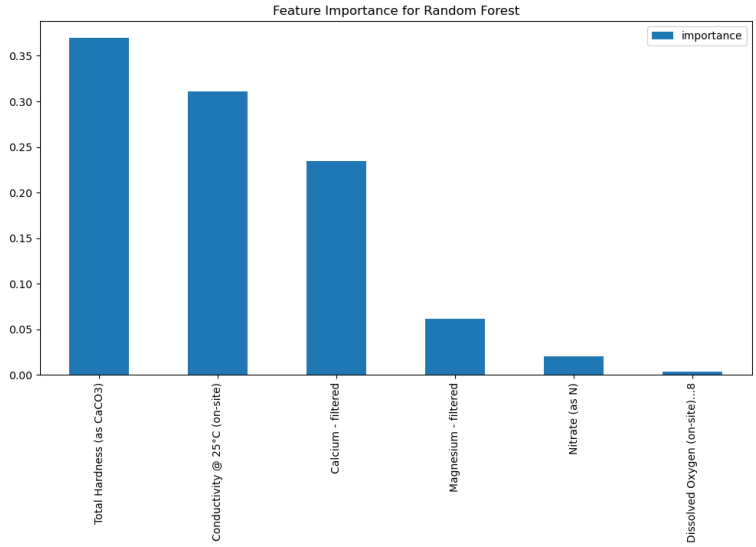

Figure 7. Random Forest Feature Importance

5.4 Extreme Gradient Boosting Implementation and Evaluation

Extreme Gradient Boosting (XGBoost) is an algorithm based on the gradient boosting decision tree method. Using regularisation and parallel computing to improve its accuracy and efficiency. XGBoost employs the gradient boosting technique, which iteratively trains a series of decision trees to gradually improve the prediction model performance (Wang et al., 2023).

Next, we modelled and evaluated XGBoost classifier and conducted hyperparameter tuning using GridSearchCV from Scikit-learn library. These parameters were learning rate (learning_rate), number of estimators (n_estimators) and maximum depth of trees. Determining an optimal combination of those parameters was performed through GridSearchCV with 2-fold cross validation. After finding the ideal parameters the final XGBoost model was trained. Finally, the target variables were forecasted using the test dataset, resulting in the following metrics:

Parameter	Accuracy	Precision	Recall	F1 Score
Alkalinity-total (as CaCO3)	0.804683	0.805669	0.804683	0.804001
Dissolved Oxygen	0.883549	0.890192	0.883549	0.884754
Conductivity	0.842267	0.843493	0.842267	0.842811
Nitrate (as NO3):	0.956254	0.960738	0.956254	0.956768

Table 5. Extreme Gradient Boost Classifier Performance

5.5 Support Vector Machine Implementation and Evaluation

SVM is a discriminative model based on building a hyper-plane to minimise errors and it can be used for both classification and regression problems (Nasir et al., 2022). SVM addresses overfitting difficulties in ML by lowering model complexity and successfully fitting training data (Malek et al., 2022).

Finally, a SVM classifier was then modelled and evaluated, as the previous models hyperparameter tuning was done using GridSearchCV from the Scikit-learn library. The parameters tuned were the regularisation parameter (C), the kernel coefficient (gamma), and the kernel type. GridSearchCV was applied using 2-fold cross-validation to determine the optimal combination of these parameters. Once the ideal parameters were identified, the final SVM model was trained using these values. The trained model was then utilised to predict the target variables from the test dataset. The model's performance was evaluated using accuracy, precision, recall, and F1 score metrics, results shown in Table 6 below:

Parameter	Accuracy	Precision	Recall	F1 Score
Alkalinity-total (as CaCO3)	0.777880	0.786162	0.777880	0.780756
Dissolved Oxygen	0.461491	0.496261	0.461491	0.437621
Conductivity	0.811768	0.818579	0.811768	0.814095
Nitrate (as NO3)	0.450709	0.464948	0.450709	0.426129

Table 6. Support Vector Machine Classifier Performance

6 Discussion

The experiments of this research were focused on predicting four groundwater quality parameters using four classification machine learning models. These models were evaluated using accuracy, precision, recall, and F1 Score metrics. The findings, evaluation, and interpretation are discussed in detail below.

Table 7. shows how Random Forest and XGBoost consistently outperformed other models in most parameters, indicating that they are the most reliable models for forecasting groundwater quality, suggesting that ensemble methods are more effective for this data. SVM perform poorly across all parameters, particularly with Dissolved Oxygen and Nitrate prediction with a F1 lower than 50%, indicating a possible overfitting or underfitting issue. For Alkalinity, Random Forest demonstrated the highest performance across all metrics, followed by XGBoost, being DT and SVM the less effective, suggesting that more extensive hyperparameter tuning might help to optimise DT and SVM performance. In the case of Dissolved Oxygen, XGBoost outperformed the other models, achieving the best accuracy (0.88) and maintaining a balanced precision and recall (0.89 and 0.88, respectively), followed by DT and RF slightly behind, and finally SVM showing poor performance (0.46 accuracy). These results indicateD issues with the dataset characteristics, and one approach to improving SVM could be to explore different kernel functions or check non-linearity. For conductivity, XGBoost showed the highest scores, followed by RF, DT, and SVM with the lowest scores,

suggesting issues with the model implementation or hyperparameter tuning. Finally, for Nitrate, RF and DT were the top performers, indicating that decision-based models were suitable for this parameter, followed by XGBoost. SVM showed significantly lower performance. One way to improve both XGBoost and SVM could be to evaluate how DT and RF performed so well and validate if this can be translated to the other two models.

Parameter	Model	Accuracy	Precision	Recall	F1 Score
Alkalinity-total (as CaCO3)	DT	0.764017	0.771619	0.764017	0.766490
	RF	0.812384	0.814800	0.812384	0.813144
	XGBoost	0.804683	0.805669	0.804683	0.804001
	SVM	0.777880	0.786162	0.777880	0.780756
Dissolved Oxygen	DT	0.882009	0.886731	0.882009	0.882903
	RF	0.881701	0.889554	0.881701	0.883216
	XGBoost	0.883549	0.890192	0.883549	0.884754
	SVM	0.461491	0.496261	0.461491	0.437621
Conductivity	DT	0.817930	0.823680	0.817930	0.819675
	RF	0.837338	0.840674	0.837338	0.838631
	XGBoost	0.842267	0.843493	0.842267	0.842811
	SVM	0.811768	0.818579	0.811768	0.814095
Nitrate (as NO3)	DT	0.959951	0.963180	0.959951	0.960275
	RF	0.959951	0.963264	0.959951	0.960285
	XGBoost	0.956254	0.960738	0.956254	0.956768
	SVM	0.450709	0.464948	0.450709	0.426129

Table 7. Machine Learning Model Comparison

Objective 4: evaluating, comparing, and contrasting the performance of the implemented models was achieved. The results showed that RF and XGBoost outperformed the other models in terms of accuracy, although XGBoost had superior precision and recall rates, completing Objective 5 with the identification of the best performance model successfully and exploring real-world applications.

The study findings provided important insights on how machine learning algorithms might be used in real-world applications, identifying how different geological changes, like, for example, erosion, sedimentation, and deposition, among others, can affect groundwater parameters and detect pollution utilising threshold values. Previous studies agreed that ensemble approaches such as Random Forest and XGBoost were successful at managing complex, nonlinear connections in environmental data. Furthermore, the results emphasised how crucial model tuning, and feature selection are for environmental data forecasting. The findings were consistent with previous research, demonstrating the effectiveness of these models in forecasting groundwater quality parameters. However, our findings revealed the limitations of SVM in this context, which contradicts several studies that showed that SVM is effective in water quality applications. This does not necessarily mean that the model cannot be used for any task that involves prediction of water quality, as demonstrated in the literature review. Rather, it emphasised the need to consider the data characteristics, more powerful tuning or alternate kernel functions.

7 Conclusion and Future Work

This study successfully demonstrated the importance of supervised machine learning in predicting groundwater quality parameters, highlighting the effectiveness of Random Forest and XGBoost. Both models exhibited better performance when forecasting the four target variables: Alkalinity, Conductivity, Dissolved Oxygen, and Nitrate, displaying robustness in the handling of the nonlinear relationships of the data. Whereas Decision Trees showed moderate performance, SVM underperformed, especially when prognosticating Dissolved Oxygen and Nitrate, indicating that model and feature selection, along with tuning, are crucial for achieving greater results. These results underscored the importance of machine learning as a strategy that can be used simultaneously with traditional monitoring methods, providing a scalable and possibly more accurate approach.

While this research has provided useful insights into the application of machine learning for groundwater quality forecasting, the study has several limitations and future work. The range of parameters was limited, impacting how applicable the models are. Future research should focus on adding external factors such as climate data and/or industrial or agricultural activities to increase metrics and generalisation. Increasing the computational resources will also allow us to explore more complex models, such as expanding the number of estimators or tree depth. Model optimisation, including hyperparameter tuning and investigating other machine learning algorithms or hybrid models like artificial neural networks (ANN), AdaBoost, or stacking, may produce better outcomes. Finally, in terms of policy implications, further research could look into how this implementation could be translated to real-world scenarios and how its application may affect monitoring techniques or environmental policies.

8 Acknowledgment

I would like to take this opportunity to express my deep regards to my supervisor for the last minute reviews and to other colleagues who have either directly or indirectly helped and guided me through this research. I would also like to express my gratitude to my family and friends that were always cooperative and encouraging, that kept me motivated during difficult times. Thanks, to all of those who supported me in any way during the completion of this thesis.

References

Bates, B., Kundzewicz, Z. and Wu, S., 2008. *Climate change and water*. Intergovernmental Panel on Climate Change Secretariat.

Craig, M. and Daly, D., 2010. Methodology for establishing groundwater threshold values and the assessment of chemical and quantitative status of groundwater, including an assessment of pollution trends and trend reversal. *Environmental Protection Agency*.

Dohare, D., Deshpande, S. and Kotiya, A., 2014. Analysis of ground water quality parameters: a Review. *Research Journal of Engineering Sciences ISSN*, *2278*, p.9472.

Donders, A.R.T., Van Der Heijden, G.J., Stijnen, T. and Moons, K.G., 2006. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, *59*(10), pp.1087-1091.

El Naqa, I. and Murphy, M.J., 2015. *What is machine learning?* (pp. 3-11). Springer International Publishing.

EPA, I., 2010. Methodology for Establishing Groundwater Threshold Values and the Assessment of Chemical and Quantitative Status of Groundwater. *Including an Assessment of Pollution Trends and Trend Reversal*.

Fewtrell, L. and Bartram, J. eds., 2001. *Water quality: guidelines, standards & health*. IWA publishing.

Groundwater, P.O., 1993. Towards Setting Guideline Values for the Protection of Groundwater in Ireland.

Kaddoura, S., 2022. Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability. Sustainability 2022, 14, 11478.

Kaiser, J., 2014. Dealing with Missing Values in Data. *Journal of Systems Integration (1804-2724)*, *5*(1).

Krushna, B.V. and Sasikala, D., 2024, January. Comparative Analysis of Machine Learning Models for Water Quality Prediction. In *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)* (pp. 1-6). IEEE.

Lo, M.H., Famiglietti, J.S., Reager, J.T., Rodell, M., Swenson, S. and Wu, W.Y., 2016. Grace-based estimates of global groundwater depletion. *Terrestrial Water Cycle and Climate Change: Natural and Human-Induced Impacts*, pp.135-146.

Malek, N.H.A., Wan Yaacob, W.F., Md Nasir, S.A. and Shaadan, N., 2022. Prediction of water quality classification of the Kelantan River Basin, Malaysia, using machine learning techniques. *Water*, *14*(7), p.1067.

McLay, C.D.A., Dragten, R., Sparling, G. and Selvarajah, N., 2001. Predicting groundwater nitrate concentrations in a region of mixed agricultural land use: a comparison of three approaches. *Environmental pollution*, *115*(2), pp.191-204.

Narasaiah, M.L., 2005. *Water and sustainable tourism*. Discovery Publishing House.

Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A. and Al-Shamma'a, A., 2022. Water quality classification using machine learning algorithms. *Journal of Water Process Engineering*, *48*, p.102920.

Parkes, M.W., Morrison, K.E., Bunch, M.J., Hallström, L.K., Neudoerffer, R.C., Venema, H.D. and Waltner-Toews, D., 2010. Towards integrated governance for water, health and social–ecological systems: The watershed governance prism. *Global Environmental Change*, *20*(4), pp.693-704.

Parnika, J., Devamane, S.B., Ramya, C.N. and Laxmi, M.S., 2023, June. Performance Analysis of Machine Learning Algorithms to Predict Water Potability. In *2023 International Conference on Computational Intelligence for Information, Security and Communication Applications (CIISCA)* (pp. 230-235). IEEE.

Robins, N.S. and Misstear, B.D., 2000. Groundwater in the Celtic regions. *Geological Society, London, Special Publications*, *182*(1), pp.5-17.

Rose, S. and Long, A., 1988. Monitoring dissolved oxygen in ground water: some basic considerations. *Groundwater Monitoring & Remediation*, *8*(1), pp.93-97.

Tutmez, B., Hatipoglu, Z. and Kaymak, U., 2006. Modelling electrical conductivity of groundwater using an adaptive neuro-fuzzy inference system. *Computers & geosciences*, *32*(4), pp.421-433.

Uddin, M.G., Nash, S., Rahman, A. and Olbert, A.I., 2023. Performance analysis of the water quality index model for predicting water state using machine learning techniques. Process Safety and Environmental Protection, 169, pp.808-828.

Wang, X., Li, Y., Qiao, Q., Tavares, A. and Liang, Y., 2023. Water quality prediction based on machine learning and comprehensive weighting methods. *Entropy*, *25*(8), p.1186.

Williams, N.H. and Lee, M., 2008. Ireland at risk–Possible implications for groundwater resources of climate change. *Geol. Surv. Irel*, *13*, pp.1-28.

Wood, W., 1976. Guidelines for collection and field analysis of ground water samples for selected unstable constituents.

Younger, P.L., 2009. *Groundwater in the environment: an introduction*. John Wiley & Sons.

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B. and Ye, L., 2022. A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health*, *1*(2), pp.107-116.